Subject: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Thu, 17 Dec 2015 09:35:06 GMT
View Forum Message <> Reply to Message

Dear Sirs,

I am collapsing Egypt DHS data on the district level in order to estimate the impact of district interventions. I run a FEs, a REs and a diff-in-diff model. Shall I use weights in this case? And if the answer is yes, how can I weigh data on the district level? At which stage should I weigh data (collapsing data, calculating district-level indicators, running the regression, etc.)?

Also, do I need to adjust for the clusters (the primary sampling units) and the strata?

Thank you.

Kind regards,
Amira

Subject: Re: Weighting district-level data
Posted by Reduced-For(u)m on Thu, 17 Dec 2015 22:32:26 GMT
View Forum Message <> Reply to Message

This is (again) a question that has no definite answer, but there are better and worse ones.

First off, since you are using these types of estimators, should I assume that you are using 2 or 3 rounds of the DHS? In that case, you have to be fairly careful when you calculate your district-level means (which become the observations in your estimates). There are also some big issues with calculating standard errors (p-values) which relates to "clustering". Here is how I would do it, and then an alternative:

Weighting: you should calculate the survey-specific district means using the survey given weights. If you do this separately by survey round, you won't have to worry at all about re-noramlizing the weights because you'll be calculating a representative mean of the district level variables (what become your observations). Also note, if you wanted to make the final regression "population representative" you could weight each district-year-level observation by the population of the district - so larger districts would get more weight than smaller ones. This may or may not be reasonable, and is up to you and your specific interests/needs.

Clustering: clustering at PSU is not sufficient in this case, at least not usually. The rule of thumb here would be to cluster at the "district" level - the level at which you collapse observations/assign "policy intervention". The usual reference is Bertrand, Duflo and Mullainathan "How Much Should we Trust Difference-in-Difference Estimates". To do this in Stata, when you define PSU in your "svyset" command, you use the district identifier (that is common across survey rounds). You can include the strata here too, but I don't think it will make much of a difference (and if it does, it should make your p-values slightly smaller).

Note: If you have fewer than 30 or 40 districts, you should also see Cameron, Gelbach and Miller "Bootstrap Based Improvements for Inference with Few Clusters" - there is a new Stata package that makes doing those "wild-t bootstraps" very easy: see "cgmreg" group of .ado files you can download.

Now - there are also a couple more ways to do this. In particular, you could do this same analysis on individual-level data with group-level covariates. Everything is the same, but now instead of weighting when you collapse, you'd have to weight in the regression. In this case, you could either follow Gary Solon in "What are We Weighting For" and argue that, with causal effects, you don't need to weight, or you can follow standard DHS recommendation and re-normalize your survey weights and then apply those in the regression analysis.

Hope some of this helps. I can follow up with details on one of these methods if you decide you like one and still aren't sure what to do.

Subject: Re: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Fri, 18 Dec 2015 16:55:42 GMT
View Forum Message <> Reply to Message

I would like, first hand, to thank you for your detailed elaboration.

Yes, I am using 3 rounds of Egypt DHS to feed the FEs/REs model. I use only 2 rounds to feed the diff-in-diff model.

The problem is that I had already run all models without weighting or clustering. I had no idea I had to weigh or cluster before running my models. My point is: if I had already calculated the district-level outcomes (dependent variables) without weighting or clustering, how can I apply what you have kindly recommended when you pointed out that "Everything is the same, but now instead of weighting when you collapse, you'd have to weight in the regression". To weigh and cluster in the regression will save me a lot of time unless it is necessary to go back to the the stage of collapsing (and apply weights and cluster then).

Please let me know should you require further elaboration.

Note: I have more than 200 districts.

Subject: Re: Weighting district-level data
Posted by Reduced-For(u)m on Fri, 18 Dec 2015 17:47:39 GMT
View Forum Message <> Reply to Message

1) It should be very easy to re-calculate the district-by-survey-round means again, this time using

the weights:

collapse Y [pweight=weight], by(district)

2) You can not do the probability weighting after you collapse - you can only adjust for population differences across districts after you collapse.

3) the clustering is easy and you don't have to do anything until you run the regressions: reg Y X, cluster(district)

4) if you have 200 "districts" does that mean that PSU is your "district"? If so, now you have an entirely new problem - namely, that "districts" are probably changing from round to round. How are you matching your "districts" from survey to survey, and what (administratively or in DHS-terms) is the "district" you are using?

So to sum up: you have to weight when you collapse (construct district-level means), you cluster in the regression, and I am not sure that you really have the information you think you do if you are using such small administrative regions (because usually these change from survey round to round, but maybe Egypt is different).

Subject: Re: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Fri, 18 Dec 2015 19:43:59 GMT
View Forum Message <> Reply to Message

1) When I try to re-calculate the district-by-survey-round indicators again using the weights, I get the same results for each district (as in the unweighted case). I think this is probably due to the way I create/calculate the district-level indicator (below):

collapse (sum) women_using_modern_methods (count) women_interviewed, by(District)
gen %women_currently_using_modern_methods=(sum_of_women_using_mo
dern_methods/count_of_women_interviewed)*100

Note: I denote women_using_modern_methods by 1 before calculating the district-level indicator.

When I use the weights as you have kindly suggested (below), the district-level indicators are the same as in the unweighted case:

gen myweight=v005/1000000
collapse (sum) women_using_modern_methods (count) women_interviewed [pweight=myweight], by(District)
gen %women_currently_using_modern_methods=(sum_of_women_using_mo
dern_methods/count_of_women_interviewed)*100

I do NOT simply create the district-level indicators as means of observations of individuals in these districts.

2) You made this point quite clear. Thank you.

3) I fail to do the clustering when I run the diff-in-diff regression below:

diff district_indicator, t (treated) p (t), cluster(District)

Stata 12 gives me the return code 198 of invalid syntax. Shall the 'District' variable (or whichever variable I use to cluster) be written down in a particular way?

4) The PSU is not the district in my case. I spatially join the displaced cluster locations of women to the GIS polygon data of Egypt's district boundaries. i.e. I allocate each woman to the relevant district.

Subject: Re: Weighting district-level data
Posted by Reduced-For(u)m on Fri, 18 Dec 2015 20:37:46 GMT
View Forum Message <> Reply to Message

Interesting. Quick replies:

1 - cool. My guess is that since it is usually 1 PSU going to 1 district, and all observations in some PSU have the same weight, that it is basically the numerator/denominator canceling each other out. But while we are on that - since it is a fraction, and since that fraction is an estimate of the population proportion, fewer observations means a less good estimate of the actual proportion. Clustering might not work here at all since you would lose the uncertainty generated in your first stage - you might have to bootstrap both stages. I'll leave that to you to decide how far you want to go down that rabbit-hole, but you may want to account for the uncertainty in your "observations" because those are estimates themselves.

2 - it is nice to help (smiley face)

3 - diff district_indicator, t (treated) p (t), cluster(District) ... I think the problem is the extra "," after p(t). Delete that, and I think it will run.

4 -I see what you are doing. That is interesting, and I could see how it would work. But you should also know that you are not necessarily comparing "apples to apples" anymore. Suppose PSU 1 is in District 1 in round 1. Then, in Round 2, District 1 contains PSU 397 (which wasn't sampled in round 1). Then you are using different people from different towns/areas to define the same "district" variable. How many PSUs per district do you have, on average? I ask because with many, I'd think a law of large numbers might apply and you'd be fine. But with only 1 or 2 PSUs per round, much of the difference across time within district is going to be do to sampling variation and not do to real changes. Again, in theory, with many N(obs), G(groups) and T(periods) you are OK, but in finite numbers you are asking a whole lot of the data.

Subject: Re: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Sat, 19 Dec 2015 11:11:35 GMT
View Forum Message <> Reply to Message

1- In my case, a district has more than one PSU. But given that the numerator/denominator cancel each other out, how can I weigh my data? And shall I cluster when I regress afterwards (as you have taught me in your comment number 3)? You pointed out that "Clustering might not work here at all".

Note: I use bootstrapped standard errors in the second stage when I regress.

3- You are right. When I delete the extra "," after p(t), the model run. Thank you!

4- I do not know for sure the number of PSUs per district (probably around 2-6 PSUs). But I know that each district has a reasonable number of clusters, not just one cluster.

Subject: Re: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Mon, 21 Dec 2015 10:48:43 GMT
View Forum Message <> Reply to Message

Good morning :-)

I would like to add something to my message below:

With reference to points number (1) and (3), when I run the FE/RE or the diff-in-diff model adding the word 'cluster' to the Stata command, the results are very close.

Thank you.

Kind regards,
Amira

Subject: Re: Weighting district-level data
Posted by Reduced-For(u)m on Mon, 21 Dec 2015 21:08:13 GMT
View Forum Message <> Reply to Message

Re bootstrap: I think to account for the imprecision in the "first-stage" - meaning the calculation of the percentage of women using modern methods - you'd have to bootstrap BOTH stages. That is, draw a bootstrap sample of observations within each district, estimate the proportions, run the second-stage regression on those aggregate values, and then repeat that whole process. But that is just a suggestion - you'd have to check how people in your field do that (there are probably other ways, say via multi-level modeling, to make the same kind of corrections.

As for why the FE and DnD are so similar - if you have just 2 rounds, the FE and DnD should be doing pretty much the exact same thing (though perhaps with the algorithm you are using, come

to slightly different numerical results). Or, if you mean that by "clustering" your estimates change very little - in effect (using the same regression model for point estimates) it should not change the point estimates at all, only the p-values/standard errors.

Subject: Re: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Mon, 21 Dec 2015 21:51:34 GMT
View Forum Message <> Reply to Message

Thank you for your assistance. And I would appreciate your tolerance as I have some more questions:

1- Do I still have to weigh the data?

2- And given that the numerator and the denominator cancel each other out in the first stage of calculating the percent of women in a district using modern methods, how I can weigh my data?

3- Shall I use clustering in the second stage (regression) if I weighed my data in the first stage?

4- I do not think I would be able to bootsrap the first stage by myself. Also, I think this will consume a lot of time. Do you think this is a necessary step in my case? Will it affect the results?

Thank you for your guidance.

Kind regards,
Amira

Subject: Re: Weighting district-level data
Posted by amira.elshal.1@city.ac.uk on Thu, 24 Dec 2015 09:00:37 GMT
View Forum Message <> Reply to Message

Dear Sir,

This is a kind reminder of my message below. I apologize for disturbing you again and I would appreciate your tolerance as I have a couple of more questions:

1- Do I still have to weigh the data? If I decide to follow Gary Solon in "What are We Weighting For" and argue that, with causal effects, I don't need to weight given that I carry out my analysis on district-level with district-level covariates (population size is one of the covariates I include), shall I cluster afterwards in the second stage (regression)? You have mentioned that "Clustering might not work here at all since you would lose the uncertainty generated in your first stage".

2- If I decide to weigh my data, how can I do so given that the numerator and the denominator cancel each other out in the first stage of calculating the percent of women in a district using modern methods?

3- Shall I use clustering in the second stage (regression) if I weighed my data in the first stage? Is clustering (in the second stage) related to my decision to weigh/NOT weigh data (in the first stage)?

4- If I have to cluster, shall I directly use reg Y X, cluster(district)? Or shall I first define PSU in the "svyset" command? I have not used the "svyset" command before.

5- I do not think I would be able to bootstrap the first stage by myself. Also, I think this will consume a lot of time. Do you think this is a necessary step in my case? To what extent will NOT bootstrapping affect the results given that I bootstrap standard errors in the second stage (regression)?

Once again, I would like to thank you for your guidance.

Kind regards,
Amira

Subject: Re: Weighting district-level data
Posted by Reduced-For(u)m on Mon, 28 Dec 2015 22:53:23 GMT
View Forum Message <> Reply to Message

Hi - sorry for the holiday-related delay. Some responses:

1. I think technically you would have to bootstrap the whole procedure (first and second stages) but, given that that is a little bit overboard, you should definitely cluster in the second stage.

2. I haven't math'd it out, but I think that if the weights cancel, you can just ignore them. If it doesn't make a difference, it doesn't make a difference.

3. You cluster regardless of your weighting choice - weighting is about getting a representative sample, clustering is about getting consistent standard error/p-value estimates.

4. You don't have to use "svyset", it is just one way to deal with it. You can just "reg Y X, cluster(district)" - or you can add your weights manually too (I also don't usually use svyset - only when trying to replicate DHS methods).

5. Your p-values will probably be a little too small if you don't bootstrap both stages, but it is probably not a huge deal. People ignore that all the time. You are just pretending that there is no measurement error in your observations (that is - pretending that your district means are actual observations and not estimates of district means). I mean - it is basically up to your field to decide what they will/won't accept, and I think 99% of smart researchers would not even consider this an issue - just me and some other sticklers.

Good luck - I think don't sweat it too much. But FYI - the bootstrap wouldn't be so hard really. You just place what you already do in a loop, use the "bsample" command (if Stata) at the beginning of the loop, and save the point estimate from each repetition. The standard deviation of

your point estimates is then used as the standard error of your estimate (I mean, it is an estimate of your standard error, but that is all anyone ever has).