

GOVT 613: Managing Your Data in Stata

Most of the time your data will not come to you ready to be analyzed. As we discussed last time, you may need to recode your variables so that they work properly in your model. Another frequent problem is that you have multiple datasets that you want to combine. Usually this occurs in one of two ways: you want to add variables to an existing dataset or you want to add observations to an existing dataset. I take you through the process of doing both below.

Merging Data

The merge command is used when you want to add variables to a dataset. The key to doing this is having some way of telling Stata which observations match from one dataset to the other. For example, the National Election Study has a unique caseid for every respondent to their survey. But sometimes you need more than one variable to properly identify individual cases.

For example, I have two datasets on congressional elections. One is located at <http://nw08.american.edu/~schaffne/incvote.dta> and includes an observation for each congressional district. The dataset has three variables: state (state of the congressional candidate), dist (congressional district number), and incvote (the percentage of the vote won the incumbent candidate won). But I am interested in examining whether the amount of money the incumbent spends in the campaign affects the percentage of the vote he/she wins. I have a separate dataset that has this information located at <http://nw08.american.edu/~schaffne/campspend2.dta>. The variables in this dataset are state, dist, and spent (for the amount of money spent by the incumbent). So, I can merge these data by state and district to add the spent variable to my first dataset.

To merge datasets, you have to make sure that both datasets are sorted by the variable(s) that you intend to use to merge them. In this case, we are merging by state and district, so both datasets must be sorted by state and district. I have already sorted the campspend2 data in this way. But open the incvote dataset and begin by sorting that data.

```
use http://nw08.american.edu/~schaffne/incvote.dta
sort state dist
```

Your dataset is now sorted so you can merge in your campaign spending dataset with the following command.

```
merge state dist using http://nw08.american.edu/~schaffne/campspend2.dta
```

When Stata merges two datasets, it creates a new variable called `_merge`. This variable indicates whether a case was successfully matched or not. The variable can take on three values:

- 1 = this observation comes from the "master" file only (the file in memory)
- 2 = this observation comes from the "using" file only (the file on disk)
- 3 = this observation was common to both the master and using files (a match)

If you tabulate `_merge` after performing the above merge you will get the following. (You should ALWAYS tabulate `_merge` after you merge two datasets to make sure that the merge was successful. If you have 1's or 2's on this variable that you cannot explain, then you must figure out why they are there.)

```
. tab _merge
```

<code>_merge</code>	Freq.	Percent	Cum.
1	35	8.05	8.05
3	400	91.95	100.00
Total	435	100.00	

The 3's are good. The 400 observations that have a 3 are the ones that were successfully matched from the two datasets. However, you also have 35 1's. This means that 35 observations that were in your incvote dataset were not in your campspend2 dataset. This makes sense when you realize that 35 members of Congress did not run for reelection in 2000, thus, they could not have spent any money. Note that there are no 2's, so there are no observations in your campspend2 dataset that are not in your incvote dataset.

Appending Data

The other common thing that you may need to do is add observations to an existing dataset. In this case, you are appending to your dataset, not merging. The `append` command in Stata essentially adds a new dataset to the bottom of your existing one. For any variables that exist in one dataset but not in another, Stata will add those variables to the dataset. The key when appending your dataset is to make sure the variable names match up exactly so that the values from the new dataset are appended into the proper columns in the existing dataset.

Just to demonstrate, I took the above dataset and broke it in half. The dataset at http://nw08.american.edu/~schaffne/incvote_nonsouth.dta includes all members of Congress in non-southern states. The dataset at http://nw08.american.edu/~schaffne/incvote_south.dta includes only members of Congress in the southern states. Start by opening the `incvote_nonsouth` dataset.

```
use http://nw08.american.edu/~schaffne/incvote_nonsouth.dta
```

Then, to add the Southern states to this existing dataset, simply use this command:

```
append using http://nw08.american.edu/~schaffne/incvote_south.dta
```

Now you should have all 435 congressional districts in a single dataset. You can append to a dataset as many times as is necessary.

Transformations

To estimate nonlinear relationships with OLS in Stata, you can usually just create a new variable that takes on the form you need to estimate.

For instance, if you want to create a variable that is the square of your independent variable, do the following:

```
gen newvar=oldvar*oldvar
```

This just tells Stata to create a variable that equals your old variable multiplied by itself. You can do the same thing by using this command:

```
gen newvar=oldvar^2
```

To create a new variable that is the log of your old variable, do the following:

```
gen newvar=ln(oldvar)
```

The square root of a variable can be generated by taking the half power of your existing variable:

```
gen newvar=oldvar^.5
```

Answering questions about Stata

When you need help in figuring out a particular command in Stata, you can turn to several sources. First, you can type **help** in the command window. Stata will then display a wide range of help options available in the program. Second, you can type **help** and then the name of a specific Stata command, if you are uncertain about how to use it. For example, the command **help sum** produces directions on how to use the **sum** command. Third, this website <http://www.ats.ucla.edu/stat/stata/notes3/default.htm> has a Stata tutorial available that you can watch in movie form. This site may also be helpful: <http://www.cpc.unc.edu/services/computer/presentations/statatutorial/>

Finally, if you still are unable to figure out a particular command, email me.