

Dear Mayank,

Thanks for the question. Tom and Mahmoud's answer works with the regular DHS. Since you also posted this query on the IPUMS DHS user forum as well, here's a similar response that uses the already-integrated data of IPUMS-DHS.

1. Can i directly use the idhspsu variable to create the requisite variables for the surveys where cluster and PSU are not the same?

Yes.

2. Do i need to use idhsstrata variable while using svyset command?

Yes. svyset would still perform the weighted estimate you do not specify the strata, but the standard errors will be wrong.

To weight IPUMS-DHS data in Stata, the command is:

```
svyset [pw=perweight], psu(idhspsu) strata(idhsstrata)
```

This establishes the weights in Stata; they are then applied to relevant commands by putting "svy:" at the beginning, such as:

```
svy: regress y x  
svy: mean(y), over(x)
```

3. If yes, how do i deal with missing strata information?

This Forum has information on how to construct strata variables when they are missing. Fundamentally, it depends on the sampling design (which you can find in the appendices to the final reports). If the sample was stratified across urban/rural areas (typical), you can replace the strata variable (idhsstrata) with the urban/rural variable (urban).

4. Can i directly use the weight perweight for this analysis?

Yes.

5. In one post on this forum i read that in multi country analysis data must be clustered at country level. Do i need to do that for this analysis. If yes, how do i cluster data at two different levels i.e., country level and then individual psu level?

Whether it's necessary to cluster at the country level, the cluster level, or both depends on how much of the variation in your dependent variable is explained by these spatial variations. You can

calculate this by running a null model, e.g.:

```
logit depvar [pweight = perweight] || idhpsu:  
estat icc
```

If the rho is large (greater than 0.15 or so), then a mixed or multilevel model is appropriate. I've seen people cluster at the country, region, and psu level. These days, the psu level seems to be more common.

If the analysis combines only a few countries, then a dummy variable for each country except one is probably the best approach, and there would be no need to cluster at the country level. To cluster a multiple levels, here are the commands:

```
regress depvar [pweight = perweight] || idhpsu: || country:
```