Subject: Creating panel
Posted by Lukresha on Thu, 10 Aug 2017 05:43:32 GMT
View Forum Message <> Reply to Message

Hello,

I want to do an analysis using data from the 2003, 2008-09 and 2014 KDHS. I have appended the files to have one big file.

I would like clarification on whether setting the xtset in stata and going ahead to carry out a panel analysis is feasible with the data I am using.

Subject: Re: Creating panel
Posted by Lukresha on Thu, 17 Aug 2017 08:31:04 GMT
View Forum Message <> Reply to Message

I have gone through various materials and realized with DHS I will have a pseudo panel.

I have gone through Verbreek's Pseudo-Panels and Repeated Cross-Sections paper. I understood the need to create cohorts.
I created my cohort as below:

cohort=2014-hv105 for 2014 DHS
cohort=2009-hv105 for 2008-09 DHS
cohort=2003-hv105 for 2003 DHS

Then appended the 3 DHSs.

Is this the correct way to do it? Or there are steps that I am missing?

My analysis focuses on impact of household land ownership on children's school attendance

Subject: Re: Creating panel
Posted by Reduced-For(u)m on Thu, 17 Aug 2017 17:03:33 GMT
View Forum Message <> Reply to Message

This is a reasonable way to do it using that recode. But essentially you are just creating "year of

birth" variables - it is unclear why this is being used to generate a "pseudo-panel". That said, if you just want to include effects for year of birth, the variable you created will work pretty well. The only issue is that some surveys (I don't know about yours) conduct interviews that cover more than one calendar year (say, Nov-Feb). Then you would be a little bit off on birth year for some households, but maybe not in a way that is problematic (it would depend a lot on how you are structuring your regressions).

All that said, your code should get you something very close to "year of birth".

Subject: Re: Creating panel
Posted by Lukresha on Thu, 17 Aug 2017 18:45:42 GMT
View Forum Message <> Reply to Message

Thank you for the response.

My thought process was that once I create the year of birth, I would get the mean of each of my variables of interest in each birth year (which serves as a cohort) which I would then use to carry out the regression.

I am still new to pseudo panels. "But essentially you are just creating "year of birth" variables - it is unclear why this is being used to generate a "pseudo-panel"". How would you suggest I go about creating the pseudo panel? I am interested in how household's ownership of land has affected school attendance of children in the household over time (from 2003 to 2014). My dependent variable is a binary variable so I intended to use a probit model.

One of my data sets covers more than one calender year (2008-09), in that case, how can I go about dealing with this problem?

Also, do you know of any papers or articles that talk about empirical implementation of pseudo panels in stata?
(I tried using help pseudo panels in stata but I didn't get much material.)

Subject: Re: Creating panel
Posted by Reduced-For(u)m on Thu, 17 Aug 2017 19:05:57 GMT
View Forum Message <> Reply to Message

I don't think you gain much by collapsing the data in that way. Why not just use the individual-level observations?

Usually when you collapse data like that it is because you are doing something like bringing in external data that is merged to the cohort (say in your case, information on when schools were built; or maybe cohort variation in exposure to some mandatory schooling laws). Even then you don't necessarily need to collapse the data down, and can just merge the variables into the

individual-level data.

As for the problem of getting the birth year wrong - does it really matter? Isn't a mother born in December of, say, 2005 very similar to one born January of 2006? It isn't clear you would be wrong to lump these two groups together in one time effect. But again, it depends a lot on the data setup, such as if Dec/Jan born women had different "exposures" to something important.

Also, once you collapse the data down into averages, it wouldn't be a 0/1 variable on the left - it would be a proportion. In which case the probit model wouldn't be right. You would want to run a probit on the individual-level data...so again I don't see the need to create this pseudo-panel. You could just run a least-squares regression of some sort on the aggregate data.

That said, if there is a reason to generate the pseudo-panel, it is straightforward to do using the "collapse" command in Stata and the "by()" option as "by(cohort)" or "by(cohort region)" or whatever is appropriate. You would also perhaps want to collapse using the DHS sample weights (to get representative estimates), which is explained in various places on the forum and on Stata help forums. But it doesn't seem clear that you really want or need to do that.

Subject: Re: Creating panel
Posted by Lukresha on Thu, 17 Aug 2017 19:39:00 GMT
View Forum Message <> Reply to Message

I initially carried out a probit analysis using only data sets from 2014. I assumed that since I am now using data sets from 3 different years (2003, 2008-09 and 2014), then I need to create a pseudo panel since I do not have a "real panel".

Does it mean that even with 3 different years, I can still use probit and get unbiased results?

Subject: Re: Creating panel
Posted by Reduced-For(u)m on Thu, 17 Aug 2017 20:35:14 GMT
View Forum Message <> Reply to Message

Whether the estimate is biased or not depends to a great extent on the setup and what exact parameter you are trying to estimate. But in general, yes, you can pool multiple survey rounds together and use a probit, just include controls for survey round. You could also do the analysis on each dataset separately, and then test whether the coefficient you are interested in is changing from round-to-round. There are lots of interpretation issues and problems with aggregation, but none of those are solved by aggregating the data as you describe. Just pool the individual-level data together and control for cohort and/or time effects as you see fit. But there is no INHERENT bias in the probit version that wouldn't be there in an aggregate regression...just the same bias you'd have either way, which depends on exactly the model you are fitting and the parameter you

are estimating.