Subject: How do I account for clustering within families?
Posted by vega25 on Fri, 11 Apr 2014 04:01:33 GMT
View Forum Message <> Reply to Message

Hello,

The children's dataset in DHS files may have some of the records for multiple children within families - for instance if the dataset has data for all children born in the five years preceding the survey and some of the female respondents had more than one children during that time. Thus many children in the dataset would be siblings.

When this is the case, how does one account for the fact that there are some siblings in the data and so they have the same set of background/household/parental variables? Is adding "cluster(caseid)" as an option at the end of the regression syntax in Stata a valid way of doing so? Would this be okay? Or just asking for robust standard errors? Alternately, is a family-fixed effects the best way of doing this? My concern with family-fixed effects is the loss of sample size.

Thank you for any advice that people may have.

Subject: Re: How do I account for clustering within families?
Posted by user-rhs on Fri, 11 Apr 2014 18:15:23 GMT
View Forum Message <> Reply to Message

It depends on what "family" is. You can cluster around caseid or hhid (concatenate v001 and v002). If you specify caseid as the cluster, I'm not sure how children in the household that are not the biological children of the woman will be counted. Using hhid would ensure that all children in the household are clustered in that household. I'm trying to think of certain contexts where one would be more appropriate than the other. I guess it all depends on your question. For example, if there is a large number of children who were orphaned who now live with their relatives, it's probably best to use HHID. If it's a largely polygynous society, then maybe caseid would be more appropriate.

HTH,
RHS

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Fri, 11 Apr 2014 18:25:03 GMT
View Forum Message <> Reply to Message

I would say that for most analyses (anything I can think of) you would want to cluster at a more aggregated level than the household. The survey design itself requires accounting for cluster sampling, so if you cluster at the PSU (primary sampling unit) level, that will subsume household and take care of both the study design and the within-household problem.

Subject: Re: How do I account for clustering within families?
Posted by vega25 on Sat, 12 Apr 2014 05:26:17 GMT
View Forum Message <> Reply to Message

Yes I see what you're saying about wanting to cluster at a level higher than the household, perhaps the cluster. That is a good suggestion and cluster-fixed effects is very much a strategy I'm considering.

But what I'm trying to do before that or in lieu of that to begin with, is essentially "control" for the fact that some of the children in the sample share the same household characteristics and are siblings. So in a manner of speaking, the total sample of children whose health status is my dependent variable, actually belong only to a much smaller number of households and therefore to a smaller number of distinct maternal or household characteristics.

In that case, is this something to do in addition to cluster-fixed effects?

Subject: Re: How do I account for clustering within families?
Posted by user-rhs on Sat, 12 Apr 2014 13:41:00 GMT
View Forum Message <> Reply to Message

Dear Vega,
If you are so inclined, you can run a mixed model where household is nested within sampling cluster. Depending on what your outcome is, you can try -xtmixed- for multilevel mixed effects linear regression or -xtmelogit- for multilevel mixed effects logistic regression. For other outcomes, use -gllamm-, which I should warn is a beast* to run, and will probably take at least half a day to complete the procedure (and sometimes, after 3 days of running, you still don't get convergence.....the point is, do NOT run -gllamm- if you're pressed for time). As I do not have Stata 13, I can't tell you whether there is a less clunky procedure that is equivalent to -gllamm- built into Stata 13. (Reduced-for(u)m, if you have Stata 13, does such a procedure exist?)

Sample syntax
xtmixed outcome covariate1 covariate2 || cluster: || hhid:

If you suspect that having random slopes makes sense for some variables you can list the variables for which you want random slopes estimated after the colons, e.g.
xtmixed outcome covariate1 covariate2 || cluster: covariate3 || hhid:

****Note, this is usually an exception rather than norm and controlling for clustering via the cluster and household fixed effects is sufficient for most purposes.

For more information, visit the documentation for -xtmixed-, -xtmelogit-, -gllamm-, and the Bristol University Centre for Multilevel Modelling:
- http://www.stata.com/help.cgi?xtmixed

- http://www.stata.com/help.cgi?xtmelogit
- http://www.gllamm.org/
- http://www.bristol.ac.uk/cmm/learning/online-course/index.ht ml

HTH,
RHS

*I have only the utmost respect for Prof. Sophia Rabe-Hesketh who wrote -gllamm- (and is co-author of the Stata Press books on multilevel modeling). GLLAMM is very flexible and powerful but takes a long time to run. The flexibility also means you have to read the documentation properly to make sure you won't get error warnings for failing to specify required "options" for the specific model you are trying to fit!

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Sat, 12 Apr 2014 21:01:09 GMT
View Forum Message <> Reply to Message

I think there is a confusion here that stems from different disciplines speaking differently, so let me clarify something.

By "clustering" I mean choosing a specification for the variance/covariance matrix of error terms that accounts for within-cluster heteroskedasticity and serial-correlation. In STATA terms, I mean something like "reg Y X, cluster(clustervar)". This is something that relates to getting your standard errors right, but will not in any way affect point estimates.

I think you're question about "controlling" for within-household effects has to do with point estimates. In that case, you may (and may not) want to include household fixed effects (or dummy variables for each household). This would mean that your estimate is based off within-household differences. It would also limit your effective sample to households with multiple children. When I say "control for household characteristics", I'm usually referring to this, which is about getting the right identifying variation for your model.

But what I was talking about before was "accounting for within-household and within-cluster similarities" in your standard errors (your estimated precision). In that case, you want to cluster at the PSU level because you get all of the benefits of clustering at HH level and the benefits of accounting for the sampling design (though not stratification, which could technically shrink your SEs back down a bit).

What the multi-level models do is, depending on what you choose, something closer to what I call "Random Effects". That is, the V/C matrix on the error terms is parametric in some way, whereas it is "nonparametric" in the clustering case. This gets technical real fast. So let me just repeat the main point:

Household fixed-effects would deal with things like selection (some parents are good, some are bad) or other omitted variable bias problems. Clustering will get the standard errors right. Two distinct problems.

That help? I can try again if it doesn't.

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Sat, 12 Apr 2014 21:12:05 GMT
View Forum Message <> Reply to Message

I haven't used the mixed-models in Stata 13, but I do have it! Here is the documentation if anyone is interested. The basic command is "mixed" and there's "meglm", "melogit", etc. too.

Basic: http://www.stata.com/manuals13/meme.pdf#memeRemarksandexampl es
More: http://www.stata.com/manuals13/me.pdf

There's also this new "gsem" command, and I see this regarding its relative speed:

Note: gllamm users will be especially interested in gsem. There is a lot of overlap in the models that gllamm and gsem can fit. Where there is overlap, gsem is faster. gsem is at least four times faster, usually it is 10 to 100 times faster, and there are examples where gsem is up to 1,000 times faster than gllamm.

http://www.timberlakeconsulting.com/Stata/?id=504

... I can't be much more helpful than that. I think back in the day I once used "xtmixed" to fit one of these to a dataset in the low thousands of obs, and it went really fast. But I'd bet it really depends on how much structure you are putting in, what kinds of prior distributions you may/not be fitting, and what particular estimation method you want.

If only Nick Cox were on the DHS Forum, we'd know what to do. Short of that, I'd say, if anyone ever has the need, drop a question on the Statalist, and then report back what Nick has to say about relative speed of the various Stata generalized linear model commands.

Subject: Re: How do I account for clustering within families?
Posted by Liz-DHS on Sun, 13 Apr 2014 22:48:22 GMT

View Forum Message <> Reply to Message

Dear User,
Here is a response from one of our experts, Dr. Tom Pullum:
Yes, there is definitely clustering at the level of the family or household or mother for outcomes such as child survival, place of delivery, childhood illness and treatment for illness, etc. The standard practice at DHS has been just to include clustering at the level of the PSU, v001 (=v021). We do that either with the option cluster(v001) or (better) with svyset, followed by svy:. The standard errors will then be robust for that level of clustering. Up through Stata 12, as I understand it, only one level of clustering can be used, and for us that would be v001. We are about to upgrade to Stata 13, and as I understand it we will be able to add a second level of clustering, which will be the household (v002).

A problem which can arise with v001, and will be much more common with v002, is insufficient density at that level. It will be necessary to specify a default. We will be able to provide better guidance soon, after we start using Stata 13. I am sure other users already have some experience with household level clustering and perhaps they can volunteer a comment.

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Sun, 13 Apr 2014 23:09:03 GMT
View Forum Message <> Reply to Message

Tom (via Liz):

Do you know what Stata does regarding estimating cluster-robust standard errors using the svy: command? If it is using the "cluster robust" sandwich estimator, clustering at PSU and Household would be the same as clustering at PSU. But if its using some random-effects-type correction (some Moulton-type paramteric specification of the V/C matrix of error terms), then multi-level clustering would be different. I've never figured out what Stata is doing with the svy. Using "reg Y X, cluster(clustervar)" uses the sandwich estimator that would subsume household in PSU, but it sounds like that is not what svy: is doing.

Thanks.

Subject: Re: How do I account for clustering within families?
Posted by Liz-DHS on Thu, 17 Apr 2014 16:42:08 GMT
View Forum Message <> Reply to Message

Dear Reduced Forum,
Tom is currently out of the country, but I will forward to him.
Thanks!

Subject: Re: How do I account for clustering within families?
Posted by Liz-DHS on Fri, 18 Apr 2014 17:51:41 GMT
View Forum Message <> Reply to Message

Dear Reduced Forum,
This is a question for Stata Corp, not for us. Tom
********** *****
I'll try following up with Stata re:
Do you know what Stata does regarding estimating cluster-robust standard errors using the svy: command? If it is using the "cluster robust" sandwich estimator, clustering at PSU and Household would be the same as clustering at PSU. But if its using some random-effects-type correction (some Moulton-type paramteric specification of the V/C matrix of error terms), then multi-level clustering would be different. I've never figured out what Stata is doing with the svy. Using "reg Y X, cluster(clustervar)" uses the sandwich estimator that would subsume household in PSU, but it sounds like that is not what svy: is doing.

Subject: Re: How do I account for clustering within families?
Posted by Liz-DHS on Thu, 24 Apr 2014 17:28:10 GMT
View Forum Message <> Reply to Message

From Stata Tech Support:
If you take a look at the section "Linearized/robust variance estimation" in the manual entry of "variance estimation-Variance estimation for survey data" [(SVY] manual), near the end of this section, it says

> "V^{G^(beta)}|beta=beta^} is computed using the design-based
> variance estimator for a total."

Then in the section "Variance of the total", you could see how PSUs and multi-stage sampling units are handled in the formulas.

A quick way to open the PDF manual is to first type

 help svy

and click the hyperlink "[SVY] svy" in blue at the top of this page. You could then navigate the Bookmarks on the left of the PDF screen and under the bookmark "[SVY] Survey Data", select "variance estimation".

Subject: Re: How do I account for clustering within families?
Posted by vega25 on Thu, 04 Dec 2014 01:33:54 GMT
View Forum Message <> Reply to Message

Hi, Thanks very much for your earlier very helpful responses. I decided in the last analysis that I was running in April to go with the cluster(psu) option. But I have run into this problem again and in discussion with some colleagues.

The background is once again that I am trying to analyze women's and children's outcomes associated with household characteristics. The DHS interviews all eligible women in the household, not just one per household. Hence the individual women's dataset has some women who share household characteristics. In Bangladesh for example, the number is 4% - not large by any means. But I am curious to see what the best strategy to deal with this issue of some women sharing household characteristics should be.

Is this too small a number of multiple women per household that I should ignore it, or is another strategy advisable? Some of my colleagues suggested that I pick one woman per household at random and then perform my analysis on the smaller individual sample so that I solve the problem in one go. My concern with this is that I am losing valuable information, and that I am no longer certain that my sample will then be representative since I cannot prove that the presence of multiple women per household is a random event. I'd also be keen to know technically in STATA how to drop cases that share household characteristics.

The alternative strategy that I was considering is clustering at the psu level - the same strategy that Dr. Tom Pullum had recommended earlier. But as we discussed earlier, that would only address the standard errors, not the point estimates.

Thoughts?

Subject: Re: How do I account for clustering within families?
Posted by user-rhs on Thu, 04 Dec 2014 01:43:21 GMT
View Forum Message <> Reply to Message

It really depends on what kind of analysis you are running. If you are running multilevel models, you will be able to specify household as a nesting variable in addition to cluster/psu.

RHS

Subject: Re: How do I account for clustering within families?
Posted by ahmed89o on Sat, 09 Jul 2016 21:14:50 GMT
View Forum Message <> Reply to Message

Dear Colleagues,

To seal this post after 3 years. Now STATA 13 and 14 are out. Could you do secold level clustering now with new STATA? would you please share the code?

Subject: Re: How do I account for clustering within families?
Posted by user-rhs on Sun, 10 Jul 2016 02:35:50 GMT
View Forum Message <> Reply to Message

Typing -help svyset- into the command window will give you the answers you seek

RHS

Subject: Re: How do I account for clustering within families?
Posted by ab803 on Fri, 22 Sep 2017 01:37:58 GMT
View Forum Message <> Reply to Message

Hi DHS Forum,

I'm using the PR dataset to examine child health outcomes for several countries between 2000 and 2014. I am not pooling countries, or surveys to get an aggregated estimate, but I am interested in whether estimates changed between each survey for a given country. For a given country, I plan to append all the surveys for that country and then examine change over time. After reading the existing threads, I have two questions I'd be grateful for your help with:

1.Do I need to re-weight or re-normalize? Based on comments in earlier threads it is my understanding that I don't need to re-weight or re-normalize as I am appending multiple years of data for a country and using i.year variables in my regressions to understand change over time. I'd be very grateful if you could confirm this.

2. In my svyset command, I would like to account for (1) the strata and (2) the non-independence and clustering of children within the household. Would the following command be appropriate?
egen stratumid=group(hv024 hv025)
svyset [pw=hv005], psu(hv021) strata(stratumid) singleunit(centered) || hv002

Many thanks!

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Sun, 24 Sep 2017 19:20:23 GMT
View Forum Message <> Reply to Message

One way to do this cleanly without having to sacrifice anything regarding survey design (clustering, stratification, weighting) would be to estimate each survey round individually, and then

use Seemingly Unrelated Regressions commands (suest and sureg in Stata) to test the equality of coefficients across survey rounds. To do that you'd just do your standard DHS svyset/svy: stuff on an individual round, save the information (see help commands for suest/sureg), do the other rounds the same way, and then directly test the coefficients.

You can do it pooled by country too, but when you make those stratumid (which I presume you do before merging) you might end up with something weird where a value of that variable repeats across survey rounds. You could just append a "001" and "002" to the end of each stratumid where "001" would mean from survey round one. You'd also need to generate new PSU identifiers, since those repeat values from survey to survey but don't represent the same PSUs.

I assume by "i.year" you are meaning survey year and not cohort, right? Also remember, the Beta/p-value on your year dummies is only relative to the omitted one, so if you have 3 rounds you'd need to test the survey round dummies against each other (not just the 0 that represents the omitted group).

Subject: Re: How do I account for clustering within families?
Posted by ab803 on Mon, 25 Sep 2017 04:43:22 GMT
View Forum Message <> Reply to Message

Thank you so much! I really appreciate your response.

I'm definitely considering suest and sureg, it's very helpful to read that you would recommend these! I am generating the stratumid before merging. Great idea re. appending 001 or 002 to the stratum ID.

Two final questions:

1. Am I accounting for the non-independence of children within households correctly using hv002 as follows:

egen stratumid=group(hv024 hv025)
svyset [pw=hv005], psu(hv021) strata(stratumid) singleunit(centered) || hv002

2. Any advice on how to generate PSU identifiers? Could I also add 001 or 002 depending on the survey here? Or, just add the survey year itself to end of the PSU id (or even the stratum ID)?

Thanks again!

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Mon, 25 Sep 2017 18:26:23 GMT
View Forum Message <> Reply to Message

Glad I could help. Re follow-up questions:

1. Yes, that is fine if you are doing the surveys separately, but if you are merging them you'd need the new identifiers that include survey year.

2. Yeah - you could just add some numbers on the end of the PSU for survey year or round or whatever...just anything that makes the values of the variable unique by survey round. But of course if you do each survey separately, you can just use the code you have (assuming that strata are defined that way in your particular surveys...sometimes strata definitions change, but it is usually done in region-by-urban groupings, which I think is what you have there.

Subject: Re: How do I account for clustering within families?
Posted by ab803 on Tue, 26 Sep 2017 05:24:58 GMT
View Forum Message <> Reply to Message

Thanks again!

In order to specify the second level of the multi-stage design and account for the non-independance of children in households, stata requires that the fpc is defined for the first level. Any advice on how to define the FPC in what follows?

```
egen stratumid=group(hv024 hv025)
svyset [pw=hv005], psu(hv021) strata(stratumid) singleunit(centered) || hv002
```

Many thanks!

Subject: Re: How do I account for clustering within families?
Posted by Reduced-For(u)m on Thu, 28 Sep 2017 20:44:18 GMT
View Forum Message <> Reply to Message

I'm not sure exactly what you are asking. If you are clustering at the PSU level, that subsumes the household, and so you are allowing for interdependence of error terms within families anyway (since you are allowing it across families in the same PSU, and no household spans multiple PSUs). You could just set the weights using svyset and specific the clustering after the regression (with something like ", cluster(PSU)", but I think what you have is fine. If you want to do an explicit multi-level model, I think you might want to switch to the new mixed-model commands in Stata under the "mixed" family (formerly called xtmixed). Questions on setting up the hierarchical structure code there would need to be sent to the StataList, not here, and I'm not an expert on that suite of commands.
