Subject: Accounting for different sampling areas over different years
Posted by UAB_user on Wed, 11 Feb 2015 04:57:02 GMT
View Forum Message <> Reply to Message

Hello,

I am using the Nepal DHS to look at factors affecting migration across the 01, 06, and 11 survey years.

I have de-normalized the weights for each year according to Ruilin's suggestions, but do I have to somehow account for the different sampling areas for each year. Would it be ok to merge all three years and use the cluster (V001) and strata (V023) variables in my analysis and assume the areas are the same for each survey round?

If I do have to adjust them, how do you recommend I go about doing so?

Thank you
Derek

Subject: Re: Accounting for different sampling areas over different years
Posted by Bridgette-DHS on Wed, 11 Feb 2015 15:44:18 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Specialists Ruilin Ren & Trevor Croft:

You need to do something to distinguish the cluster numbers (V001) that have the same value, but they actually came from different surveys. Each cluster from each survey should stand alone as a cluster in your merged file.

You can create a new cluster number as follows:

```
gen year =.
{ Convert Nepali years to unique survey years }
replace year=2011 if v007 == 2067 | v007 == 2068
replace year=2006 if v007 == 2062 | v007 == 2063
replace year=2001 if v007 == 2057 | v007 == 2058
egen newcluster = group(year v001)
```

Subject: Re: Accounting for different sampling areas over different years
Posted by UAB_user on Wed, 11 Feb 2015 23:14:13 GMT
View Forum Message <> Reply to Message

Great!

Thank you
Derek

Subject: Re: Accounting for different sampling areas over different years
Posted by UAB_user on Tue, 17 Feb 2015 20:56:01 GMT
View Forum Message <> Reply to Message

Would i have to do this to V023 as well?

Subject: Re: Accounting for different sampling areas over different years
Posted by Trevor-DHS on Wed, 18 Feb 2015 12:58:25 GMT
View Forum Message <> Reply to Message

While the strata in v023 are consistent across the 3 surveys and represent the same areas (unlike v001 which are different clusters in each survey year), I would recommend following the same procedure to create a separate strata for each survey year as for v001.

Subject: Re: Accounting for different sampling areas over different years
Posted by mmr-UMICH on Thu, 16 Apr 2015 17:08:59 GMT
View Forum Message <> Reply to Message

Strata are consistent across surveys for a country indicates that the codes/values of strata variable (after combining region and residence variables) are the same across the survey waves (e.g, 2001, 2006, 2011). If country has 5 regions and urban/rural, so there are 10 strata codes (say, 1 to 10) for each survey year. My understanding is that in pooled data set the number of strata is still to be 10. Because the stratification was the same but the sampling of clusters within stratum was different for each survey year, so cluster codes must be the different for identical strata across the survey waves. If we treat strata codes different across the surveys, the variance estimation is not only affected but also the degrees of freedom, confidence intervals, and p-value calculations.

Subject: Re: Accounting for different sampling areas over different years
Posted by Reduced-For(u)m on Thu, 16 Apr 2015 20:05:41 GMT
View Forum Message <> Reply to Message

My intuition is that you would want to use different strata too - the idea being that the stratification was done separately by survey round, even if they overlap - but I think this is probably, if not an open question in the survey analysis literature, at least sufficiently esoteric that there is no agreed-upon course of action. That said, I do have two points I'm more sure about:

1 - you say "If we treat strata codes different across the surveys, the variance estimation is not only affected but also the degrees of freedom, confidence intervals, and p-value calculations." But variance estimation will always affect CIs and P-values, and the effect of the loss of DF should not affect critical values, given the large number.

2 - depending on your variables of interest and how those are constructed, you might want to use a standard error estimator that accounts for more robust correlations than those you would use if you were just looking at a single, individual-level covariate from one survey. Error terms are likely correlated across time within region (worse if you are using aggregated or constructed variables on the right hand side of your regression) and the standard DHS method won't account for this, but clustering by spatial region across survey rounds would.