Subject: Pooled Datasets - Use of Svyset & regional controls
Posted by lukassg on Fri, 03 Jan 2014 13:48:01 GMT
View Forum Message <> Reply to Message

For a research paper we have access to 'Demographic and household surveys' from several different countries taken at different years, e.g. surveys for Uganda for the years 1996, 2002, 2006 and 2011. We then want to pool all surveys from one country to a single dataset.

We experience the following two challenges:

1)
In order to account for the complex survey design we think we have to correctly specify the weights, stratification and clusters for each survey. Even though each survey is from the same country, they can differ slightly depending on the year.
Thus when we pool them, we still want to correctly specify the survey design. However now the question arises how to do it. Before when doing each year by its own, we used code along the following line:

```
gen weight = v005 / 1000000
egen stratid = group (v024 v025), label
svyset [pweight=weight], psu(v021) strata(stratid)
```

The main thing that differs between the surveys is the stratification variables. Sometimes there exists already a stratification variable, sometimes we had to create one like above. Also sometimes the variable v024 (region) for example has 6 values in one year and 10 in the next year. Is it even possible to correctly stratify our dataset when we pool different surveys?

2)

Since we also want to control for regional / community effects later on in our regression models (using svy: reg or svy: logit/clogit) it can be problematic if the defined regions and clusters differ between the surveys.

The only solution we see, is performing single regressions for each year/survey. The drawback is that one cannot directly see whether differences in the constant term or the coefficient of maternal education between the different years/surveys are significant.

Is there any other statistical method that could deal with this dilemma?