Subject: Clarification on Variables for svyset in STATA and generating stunting variable
Posted by 616blue on Wed, 29 Apr 2015 17:18:54 GMT
View Forum Message <> Reply to Message

Hello,

I've been through numerous threads on these two topics and would appreciate some additional clarification for my project.

#1. For using svyset in STATA, most threads note we need a psu, pweight and strata. There is consensus on psu=v021.
For strata, generally we would use group (v024, v025)-but this would differ by country what the strata is (and we would NOT use v022 in general). Where could I check what the strata details are specifically for my country of interest (ETHIOPIA)?
For sampwt, I've read both that we should use v005 directly/ use v005/1000000. Could I receive clarification on what I should use and why?

I have:
```
gen psu =    v021
egen strata = group(v024 v025), label
tab strata
gen sampwt = v005/1000000 //as per DHS  instruction//
svyset psu [pweight = sampwt], strata(strata)
```

#2. For generating the stunting variable, I previously used the coding:
```
codebook hw70
tab hw70 if hw70>9990,m
tab hw70 if hw70>9990,m nolabel
gen HAZ=hw70
replace HAZ=. if HAZ>=9996
histogram HAZ
gen stunted=.
replace stunted=0 if HAZ ~=.
replace stunted=1 if HAZ <-200
tab stunted
regress stunted
regress stunted [pweight=v005]
```

However, I found a response at: http://userforum.dhsprogram.com/index.php?t=tree&goto=3952&S=09ed0022cd4173b993b147e4fdc88183&srch=svy+stata #msg_3952
Re: svy, subpop [message #3952 is a reply to message #3942]
that suggests differently:

Following is a response from Senior DHS Specialist, Tom Pullum:

Here are the lines you need for a logit regression of stunting on the wealth index for children age 0-23 months.

```
* logit regression with stunting in months 0-23 as outcome
* use the PR file; KR file is limited to children living with mother
* usually limit to de facto residents, i.e. hv103=1
keep if hv103==1
* hc70 is the WHO haz score, already edited
gen stunting=0
replace stunting=1 if hc70<-2
replace stunting=. if hc70>600
* the cluster id is hv001=hv021
* stata will normalize the weights; no need to divide by 1000000
* hv022 may be the strata variable but always check
* always a good idea to include one of the singleunit options
svyset hv001 [pweight=hv005], strata(hv022) singleunit(centered)
* it is normal to use hc1 (=hv008-hc18) as age
svy: logit stunting i.hv270 if hc1<24
```

Questions are as follows:
Why would I use the PR file vs the BR file? Or asked differently, if I am interested in the mother's education on child's stunting outcomes, would it be correct to use the BR file-since it has the birth data for each child and mother info? Or would it even be the KR file?

Why would we limit to de facto residents? Is the svyset command not controlling for this?

Why would we use hc70<-2 vs hc<-200? I think on different threads I found that we need to multiply by 100.

I assume replace "stunting=. if hc70>600" is equivalent to "replace HAZ=. if HAZ>=9996" since the range for plausible values is up to 600 and anything else beyond is coded in the 9000's to indicate missing etc.?

The following four lines confuse me.

```
* the cluster id is hv001=hv021
* stata will normalize the weights; no need to divide by 1000000
* hv022 may be the strata variable but always check
* always a good idea to include one of the singleunit options
svyset hv001 [pweight=hv005], strata(hv022) singleunit(centered)
```

Why is the psu, weight and strata different from what is used in other examples, as illustrated in #1?
In what combination do we divide weights by 1000000 or not divide?

Thank you in advance!

Subject: Re: Clarification on Variables for svyset in STATA and generating stunting variable
Posted by user-rhs on Wed, 29 Apr 2015 17:44:41 GMT
View Forum Message <> Reply to Message

I'll answer the Stata and more general parts of your question, and leave the anthropometry-specific indicators for the experts:

Quote:Where could I check what the strata details are specifically for my country of interest (ETHIOPIA)?
Check the DHS final report for that country and survey year

Quote:The following four lines confuse me.
* the cluster id is hv001=hv021
* stata will normalize the weights; no need to divide by 1000000
* hv022 may be the strata variable but always check
* always a good idea to include one of the singleunit options
svyset hv001 [pweight=hv005], strata(hv022) singleunit(centered)

Why is the psu, weight and strata different from what is used in other examples, as illustrated in #1?
In what combination do we divide weights by 1000000 or not divide?

I assume the PSU question has to do with why HV001/HV021 was used instead of V001/V021. Recall that in the DHS datasets, different datasets have different prefixes for the variables. For example, in the women's recode dataset, the woman variables start with V (but different modules have different prefixes, e.g. B for birth hx questions). In the household recode, the prefixes are generally H-something (HV, HC, HW). Therefore, you will have the same variables with different prefixes between the datasets, i.e. HV001==V001, HV002==V002. You will need to create a uniform name for these variables when you want to merge the datasets together.

Re: the weights. The DHS final report for the survey year will have the instructions what to divide the weight variable by (I think I've seen instructions to divide by 100,000 somewhere before). In the grand scheme of things, using the weight variable as-is will give you the same means and LSEs, proportions, parameter estimates, tests of significance, etc. as using weight/100000. The only difference you will note is in the number of observations, where it will be 1,000,000 more than the cell size, number of obsns (in a regression) than if you had divided by 1,000,000.

Example from one of the Indonesia datasets:

Unweighted:
tab v025 rural

```
   Type of |
  place of |       rural
 residence |         0          1 |     Total
-----------+----------------------+----------
     Urban |     6,994          0 |     6,994
```

```
    Rural |         0      8,268 |     8,268
----------+----------------------+----------
    Total |     6,994      8,268 |    15,262
```

Weighted with weight-as is (svyset [pw=v005]):
svy: tab v025 rural,count format(%12.1g)
(running tabulate on estimation sample)

```
Number of strata   =         1              Number of obs      =        15262
Number of PSUs     =      1827              Population size    =  14782036227
                                             Design df         =         1826
```

```
-------------------------------------------
Type of   |
place of  |              rural
residence |         0          1       Total
----------+--------------------------------
   Urban | 7357950946          0  7357950946
   Rural |         0  7424085281  7424085281
         |
   Total | 7357950946  7424085281     1.5e+10
-------------------------------------------
  Key:  weighted counts
```

Weighted with weight/1000000 (svyset [pw=wt]):
svy: tab v025 rural,count format(%12.1g)
(running tabulate on estimation sample)

```
Number of strata   =         1              Number of obs      =     15262
Number of PSUs     =      1827              Population size    = 14782.036
                                             Design df         =      1826
```

```
------------------------------
Type of   |
place of  |       rural
residence |     0      1  Total
----------+-------------------
   Urban |  7358      0   7358
   Rural |     0   7424   7424
         |
   Total |  7358   7424  14782
------------------------------
  Key:  weighted counts
```

Subject: Re: Clarification on Variables for svyset in STATA and generating stunting variable
Posted by duke2015 on Sun, 14 Jun 2015 13:00:15 GMT
View Forum Message <> Reply to Message

I have a similar follow-up question to this one.

You said to use the hc 70 variable to calculate stunting. I could only find hc_1, hc_2, etc. Which data file is this from?

If I am doing a regression looking at the predictors of child stunting, and I want to include mother characteristics AND household characteristics AND child characteristics in my prediction model, why wouldn't I use hw70 from the children's recode file? Is there an advantage to using the hc_70 variable?

Thanks!

Subject: Re: Clarification on Variables for svyset in STATA and generating stunting variable
Posted by Reduced-For(u)m on Wed, 17 Jun 2015 21:31:31 GMT
View Forum Message <> Reply to Message

I believe that HC70 is the WHO anthropometric variable, while the lower numbers use the CDC standards. The WHO standards are only available for the latest rounds of the DHS (they were defined in 2006 or so). You can use the lower HC (or HW) numbers if you don't care which standardization you use, or you can either a) merge in the WHO anthropometrics from the appendix datasets available online; or b) calculate your own using a package like "zscore06" which can be downloaded into Stata.