STATA Help

17.842
TA Jiyoon Kim

December 4, 2002

I hope this handout help you with using Stata. The first thing that you need to do for your paper is **obtaining and cleaning data set**. If you have done this part, you have done a half of your work. Data operation is painful thing, but good to do when your brain does not work for a serious thinking.... it is basically labor-intensive work. Anyway, I am going to explain some basic data handling commands in Stata with Michigan data that Steve gave us.

Let's say that you are interested in looking at the Michigan election data in 1960. And you want to see whether the income level determines the democratic win of a district. (I created income data for this. This is totally manipulated data set, so you shouldn't believe this in real ....) The data has several different variables based on each district. The candidate names were coded as two different variables, dname_60 and rname_60. Also, votes each candidate obtained were also coded as dvote_60 and rvote_60. These are good to look at the comparison, but not really helpful for statistical work. You need to create a new data set with one variable of candidate's name and another variable of party. In this case, we use "reshape" command.

Reshape command is used to transform your data from long to wide form or vice versa. In this case, I want this data to be long. In a newly transformed data set, I want candidate's name, party and whether or not the district was given to Democrats. Okay, let's start the work.

Variables that I would like to reshape are dvote_60, rvote_60, dname_60 and rname_60. In a new data set, I want these four variables to be arranged under vote, name and party variables. How can I dissect "d" and "r" from these?

When you ask Stata to reshape, it usually takes the rest of a newly assigned variable name as values. (I don't know how to explain it more accurately, so please work on this on your own...) In order to make reshaping easier, I renamed those variables as shown in do-file.

Before reshaping the data, I recognized that it would be eaiser for me to calculate the vote difference between two candidates in a district. So, I generated "diff" variable which substracts republican votes from democratic votes. So, if the difference is positive, the democratic candidate earned more votes, therefore the district is going to Democrats. (and of course, vice versa.) With these, I am ready to transform the data set.

After reshpaed it, I found a small problem. Party variable, which is string, is not very helpful as it stands. So, I convert this into numeric. Democrats are coded as 1 and Republicans are coded as 0. Please be careful with the error message of "type mismatch" when you are dealing with string variables. This error message pops up when you mix string variables with numeric ones. If your command uses string variable, you need to use quotation mark. Please take a look at the do file attached.

So, now party variable is coded either 1 or 0, but it is still string. Stata is still recognizing it as a letter of 1 or 0, not a number. Therefore, I use the command of "destring" to make this variable numeric. (If any one observation has a non-numeric code, e.g. -, ? etc., Stata would give you an error message. Then you need to check out your data set and make sure that everything is numeric transformable.)

Okay, now I create a dummy variable for a district where Democrats win. I named it as "dem_win" and if it wins, it obtains a value of 1. If not, it is coded as 0. We can use the variable of vote difference - remember? If the democratic party wins, it should have positive value for "diff". Use this fact and generate the variable.

One part of data set is now complete. But, I told you that I would like to create the data set with which I can examine the effect of income on democratic win of a district. I have a (bogust) data file for income of Michigan in 1960. (named mi_60_inc.dta) It has fipscode, party and income data. The next step that I need to do is to incorporate this data file with my election data file.

We can match for each observation based on its id, which is differentiating district. (the income data I created is also at the level of district.) Before you merge your data set, you need to sort both data. Since you want to match your two data sets based on your district level, which means id variable, you sort your data with id. The current data file you are in should be a master data file.

After you merge your file, you need to check out whether all observations merged appropriately. Stata create a new variable of "_merge" whenever it finishes merging. If it is coded 1, it means that the data observation exists in a master file, but not in a source file. 2 means the opposite. When an observation exists in both files, it is coded as 3. Therefore, the way to check the appropriateness of merge is to use "tab _merge" command. By typing this, you can see a table for a _merge variable. If your table for merge shows all 3, you are pretty safe.

Then, regression...

Simply use the command "reg dependent variable independent variable". Since I want to look at the influence of a district's income level on the probability of the democratic party's win, my dependent variable is whether or not the democratic party won a district. My independent variable is income.

The result of Stata output shows something. Coefficient of income is negative and p-value is pretty small - almost zero. The absolute value of t-score is very high, so we can say that as the income level of a district increase, the probability of the democratic party decreases. And this is statistically significant at 95% level.

I explained how to use "merge" command to incorporate two different data files. Merge is very useful when you want to add more variables to a master data. But, what if you want to add more observations? Then, you use "append" command. The command line is:

append using source file name.dta

One thing to remember - same as in merge case, if the variables in your master file and source file does not match completely, Stata would create a new variable and code cells which lacks datapoints as missing variable. For instance, if your source file (say it is data for 1970) has population variable but your master file (and this is for 1960) doesn't, after appending files, your new file would still have population variable, but all the data points for the year of 1960's population would be coded as missing variables.

Oh well, I tried to explain as much as possible, but still you have questions. If you do encounter any problem or question, please let me know...