Subject: accounting clustering effects of women's data when using baby-based analysis
Posted by rkinoshita on Fri, 27 Sep 2013 12:42:43 GMT
View Forum Message <> Reply to Message

Dear colleagues,

I am a master's student in Epi and currently finishing my master's thesis.

I am using DHS data from Nicaragua looking at the relationship between partner violence and newborn health outcomes, primarily neonatal, infant and under5 mortality. I am using a dataset in which women's and children's datasets were merged. When I merged two datasets, they created duplicated women's records because I am doing birth-based analysis (each birth needs to have their mother's record).

I've run into trouble with a birth-based analysis because the data for mothers who have more than one birth is duplicated, ending up over-representation of women who had more than one child in the analysis of regression. We could limit the analysis to one birth per woman but in this case with a mortality outcome, I prefer to keep all the births data to get adequate power.

My supervisors at university suggested to use vce option of regression in STATA. other suggested options are random effect modeling and use of a hierarchical model. For none of the methods, I am familiar with, and my deadline for submitting thesis is coming soon.

could you please let me know what would be the easiest and most appropriate method in this case, to minimize the over-representation of women with multiple births in my analysis? I am concerned because parity could be associated with other confounder or effect modifiers that I am looking at, and need to control in the final model.

thanks in advance.
Rinko

Subject: Re: accounting clustering effects of women's data when using baby-based analysis
Posted by Reduced-For(u)m on Fri, 27 Sep 2013 18:51:10 GMT
View Forum Message <> Reply to Message

R,

I think there are two things worth considering.

1 - Standard errors/confidence intervals: In my field, we would use a "clustered" standard errors approach here (one of the "vce" options). To deal with just the multiple-observations-from-same-woman problem, you would want to cluster on the household ID number. But, since the survey is done in a cluster-randomized framework, you would probably want to cluster at a higher (bigger) level to subsume the survey effects, probably PSU (primary

sampling unit - its a variable in your data"). If you use the DHS recommended standard errors (see the FAQs on the measureDHS site), that will account for survey clustering too, but I've had trouble figuring out if that is a random effects method or a "clustering" correction (non-parametric V/C matrix). So I would just do it by hand: reg Y X [pweight=weight], cluster(PSU) . Note that by clustering on PSU, you are subsuming the woman, so you take care of two problems at once.

2 - point estimates: since you have the same woman in the data multiple times, you will have to adjust your weights to account for this. The weights from the "birth recode" instead of the "woman recode" should take care of this (anyone disagree? I haven't used those weights). You could compare weights and maybe get some idea if this is right.

Anyway, I would cluster at the PSU level for your standard errors and make sure you weight the data to account for the multiple women. That is standard in my field (applied microeconomics), and is increasingly common in other fields. Other people like the hierarchical model, but point-estimate-wise you should get very similar results either way, and the clustering method should be slightly more conservative and easier to implement and defend.

Subject: Re: accounting clustering effects of women's data when using baby-based analysis
Posted by rkinoshita on Sat, 28 Sep 2013 09:43:12 GMT

View Forum Message <> Reply to Message

Hi. Thank you so much for your prompt response. This is indeed very helpful. I really appreciate it.

sorry for being insistent, and asking more questions but I do have a few follow up questions for you.

1) I run the cluster command that you mentioned in your point 1, and it does not work for some reason. I past my STATA output here. IPVlife- my exposure variable, Infant- infant deaths for all births. HHCLUST= PSU. I thought that it did not work because this command is only for regression rather than logistic, but it still did not work. any thoughts would be great.

```
. xi: svy: logistic IPVlife infant [pweight=weight], cluster(HHCLUST)
weights not allowed
r(101);

. xi: svy: logistic IPVlife infant [pweight=weight2], cluster(HHCLUST)
weights not allowed
r(101);

. help cluster

. xi: svy: logistic IPVlife infant [pw=weight2], cluster(HHCLUST)
weights not allowed
r(101);
```

```
. xi: svy: regres IPVlife infant [pweight=weight2], cluster(HHCLUST)
weights not allowed
r(101);
```

2) I do not fully understand what is the difference between -- a random effects method and a "clustering" correction (non-parametric V/C matrix) that you mentioned in your first suggestion. But if I understand well, you are not recommending vce option but using reg Y X....that you suggested, and the reason for this is that vce will only consider the clustering at individual households but not at the bigger level (e.g. PSU level), correct? Does this mean vce option cannot use for clustering at PSU, or in other words, I cannot use vce option with varname PSU??

3) for your second suggestion, I do not understand the weight in "women recode" that you mentioned. in my merged dataset, I have pesomef and pesonino - two weighting variables- one for women and children. Are you talking about these variables? how do you compare these variables (obviously I cannot list it) and how do you use adjust the weight? Do I create another weight variable using pesonino in the merged dataset? FYI, before merging the two datasets (women and baby), each dataset was already accounted for weight.
below is what I did for women, and the same thing in baby's dataset using pesoNino

```
gen weight2= PesoMEF
svyset HHCLUST [pw=weight2], strata(HHDEPAR)
```

thanks very much again for your help.
Rinko

Subject: Re: accounting clustering effects of women's data when using baby-based analysis
Posted by rkinoshita on Sat, 28 Sep 2013 10:22:43 GMT
View Forum Message <> Reply to Message

hi, and if you could look at below outputs from STATA and let me know what you think. When I used vce option, the results for vce (robust) and vce (cluster...) are very different. do you know why?
thanks

```
. regress IPVlife neonatal, vce (cluster  HHCLUST)

Linear regression                               Number of obs =   32115
                                        F(  1,   730) =    2.54
                                        Prob > F      =  0.1117
                                        R-squared     =  0.0001
                                        Root MSE      =  .42334
```

```
                     (Std. Err. adjusted for 731 clusters in HHCLUST)
------------------------------------------------------------------------------
             |               Robust
     IPVlife |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    neonatal |   .0337656   .0212035     1.59   0.112    -.0078615    .0753928
       _cons |   1.699111   .0427602    39.74   0.000     1.615164    1.783059
------------------------------------------------------------------------------
```

```
regress IPVlife neonatal, vce(robust)

Linear regression                                      Number of obs =   32115
                                                       F(  1, 32113) =    3.34
                                                       Prob > F      =  0.0674
                                                       R-squared     =  0.0001
                                                       Root MSE      =  .42334

------------------------------------------------------------------------------
             |               Robust
     IPVlife |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    neonatal |   .0337656   .0184639     1.83   0.067    -.0024243    .0699556
       _cons |   1.699111   .0366966    46.30   0.000     1.627184    1.771038
------------------------------------------------------------------------------
```

Subject: Re: accounting clustering effects of women's data when using baby-based analysis

Posted by Reduced-For(u)m on Sat, 28 Sep 2013 21:40:46 GMT

View Forum Message <> Reply to Message

Hey R,

  Glad I could be some help, and can answer a few of your questions. The results below I'll respond to in another post because there is something kinda awesome (awesome from an applied stats perspective, not for your research) about that. Anyway, on these questions:

1 - I think you figured it out, but basically you can't use both "svy" and specify a V/C matrix - the svy command is telling stata how you want to weight and compute standard errors, so you can't then tell it to use some other standard error calculation. So when you want to weight and specify

an SE computation too, then drop the "svy".

2 - This is tricky, but here is the basic idea: both are specifications of the V/C matrix for error terms - errors within a "group" are allowed to be correlated in some way and have some heteroskedatasticity. In the RE model, you are parametrically modelling this V/C matrix - the within-group off-diagonals are a parameter you estimate that is constant (in some way) within a group. You are giving structure to how the matrix should look, because of your assumptions about the data.

The "clustering" V/C matrix (called an "arbitrary" V/C by some) sets up "groups" similar to an RE estimator, but now you make no parametric assumptions "within group" about the structure of the V/C matrix. Each off diagonal is just E'E (residuals squared or person A's residual times person B's for A,B in same cluster). It allows the model more freedom within group. This tends to lead to higher standard error estimates than the RE model because you are imposing less structure (tradeoff between SE size and believability).

A really good, and fairly readable paper on this is Cameron and Miller's "Robust Inference with clustered data" which goes over REs and clustering and bootstraps.

3 - hmmmm... I'm not sure about these variables, so I'll just say that my main point was that you want a weight that is designed to be used at the level of your analysis, in this case the child. So I would use the weight that comes with the child recode, and not the one from the woman recode. Maybe a DHS staffer can help with this better than I can.

OK. More below on your regression results.

Subject: Re: accounting clustering effects of women's data when using baby-based analysis
Posted by Reduced-For(u)m on Sat, 28 Sep 2013 21:52:08 GMT

View Forum Message <> Reply to Message

So I'm guessing the big concern here is that using an RE model you get "significant" results, and using clustering you don't. My honest opinion - I'd believe the clustering p-values above the RE ones. On the other hand, in Epi all previous results have p-values from RE models, so you'd be subjecting your results to a tougher test than is common. Common is probably wrong in Epi (meaning that my experience is most of their SE estimates are too small in general), but it is also the baseline. So which results you decide to go with is totally up to you.

As for why this is happening - well, since clustering imposes less structure on the data (and to repeat what I said above) and so almost always produces larger SEs than RE models. In your case, that causes the 90% CI to go from excluding 0 to including it. That sucks for you, but is probably more "right". Notice that the point estimates are exactly the same, so all that is changing is that when you don't impose structure on your residuals, your residuals (within-group) tell a different story about precision.

So I guess my point is: clustering will lead to larger SEs than RE models, but these SEs are probably more "correct" or closer to "correct" (meaning produce reliable rejection rates) than the RE estimates.

One last thing - how many "clusters" do you have? If it is less than 50 or so, even these cluster SEs might be too small. Otherwise, they should be good.

Sorry to bear what I think is probably bad news for your research, but I also think your result is probably still really interesting, since I think of "statistical significance" as something much less important than other people do. Getting precision estimates right is important, but which side of a threshold it falls on is not that important (to me). Good luck. Let me know if I can help more.

Subject: Re: accounting clustering effects of women's data when using baby-based analysis

Posted by rkinoshita on Sun, 29 Sep 2013 14:21:05 GMT

View Forum Message <> Reply to Message

Hi. Thank you so much for all of useful inputs- they will be fruits for thoughts for my paper when I discuss about limitations and interpretations of the significant (or non-significant) results.

As for the number of clusters, please see below, women by area but limiting to those women who have ever married or in union because they also responded to my two violence variables (main exposure and outcome variables). Number of strata is 17 and PSU is 731. I am guessing that this would probably make less justifiable to use cluster model (which makes more precise estimates of SE than Robust model)- then again, my question is for which model do I need to use now???

I was discussing with my supervisor and we agreed that we will stick to birth-based analysis for associations between intimate partner violence and infant/child outcomes and discuss possible biases arising from repeated data of women with more than one birth.

My supervisor also found that svy + logistic is the same as using just logistic with vce(cluster) as an option, but with added value of weighting too. she suggested me to stick to svy + logistic because in order to pass this thesis and get my degree out, I don't need to learn new methods. Later on, I would need to look at this though, as I am hoping to publish this.

anyway, thanks a lot, again!
Rinko

```
. svy: tab Area if QW814F==1, obs cell count format(%10.4f)
(running tabulate on estimation sample)

Number of strata   =        17                  Number of obs      =     12065
Number of PSUs     =       731                  Population size    =  16178.63
                                                Design df          =       714
```

```
----------------------------------------------
Area de   |
Residenci |
a         |      count  proportions        obs
----------+-----------------------------------
   Urbano |   9419.3941      0.5822    5935.0000
    Rural |   6759.2356      0.4178    6130.0000
          |
    Total |  16178.6297      1.0000   12065.0000
----------------------------------------------
  Key:  count     =  weighted counts
        propor~s  =  cell proportions
        obs       =  number of observations
```

Subject: Re: accounting clustering effects of women's data when using baby-based analysis
Posted by rkinoshita on Sun, 29 Sep 2013 14:27:46 GMT
View Forum Message <> Reply to Message

thanks so much for your detailed explanation and also recommended reading. I would definitely look into this

yes, it would be great if DHS staff could help me with weighting variables.

When I am using birth-based analysis (that is, merged birth history and women's datasets, creating one women's record per each birth), I assume I do not need to do svyset again in the merged dataset, because I have already done this in birth and women's datasets, using their respective weighting variable. If someone could please confirm this, that would be great. Here is what I did separately in women's and birth datasets before merging them using "joinby....."
thanks
Rinko

```
gen weight= PesoMEF
svyset HHCLUST [pw=weight], strata(HHDEPAR)

gen weight= NinoMEF
svyset HHCLUST [pw=weight], strata(HHDEPAR)
```

Subject: Re: accounting clustering effects of women's data when using baby-based analysis
Posted by Liz-DHS on Mon, 27 Jan 2014 22:24:08 GMT
View Forum Message <> Reply to Message

Dear User,
Please let us know if you still need assistance with this. Thank you Reduced-For(u)m for providing so much support. If support is still needed, please let us know and I will try to find one of our staff experts to assist with this. In the meantimne, here is a link with more information about using weights http://www.measuredhs.com/data/Using-DataSets-for-Analysis.cfm#CP_JUMP_14042