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
```