
Subject: Re: Using weights in regression analysis
Posted by [Reduced-For\(u\)m](#) on Sun, 20 Oct 2013 23:09:17 GMT
[View Forum Message](#) <> [Reply to Message](#)

Here is some discussion of the problem, which continues in more (and helpful) detail if you follow the link.

<http://www.stata.com/support/faqs/statistics/stratum-with-on e-psu/>

Having a stratum with a single PSU is a fairly common problem. When there is only one PSU within a stratum, there is insufficient information with which to compute an estimate of that stratum's variance. Therefore, it is impossible to compute the variance of an estimated parameter when the data are from a stratified clustered design. There are two solutions. The first solution is to simply delete the stratum with the singleton PSU from your sample. The second solution is to treat the data from that stratum as though it is from another stratum. In order to implement either solution, one must first identify which strata are affected and which observations in the dataset belong to those strata. The `svydes` command will identify the strata with singleton PSUs by placing an asterisk next to the stratum identifier. For example, in the output below, stratum 1 is identified as having only 1 PSU.

The other possibility (I think) is to use the `subpop` command, which is discussed in another context here:

<http://www.icpsr.umich.edu/icpsrweb/CPES/support/faqs/2011/04/how-should-i-detect-and-handle-single>

I really wish I understood better what kind of estimator this particular "svy" command is using, but I've still not found good documentation describing it, so I can't explain exactly why this is a problem in a mathematical/statistical sense. One other thing people have worried about here is the weighting - since you are only using people who have tested positive for HIV, you are pretending like HIV + is orthogonal to sampling probability, and I'm pretty sure it wouldn't be (because HIV is not distributed randomly across geography and SES class). But I wouldn't think it makes that much difference.

One alternative strategy would be just to give up on the weights and cluster at some larger-than-PSU geographic level - say maybe region if there are many regions (if there are few regions, the wild-t bootstrap would work and I would think you would "cluster" those by strata, because I'm guess that is something like region-by-urban status). Something like:

```
logistic unmetneed i.v106 if hivtest_result ==1, cluster(region)
```

Let me know if this helps.
