Subject: When weights are not supported
Posted by Yohannes on Tue, 04 Nov 2014 01:06:31 GMT
View Forum Message <> Reply to Message

I have DHS data for over 30 countries which I plan to use for a multi-country analysis using a STATA platform. I have already de-normalised the weights (following DHS suggestion) but the model I want to fit, multi level probit/logit model, does not support weights. The question I have is what do you do when the platforms you use for analysis do not support weight, but you know that ignoring weights is not an option because sample sizes vary enormously between countries, and in some cases countries that have half the size population than the larger countries have two or three times larger samples.

Subject: Re: When weights are not supported
Posted by Bridgette-DHS on Wed, 05 Nov 2014 13:20:58 GMT
View Forum Message <> Reply to Message

Following is a response from senior DHS Specialist, Tom Pullum.

I really don't think there is an easy answer to this question. If the software doesn't allow you to use weights--and there are several complex statistical procedures for which a weighted version has not yet been written--then your results for each country, let alone all countries combined, will be biased toward the over-sampled sub-populations.

I recommend that you include fixed effects for country or survey, and fixed or random effects for all the strata, in your models. Most of the variation in the weights (v005 and the survey-specific factors that you could calculate) will be explained by those effects. Perhaps other users have different solutions to this problem.

Subject: Re: When weights are not supported
Posted by Yohannes on Wed, 05 Nov 2014 23:54:49 GMT
View Forum Message <> Reply to Message

Thank you for the suggestion. Just a quick follow up question. Actually, I was going to fit random effects at country and strata levels but for a different reason - i.e. to capture unobserved strata and country level differences in my response variable -, so looking form the comment does this now mean that the random effects will not measure these effects, but rather show only the 'variation in weights'? As part of my random effect model I was also going to include some explanatory variables on the variance function (such as GDP for the country level random effect that are not necessarily related to sample size), and given the comments on the links between random effects and weights, not sure how these things play out.

Subject: Re: When weights are not supported
Posted by Reduced-For(u)m on Thu, 06 Nov 2014 05:09:13 GMT

Yohannes,

I don't really understand the context of your research well enough to guarantee that any of these suggestions will be helpful, but here are some options (and some comments on the fixed/random effects question).

1 - A very simple solution would be to use a "linear probability model" instead of a logit/probit - meaning an OLS regression on a 0/1 outcome. In big samples where you are interested in the effect of some particular covariate, this should give you a result very similar to the probit (unless you have someth really rare outcome, in which case maybe a Poisson regression). These are easily weighted and can deal with survey effects in any way you choose.

2 - Dis-aggregation: You could estimate each country separately using any method you want, and then weight the coefficient estimates by population size to get a single, overall estimate - or just present the distribution of point estimates. This has a bit of a different interpretation than doing it all at once, and you are implicitly allowing each country to have its own (totally unconstrained relative to other countries) effect of each covariate. It will cost you power (efficiency). There may be a Bayesian hierarchical way to estimate this too all at once, but I don't know it, and I don't think Stata would do it.

3 - Country Fixed Effects. I believe the recommendation for country level FEs had to do with establishing a commonality among countries in "levels". That is - if you use random effects, your model will still identify the co-efficient of interest using both within-country variation and across-country variation. So if the levels of your Y or X variables are majorly different across countries (compared to within-country) you will be estimating you coefficient mostly on differences between countries that may or may not be reasonably comparable. Using country fixed-effects de-means everything so that only the difference from the country level mean will identify the coefficient. In certain cases, this would make the weighting problem less severe (suppose all your observations were from very high or very low level country's for your covariates and outcome - the across-country variation would be terrible, driven by lots of observations in the tails, but the within-country might still be OK).

4 - Depending on what your covariate of interest is, you will likely want to estimate your standard errors in ways that are far more conservative than those recommended for using one survey by the DHS. This depends a great deal on the particular empirical question you are asking, in particular on where the variation in your right-hand-side is (for instance, is it a response to some question in the DHS or some other data you are merging in). The paper linked below gives a good, if somewhat technical, overview of clustered standard errors and the problems you might face. I can offer better suggestions here if I understood your context a little better.

http://cameron.econ.ucdavis.edu/research/Cameron_Miller_Clus ter_Robust_October152013.pdf

Subject: Re: When weights are not supported
Posted by Yohannes on Thu, 06 Nov 2014 23:39:41 GMT

Thank you for the attention to my question and for the suggestions. Thank you also for the literature and the offer of assistance. Basically, what I am trying to do is look at the covariates of child well being in low and middle income countries in a multi-country context using primarily DHS data, but also supplemented by relevant national level indicators obtained from international databases, namely the World Bank. One of these child well being measures which I am trying to model is primary school attendance which is assumed to be a dichotomous outcome but can also take a count data form if level is taken in to account, and can be fitted using inflated or hurdle variants of Poisson or negative binomial models. I am also assuming that the response variable is affected by four set of factors: (1) child level characteristics (such as sex, age birth order etc); (2) household level characteristics (such as religion, parental education, and wealth status (I will come back to this to show how it is handled); (3) area level characteristics derived from the DHS data itself (such as urban rural residence, and other characteristics derived from the mean values for the strata in which the child lives); and (4) national level covariates (such as GDP, health expenditure per capita, external aid, indices of governance etc..) obtained from the World Bank. Given these, I set out to test two set of models, and in the first wanted to put all the four level variables in the main model and then add two random terms (without covariates) at country and strata level. In the second model, I wanted to keep only the child level and household level factors in the main model, and include the strata level and national level variables in the respective variance functions. For the wealth index variable, given that the DHS generated index is both country and survey specific, I went on and re-calculated the index using the data for all countries combined, and reclassified households based on this new index rather than the original measure developed by the DHS.

So coming back to point 1 of the suggestion:

The data I am using for the study has over 200000 observations (unweighted) and the variable of interest, child schooling, is also not a rare event as such, so I can revert to a "linear probability model" easily. But my challenge is the available multilevel commands (such as the one on STATA which I use extensively for my research) do not support weights for any type of response variable specification. I am yet to go through the attached publication perhaps it may have a way to resolve my problem (??)

Yes, I can also do my analysis on country by country basis but as you noted that will change the research question. Primarily, I will not be able to capture between country effects. Besides, I am not exactly sure how I will be able to capture country level indicators in such analysis.

On the fixed random effects: The countries in the analysis vary widely both in terms of the response variable and explanatory variables as well as in DHS sample sizes. And given that my interest is also to look at both within and between country differences I was thinking that a random effect model that has the ability to correct for sample weights may be the best way to go. And given that the multi-level commands do not allow this, and that the fixed effect approach only partially addresses what I would like to do I am somehow torn between approaches neither of which appear to be ideal.

Subject: Re: When weights are not supported
Posted by Reduced-For(u)m on Fri, 07 Nov 2014 21:22:17 GMT
View Forum Message <> Reply to Message

Well... we are starting to get pretty far from area of expertise now (which is more about causal inference on a single RHS variable and not multi-level modeling).

That said - I wonder if the new class of -sem- (structural equation modeling) commands in Stata might help. They support the "svy" prefix with weights and strata, and I think you can model what you want using those commands, but like I said, getting pretty far from my expertise here.

http://www.stata.com/manuals13/sem.pdf [maybe start with intro 10 -- Fitting models with survey data (sem only)]

Let me know if these work - this problem has come up in enough contexts now that I would like to have an answer for people.

Oh - and out of curiosity, are you matching your country level covariates to the survey timing (so GDP in survey year) or to the birth or age timing (so relevant to when you were born/how old you are). For schooling, I'd say there is hazard of drop-out every year, and if a kid dropped out two years ago, then current condition may not mean much as a predictor. I'm just sort of curious as to how people are thinking about these things. Thanks! Good luck.

Subject: Re: When weights are not supported
Posted by Yohannes on Sat, 08 Nov 2014 12:24:04 GMT
View Forum Message <> Reply to Message

Thnank you for the suggestion. I will look in to the SEM routine. Currently I am also exploring two user written commands (cmp and gllamm) and hopefully will get something useful between these routines.

Regarding the macro level variables I matched with the DHS data, what I used is the latter procedure (i.e. macro level variables were matched with birth year rather than year of survey) but I fully agree that this approach may have some issues for children who did not have a 'linear' progression with their schooling experience (either because they did not start school at the 'right' age and/or dropped out of school Were re-entrants). I think with some additional information on the timing of school events it may be possible to align events and covariates better than I was able to do or be able to use such information to generate lag variables.

Subject: Re: When weights are not supported
Posted by Bridgette-DHS on Tue, 11 Nov 2014 17:26:54 GMT
View Forum Message <> Reply to Message

Another response from Tom Pullum:

I agree with what Professor Cameron said. The fixed and random effects should partially remove the influence of the weights, as well as adjust for unmeasured sources of variation.

I suggest that you first apply, to the same data, a simpler statistical model--single-level rather than multi-level-- that allows you to use weights. Apply that model with weights and then without weights to see how robust the estimates are--that is, to see whether there is a dramatic difference between the weighted and unweighted estimates when the fixed/random effects are included. (The fixed/random effects would be included in both the weighted and unweighted versions.) If there IS a dramatic difference, then maybe you should wait until the multi-level model has been revised to include weights.

There is another way to check robustness. I call this the construction of "fake" weights. First adjust v005 in the way you would like, for example so that the weighted number of cases in each survey is the same or is proportional to the population of the country, etc. Then think of v005 as an fweight rather than a pweight. You could remove a factor of 10,000 so that the arbitrary inflation factor is 100 rather than 1,000,000.

Say that the adjusted v005, still with the factor of 1,000,000, is called v005r. Then try these lines:

First run the model without any weights.

Then:

replace v005r=round(v005r/10000)
expand v005r

and then re-run the model on these data (without any weights).

Neither model will actually use weights, but the second data set has been artificially weighted through the expand command.

There is a substantial issue here--the expanded data file will be about 100 times bigger than the original file and probably too big. To make this strategy feasible you could do some sub sampling. OR you could work from a sample of the original data. The main objective would be to get a sense of how much distortion there will be in the estimates if you use fixed/random effects for countries and strata and do not weight.

Subject: Re: When weights are not supported
Posted by Reduced-For(u)m on Wed, 12 Nov 2014 01:19:58 GMT
View Forum Message <> Reply to Message

I almost pitched the same weighting work-around idea - just make your dataset sample size proportional to population by expanding the number of observations (and within-survey, you

expand relative to the probability weights). I would just add that, if you use the "cluster" command and cluster at a level above the individual, then you should get proper inference because the model will "know" that there is no added variation from the extra observations.*

*Note: a great way to convince yourself of this is to generate some fake data, add an id number, run a regression, get the standard errors, and then replicate each observation 100 times (expand) and "cluster" on the id variable. You'll get back the original standard errors (while OLS on the expanded sample will produce SEs that are far too small). You could also do this using an original DHS dataset.

Subject: Re: When weights are not supported
Posted by Yohannes on Wed, 12 Nov 2014 05:36:18 GMT
View Forum Message <> Reply to Message

Thank you both for the suggestion on using "replications" as a possible way to get around 'weights' for statistical routines that do not support weights. I had initially entertained similar thoughts but concern with highly significant coefficients that 'inflated' data may lead to have made me refrain from going ahead. I am glad to hear that it can still be used (with some adjustment) as a 'valid' potential approach to inject weights into formal regressions analysis. I will do a few simulations on a sample data and see how it goes. Thanks again for the thoughts and suggestions!

Subject: Re: When weights are not supported
Posted by lhuri on Sun, 21 Dec 2014 10:43:04 GMT
View Forum Message <> Reply to Message

Dear all,

I'd like to know which variables in special DHS (I'm working on Indonesia Adolescent Reproductive Health) to use for weighting purpose.
Because I notice that they are different from those of standard DHS. I find "v0005" equal to "aweight" in special DHS. But I'm not sure the equal variables for PSU and strata. Does anyone here ever use special DHS dataset and know how to properly use sample weights?
Thank you,
Lhuri

Subject: Re: When weights are not supported
Posted by Trevor-DHS on Fri, 26 Dec 2014 16:14:54 GMT
View Forum Message <> Reply to Message

The Indonesia Adult Reproductive Health Survey is in a raw data format and not a recode format, and so does not have the same variable names as the recode files. For this file, the following are the variables to use:

weight - aweight/1000000
PSU - acluster
strata - a combination of aprov (Province) and atype (Urban/Rural), e.g.
egen strata = group(aprov atype)

Subject: Re: When weights are not supported
Posted by lhuri on Sat, 03 Jan 2015 15:54:08 GMT
View Forum Message <> Reply to Message

Thank you Trevor. Really appreciate it.
Happy new year.
