

# Using Big Datasets in Stata

Josh Klugman and Min Gong  
Department of Sociology  
Indiana University

On the sociology LAN, Stata is configured to use 12 megabytes (MB) of memory, which means that the user can load up datasets that are no more than 12 MB in size.<sup>1</sup> There are two ways to get around this if you are working with a large dataset: increase Stata's memory usage or use DBMS/Copy to make a smaller dataset. Which method you use depends on the size of the dataset; the larger the dataset, the slower Stata is going to run. For Stata SE, we recommend that you set up your computer's mem to a number larger than the dataset you use, so you won't have problem opening the dataset or running analyses. If your dataset is very large, you may want to consider making your dataset smaller is using DBMS/Copy instead.

## Increasing Stata's Memory Usage

The easiest way to use a large dataset is to increase Stata's memory usage. This is simply done by using the `set mem` command.

```
set mem #
```

In this command, # is the number of kilobytes (KB) of memory that you want Stata to use. Choose a number that is somewhat larger than the size of the dataset you want to use (the reason is that you may need to add variables during your analyses, which will increase the size of the dataset). For example, if you want to use a 30 MB dataset, tell Stata to use 35 MB:

```
set mem 35000
```

You may abbreviate the number of MB you want Stata to use by dividing the number by 1,000 and adding a lower-case "m" next to the # of MBs.

```
set mem 35m
```

## Using DBMS/Copy

DBMS/Copy is a nice program that lets you copy datasets and translate datasets from one statistical package to another. If you want to create a smaller dataset that Stata can use, you can use DBMS/Copy to copy the dataset but to delete a certain number of variables or cases.

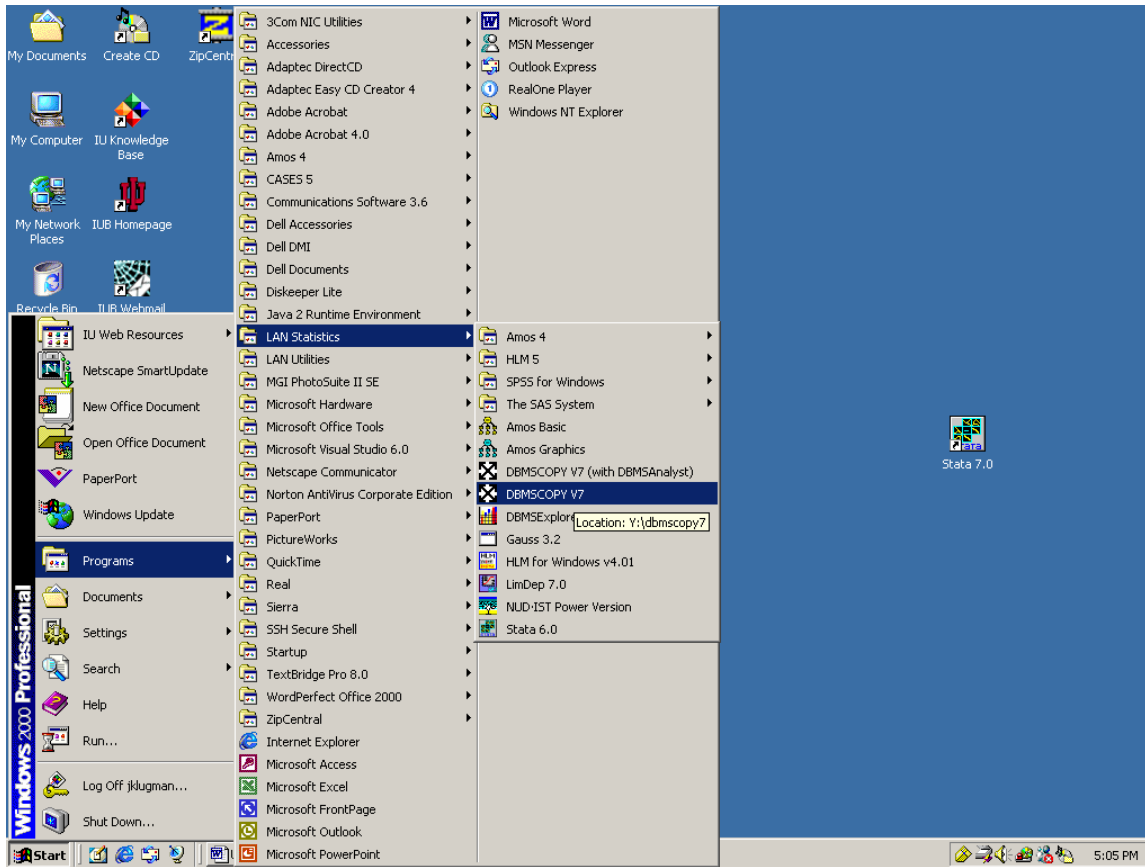
Here is how DBMS/Copy works:

Step 1: Load up DBMS/Copy. You can access the program through the Start menu, under Programs/LAN Statistics. Choose DBMS/Copy → DBMS/Copy 8. As far as we

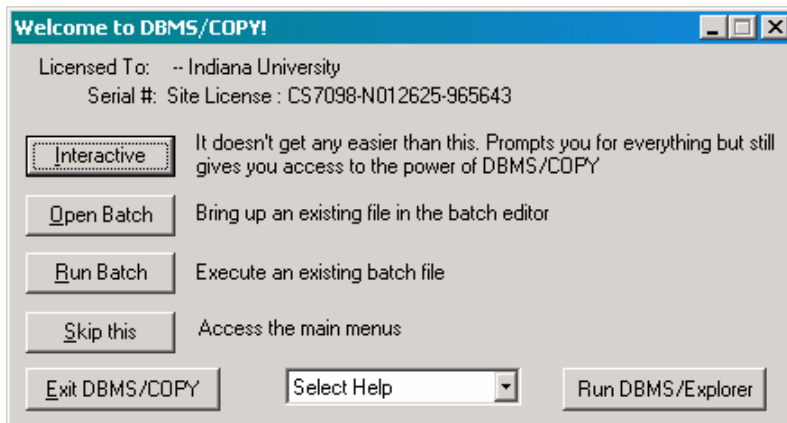
---

<sup>1</sup> You can tell how large a file is by clicking on it in a Windows window; Windows will tell you the size in kilobytes (KB) or megabytes—a megabyte is approximately 1,000 KB (e.g. 12,000 KB = 12 MB).

know, it doesn't matter if you choose either "DBMS/COPY V8 (with DBMSAnalyst)" or just "DBMS/COPY V8".<sup>2</sup>

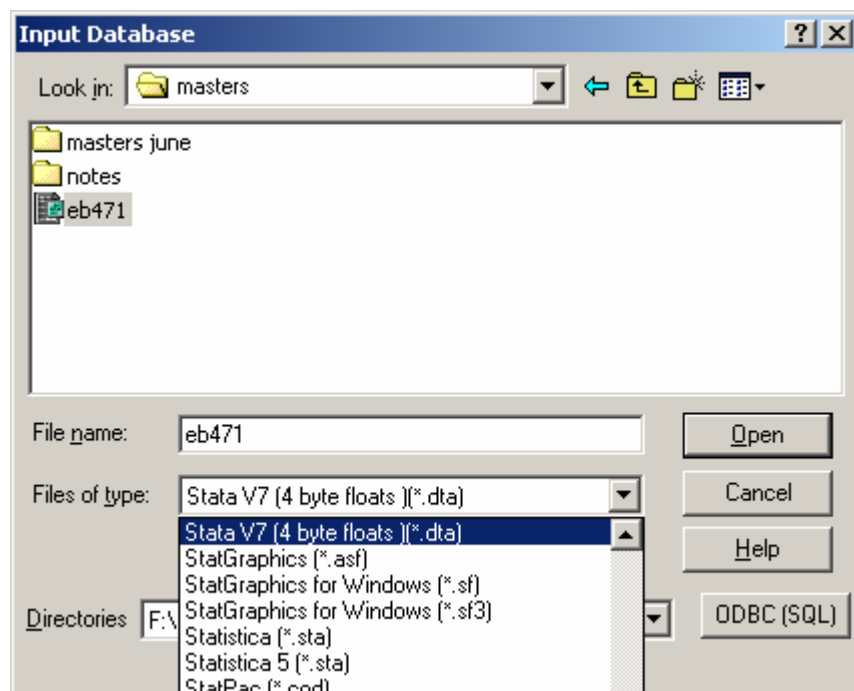


You should get the following menu choices:



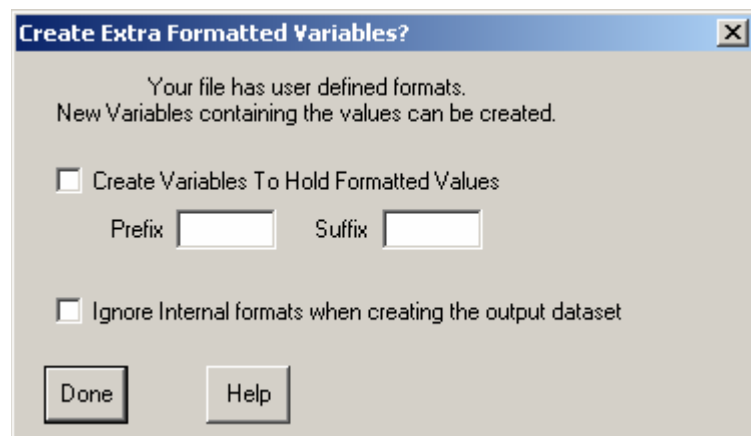
Click on the "Interactive" button. A window should pop up where you choose the "input" database.

<sup>2</sup> Following graphs show "DBMS/Copy V7." We have "DBMS/Copy V8" now.

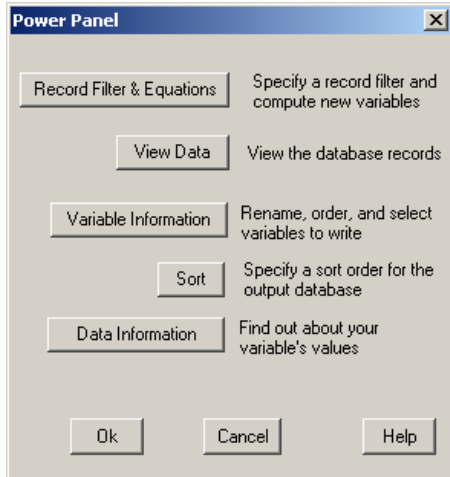


Since you're dealing with Stata datasets, make sure the program is looking for Stata SE datasets. If you want to transform SPSS or SAS data sets, look for "SPSS 12.0 for windows" or "SAS 9.1 for windows." (We use the Eurobarometer 47.1 dataset as an example.)

This window pops up. Ignore this window and press the "Done" button.



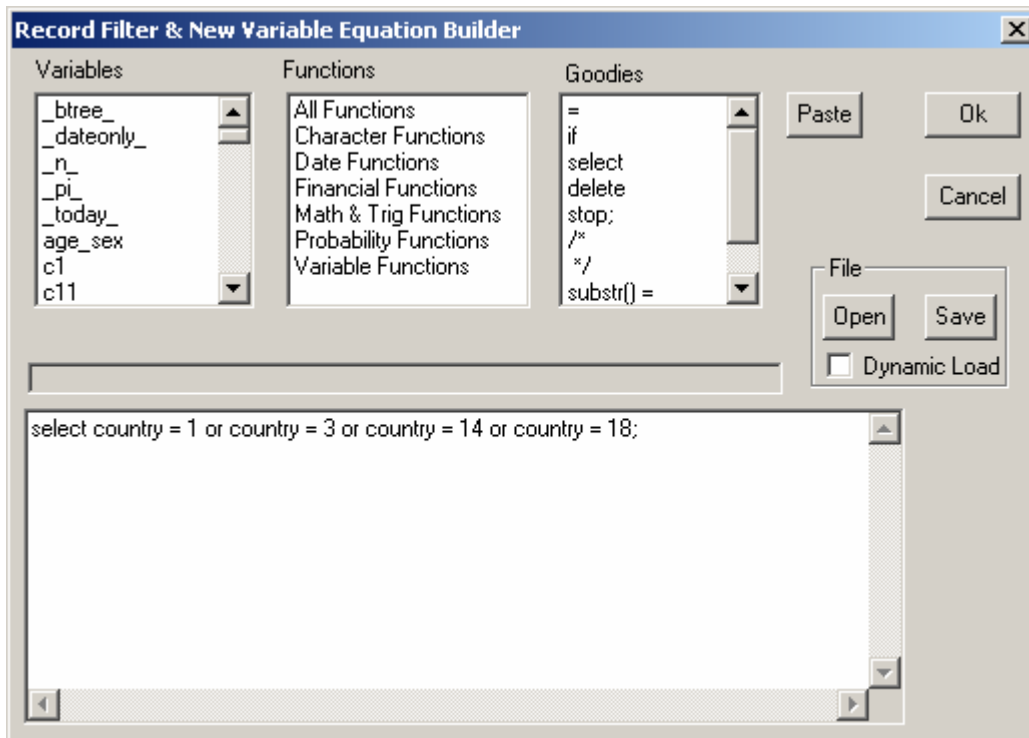
Then this window pops up:



Step 2: We will show you how to delete cases or delete variables next.

(1) To select cases, press “Record Filter & Equations;” (2) to delete variables, press “Variable Information.”

(1) Selecting Cases:



In this example, we are telling DBMS/Copy to select cases where the respondent is from Belgium, W. Germany, E. Germany, and Austria. This implies that we are deleting all the other cases that are not selected. You can either double click “select” in the box of

“Goodies” or type “select *variable=value*” in the last box. The program’s syntax is pretty simple; you can also “drop” cases as well. For more help, look at DBMS/Copy’s help file on equations (Y:\DBMSCOPY 8\program files\DataFlux\ dfPower Tools\8.0.hlp).

After you are done typing in your equation, click on “OK”

	Name	Rename	Pos	Keep	Type	Length	Dec	Label	Format	FormLen	FormDe
424	q4804		317	<input type="checkbox"/>	Float			d) Culture	fixed	1	
425	q4901		318	<input checked="" type="checkbox"/>	Float			1. In schools where there are too many children from these m	fixed	1	
426	q4902		319	<input checked="" type="checkbox"/>	Float			2. People from these minority groups get poorer housing, lar	fixed	1	
427	q4903		320	<input checked="" type="checkbox"/>	Float			3. People from these minority groups abuse the system of soc	fixed	1	
428	q4904		321	<input checked="" type="checkbox"/>	Float			4. Without people from these minority groups, (COUNTRY) woul	fixed	1	
429	q4905		322	<input checked="" type="checkbox"/>	Float			5. The authorities should make efforts to improve the situat	fixed	1	
430	q4906		323	<input checked="" type="checkbox"/>	Float			6. People from these minority groups are enriching the cultu	fixed	1	
431	q4907		324	<input checked="" type="checkbox"/>	Float			7. The religious practices of people from these minority gro	fixed	1	
432	q4908		325	<input checked="" type="checkbox"/>	Float			8. People from these minority groups pay more into our socia	fixed	1	
433	q4909		326	<input checked="" type="checkbox"/>	Float			9. Where schools make the necessary efforts, the education o	fixed	1	
434	q4910		327	<input checked="" type="checkbox"/>	Float			10. The presence of people from these minority groups is a c	fixed	1	
435	q4911		328	<input checked="" type="checkbox"/>	Float			11. People from these minority groups are given preferential	fixed	1	
436	q4912		329	<input checked="" type="checkbox"/>	Float			12. People from these minority groups do the jobs which othe	fixed	1	
437	q4913		330	<input checked="" type="checkbox"/>	Float			13. When hiring personnel, employers should only take accoun	fixed	1	
438	q4914		331	<input checked="" type="checkbox"/>	Float			14. People from these minority groups keep entire sections o	fixed	1	
439	q4915		332	<input checked="" type="checkbox"/>	Float			15. The presence of people from these minority groups increa	fixed	1	
440	q4916		333	<input checked="" type="checkbox"/>	Float			16. People from these minority groups are being discriminate	fixed	1	
441	q4917		334	<input checked="" type="checkbox"/>	Float			17. Discrimination in the job market on grounds of a person'	fixed	1	
442	q4a		23	<input type="checkbox"/>	Float			a) 1st	fixed	1	
443	q4b		24	<input type="checkbox"/>	Float			b) 2nd	fixed	1	
444	q50		335	<input type="checkbox"/>	Float			Again, speaking generally about people from minority groups	fixed	1	

In this window, we are specifying which variables to keep and which to delete. The first column “Name” gives the variable name; under the “Rename” column you can type in a new variable name for the output dataset; in the “Keep” column the user indicates what variables are to be retained for the output dataset; and the “Label” column gives the variable labels.

Importantly, whether you choose to “Drop Check” or “Keep Check” depends on whether you want to work with a small (“Keep Check”) or large (“Drop Check”) fraction of the variables in the dataset. Notice that “Keeping” and “Dropping” have the same effect— they trim the output database.

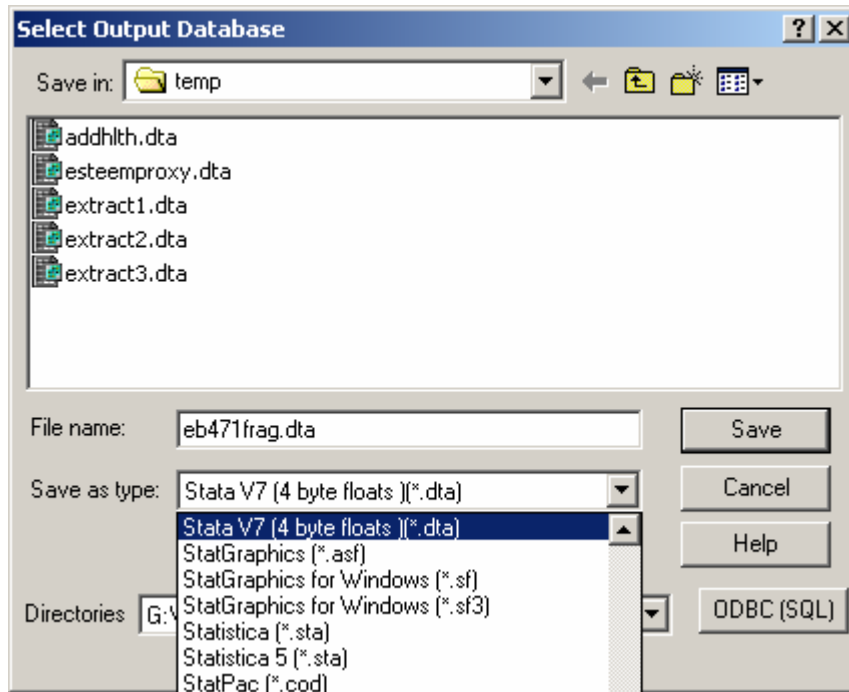
In this case, we are telling DBMS/Copy to keep the variables q4901-q4917, which are a series of statements about immigrants in Western European societies. You can tell this because those variables are checked.

## (2) Deleting Variables:

Alternatively, you can drop variables from the output dataset. In that case, you would specify “Drop Checked” in the button third from the left at the top (which currently says “Keep Checked”). The “Keep” column would then become a “Drop” column, and any variables you selected would be eliminated from the output dataset.

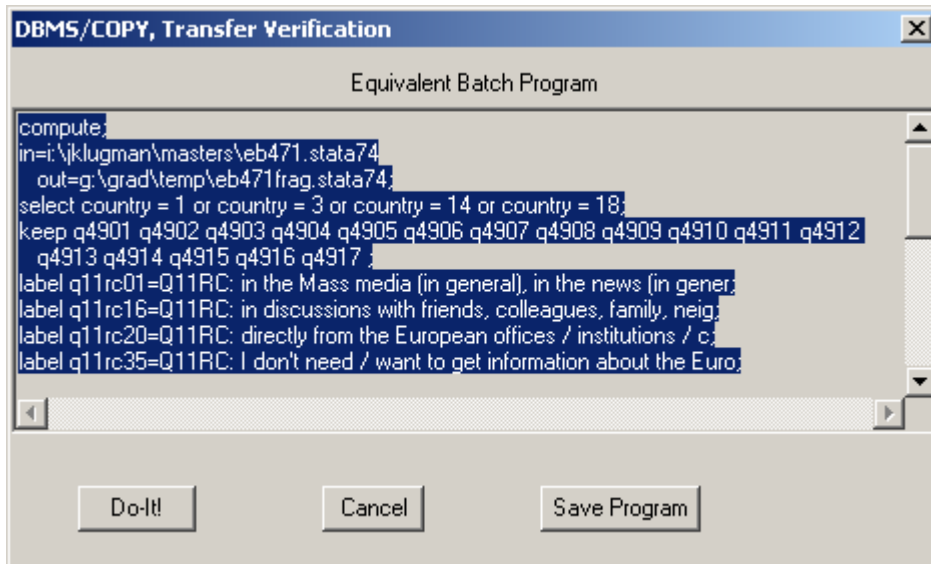
After you are done specifying which variables should be kept or dropped, click on the “OK” button.

Step 3: After selecting the variables, you are back at the Power Panel, shown above. Click the “OK” button.



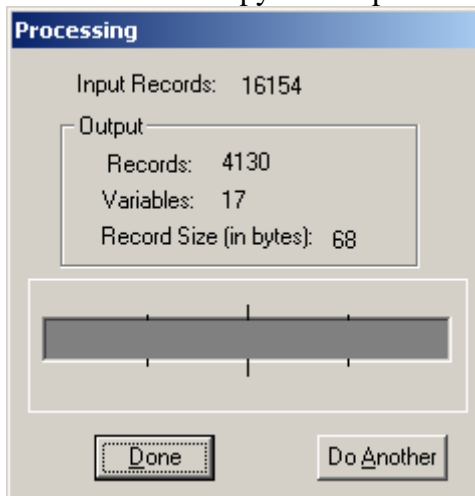
Here you specify what your output dataset is going to be. Make sure that you specify Stata SE dataset (You can also select SPSS 12.0 for windows or SAS 9.1 for windows if you prefer to use these two statistical packages). After you finish this procedure, click the “Save” button.

You will be taken to this window:



This window gives you the DBMS/Copy’s syntax for all of your specifications. You can save your program by pushing the “Save Program” button—this is a good idea to keep track of your data manipulations. For now, click the “Do-It!” button.

DBMS/Copy is now processing your commands.



When it’s done, you’re ready to open the Stata/SPSS/SAS file that contains the variables you want. Congratulations!!