Subject: Merging data files in Stata
Posted by DHS user on Thu, 04 Apr 2013 18:26:13 GMT
View Forum Message <> Reply to Message

I am currently working on the 1998, 2003, and 2008 Philippines DHS datasets in Stata. I am hoping to analyze the data for trends over the 3 different years, and would like to merge the datasets across years. Each year appears to have different PSUs, but consistently stratified by the 17 regions in the Philippines. In order to merge the datasets, is there a method to handle the different sampling and strata across the different years of the survey?

Thank you very much in advance for your assistance.

Subject: Re: Merging data files in Stata
Posted by Bridgette-DHS on Thu, 04 Apr 2013 18:28:54 GMT
View Forum Message <> Reply to Message

Here is a response from one of our DHS Stata experts Tom Pullum, that should answer your question.

In Stata, this would be done with an append command, rather than merge. I suggest that you "use" (open) the 2008 file, then "append using" the 2003 file, "append using" the 1998 file, and then save with a different file name.

I have encountered situations in which starting with the earlier file and then appending later files will not work; it's safer to start with a later file and then append an earlier file. I think Stata likes to start with the file that is larger, in terms of the number of variables, and with DHS that would usually mean starting with a later file.

You are correct that it is necessary to re-number the clusters. There are two ways to do this. One is to add some larger number to the codes. For example, you could have the original id numbers in the first survey, but in the second survey, add 1000 to the id numbers and in the third survey add 2000 to the id numbers. An alternative would be to use the "egen group" command. For example, if you had a line "egen v001r=group(v000 v001), it would completely renumber the clusters, consecutively, from 1 to the total number of clusters in the three surveys. This is elegant but will make it just a little harder to figure out what was the original number of the cluster if you ever needed to do that.

The strata may be the same across surveys, and in that case you would want to just make sure they have the same id numbers. For example, Metro Manila should have the same number in all three surveys.

The weights should be ok. Sometimes surveys from several countries are pooled, and then the weights may need to be changed by a different multiplier for each survey.

We will often combine multiple surveys into a single file, but you have to be careful when you do this. For example, I would advise against treating them as a single survey and calculating the

mean of some variable in all three surveys. Looking at differences or changes between surveys is fine.

Bridgette-DHS

Subject: Re: Merging data files in Stata
Posted by emijo on Thu, 18 Apr 2013 13:41:01 GMT
View Forum Message <> Reply to Message

I am interested doing a pooled analysis across different countries (using only child recode files), and wondering about how to deal with weights in the pooled analysis? The previous message mentioned a multiplier. Could you explain more?

We are using Stata and running a multilievel logistic regression model with individuals nested within clusters (country included as random coefficient). Do I need to handle weights since the multilevel model should handle the clustering effect?

Subject: Re: Merging data files in Stata
Posted by Reduced-For(u)m on Fri, 19 Apr 2013 05:12:01 GMT
View Forum Message <> Reply to Message

http://userforum.measuredhs.com/index.php?t=tree&th=54&a mp;goto=82&#msg_82

This might help. There is a note from one of the DHS statisticians that describes how to re-scale your weights. I think it applies to your problem.

Subject: Re: Merging data files in Stata
Posted by ahmed89o on Sat, 09 Nov 2013 17:06:39 GMT
View Forum Message <> Reply to Message

I would like to know why you would advise against treating multiple surveys as one survey?

Subject: Re: Merging data files in Stata
Posted by Bridgette-DHS on Fri, 22 Nov 2013 18:38:01 GMT
View Forum Message <> Reply to Message

Here is a response from DHS Stata expert Tom Pullum:

It can be useful to pool surveys and/or ignore weights for exploratory work or model building. However, you have to be careful in a final analysis. The general goal that I would have is to produce rates, means, proportions, coefficients, etc. that are unbiased estimates of what you would get if you had complete data from a well-defined population at a specific point in time.

If you pool multiple surveys from a single country, the time points or time intervals should be taken into account, at least as covariates.

If you pool surveys from multiple countries, then the pooled sample probably cannot be described as a sample of a well-defined population. The dates of data collection will differ, but beyond that there will be many historical, cultural, ethnic, etc. differences. You probably will not have data from ALL the countries in a recognized region, e.g. "East Africa".

If you pool, consider using a multi-level model, and allow for variation in the effects, not just in the intercepts.

If you CAN justify pooling, then you have to re-weight somehow, at least to compensate for the arbitrary variation in the sizes of the samples. The two main options would be to re-weight so that the total weight is the same for each country or survey (as with the U.S. Senate) or the total weight for each survey is proportional to the population of the country, e.g. the estimated number of women 15-49 at the time of the survey (as with the U.S. House of Representatives.

Subject: Re: Merging data files in Stata
Posted by ahmed89o on Sun, 04 May 2014 16:45:06 GMT
View Forum Message <> Reply to Message

Dear DHS,
In your last response, you advise considering multilevel model. I am aiming to pool survey from the several rounds of survey of the same country. What is wrong with using OLS in such case? I prefer simplicity than fancy modeling in addressing research objectives.

Subject: Re: Merging data files in Stata
Posted by Bridgette-DHS on Tue, 06 May 2014 13:33:52 GMT
View Forum Message <> Reply to Message

Here are comments from Tom Pullum:

Dear Ahmed--I do not like unnecessary complexity either, but the choice of the model depends on many things, including the level of measurement of the outcome variable (for example, is it an interval-level variable, or binary, or a count), and the sampling design (for example, is it a multi-stage cluster sample and were weights used). OLS normally refers to an interval-level outcome and a simple random sample (srs)--that is, the simplest possible situation. For most DHS users the outcome is NOT interval level, and the sample is never srs.

Please say a little more about what you propose to do after you pool several surveys from the same country? What is your research question? What is your outcome variable?

Subject: Re: Merging data files in Stata
Posted by ahmed89o on Tue, 27 May 2014 17:25:32 GMT
View Forum Message <> Reply to Message

Am doing trend analysis to explore the inequality in access to maternal health care, I am pooling datasets from 4 rounds of surveys of same country and doing interaction terms of the year survey dummy and variable of interest over the three periods (first round is the base year) using Linear Probablity model not even probit or logit. my main outcomes are the interaction terms.

Subject: Re: Merging data files in Stata
Posted by ahmed89o on Tue, 27 May 2014 17:31:25 GMT
View Forum Message <> Reply to Message

and forget to mention my dependent variables is binary variable whether women received prenatel care or not, whether skilled attendant assist delivery and has child dies in first year of life. so basically assessing inequalities related MDG using linear probablity model. Many thanks for your interest

Subject: Re: Merging data files in Stata
Posted by Bridgette-DHS on Wed, 28 May 2014 13:21:46 GMT
View Forum Message <> Reply to Message

Following is another response from Tom Pullum:

With four surveys from the same country, my suggestion of multi-level analysis would not apply.

Your outcome or dependent variable is whether the child died before the first birthday (0=no, 1=yes). Some predictors of interest--birth attendants and place of delivery--are usually available for all children born in the past five years. Others, such as antenatal care, are usually only available for the youngest child born in the past five years. You have to be careful because of the fact that the subgroup of youngest children usually has a better chance of surviving than the other children born in the past five years (the youngest child tends to have a longer preceding birth interval). You also have to take into account that children born less than a year before the survey have not had full exposure to the risk of dying before the first birthday, i.e. are censored.

You should not do just a linear probability model. This is never recommended for a binary outcome, although the fact is that if the fitted probabilities are in the range from .3 to .7, approximately, then a model that is linear in the probabilities will agree pretty well with a model that is linear in the log odds (or logits). Since your probabilities are small, you should do a logit (or

probit) model or a log probability model, which is basically equivalent to a hazard model or survival model. I assume you are using a statistical package, and it's easy to do it the right way.

Last year DHS did a Further Analysis report on neonatal mortality in Rwanda using the 2000, 2005, and 2010 surveys: http://dhsprogram.com/pubs/pdf/FA88/FA88.pdf. It may help.

Subject: Re: Merging data files in Stata
Posted by ahmed89o on Fri, 30 May 2014 13:51:23 GMT
View Forum Message <> Reply to Message

Many thanks for this. I have some comments here.

I am using Egypt DHS from 1995 to 2008 so if antenatal care is only for youngest child born in the last five years. How am getting the same number of observations for m2a and m3a? Am assuming that m2a must be less in such case but they are not?

"the subgroup of youngest children usually has a better chance of surviving than the other children born in the past five years (the youngest child tends to have a longer preceding birth interval)." can you please clarify a bit more.

LPM vs probit, my reference is Wooldridge (introductory to economitrics) he still recommending with some adjustments like Weight least square. I economics student not public health, as you may know that we are more into marginal effects than odds ratio. I found it complicated to calculate marginal effects for interaction terms either in stata or SPSS (I think SPSS cannot do it all). If you can advise how to get marginal effect correctly in stata I will be grateful.

Subject: Re: Merging data files in Stata
Posted by ahmed89o on Fri, 30 May 2014 13:53:06 GMT
View Forum Message <> Reply to Message

Many thanks for this. I have some comments here.

I am using Egypt DHS from 1995 to 2008 so if antenatal care is only for youngest child born in the last five years. How am getting the same number of observations for m2a and m3a? Am assuming that m2a must be less in such case but they are not?

"the subgroup of youngest children usually has a better chance of surviving than the other children born in the past five years (the youngest child tends to have a longer preceding birth interval)." can you please clarify a bit more.

LPM vs probit, my reference is Wooldridge (introductory to economitrics) he still recommending with some adjustments like Weight least square. I economics student not public health, as you may know that we are more into marginal effects than odds ratio. I found it complicated to calculate marginal effects for interaction terms either in stata or SPSS (I think SPSS cannot do it

all). If you can advise how to get marginal effect correctly in stata I will be grateful.