Subject: Re: Using weights in regression analysis
Posted by mnicolson on Fri, 14 Jun 2013 20:18:11 GMT
View Forum Message <> Reply to Message

Hi, I have a couple of questions about weighting which I was hoping someone might be able to help with?

Weighting, clustering and stratification for regression

An earlier poster has responded by saying:

"From the DHS FAQs (under "using data files": http://www.measuredhs.com/faq.cfm):

***First, use the svyset command to tell Stata how your data is set up:

```
*generate weight
generate weight = v005/1000000
```

```
*make unique strata values by region/urban-rural (label option automatically labels the results)
egen strata = group(v024 v025), label
*check results
tab strata
```

```
*tell Stata the weight (using pweights for robust standard errors), cluster (psu), and strata:
svyset [pweight=weight], psu(v021) strata(strata)
```

****Now for a regression - if you prefix regress with "svy:" Stata will now know how to weight your data and compute the right standard errors

```
svy: reg Y X
```

***Quick note: computing standard errors in this way is probably not OK for a lot of regressions. Without getting off track or all statsy, a good way to think of this is that this standard error calculation is alright IF the error terms and covariates are independently and identically distributed across observations, other than as operating through the sampling procedure (the stratification and clustering prior to randomization that produces the particular sample you have). I tend to think of these standard errors as the smallest the "true" standard errors could possibly be, but I'm kind of on the conservative/stickler end of this debate, and others would surely disagree."

I have followed this and all works fine - however, I have two questions.

(1) It seems that the command given above assumes that the data has been collected using one-stage design.

The Stata Manual defines one-stage design as follows:

"A commonly used single-stage survey design uses clustered sampling across several strata, where the clusters are sampled without replacement."

However, when I read the DHS country manuals, it suggests that samples were selected in two or more stages depending on whether the respondent comes from a rural or urban area. The Stata Manual states that we then have to use a different command, one which accounts for the multiple stages of sampling.

It gives the example:
"We have (fictional) data on American high school seniors (12th graders), and the data were collected
according to the following multistage design. In the first stage, counties were independently selected
within each state. In the second stage, schools were selected within each chosen county. Within each
chosen school, a questionnaire was filled out by every attending high school senior."

The stata command it suggests is:
svyset county [pw=sampwgt], strata(state) fpc(ncounties) || school, fpc(nschools)

Is the command -svyset [pweight=weight], psu(v021) strata(strata)- the correct way of dealing with DHS survey data? Or should I be using a command that takes into account the multiple-rounds used to collect DHS data?

I realise that there is also another factor to consider - namely, whether the clusters are sampled *with* or *without* replacement. Does DHS survey with replacement therefore making it unnecessary to account for the second-stage clusters?

(2) The comment suggests that this way of calculating standard errors ('the way' - accounting for weighting and stratification) won't be appropriate for a lot of regressions. Why is this? If it's not appropriate, does that mean an alternative way is to simply run a regression without weighting or accounting for sampling?

****

Even though I am using the sample weight, my tabulations differ from those in the country tables

I am analysing the India DHS dataset. My unit of analysis is the individual and I have appended all three DHS India datasets into one large dataset.

I am using the following command in order to attempt to replicate the total contraceptive prevalence rate given on p. 170 of the DHS-3 India country report:
tab cpr [iweight=weight] if (v025==0 | v501==1 | year==2005 & 2006)

I created the weight variable using the following command (given above)
generate weight = v005/1000000

cpr is a dummy variable that I have created from v313 (cpr=1 if any method is used; cpr=0 when no method is used)

v025==1 (means that the household type = urban)
v501==1(means that marital status = married)

My tabulation states that the CPR=45.43%
The country table states that the CPR=64%

Could the difference between my figure and the figure in the country table be due to the fact that I have appended the three datasets into one?

Or do you calculate the contraceptive prevalence rate differently from me? If so, how do you do it?

***

My apologies for the length of this post - I hope it all makes sense and I look forward to any responses

Thanks.