Subject: Query on Cluster-Level Modeling with DHS Data and Sampling Weights
Posted by sayianka on Mon, 16 Sep 2024 08:21:46 GMT
View Forum Message <> Reply to Message

Greetings DHS Forum,

I'm working on modeling the prevalence of health outcomes (e.g., diarrhea) as a function of covariates such as literacy and wealth index using DHS data. My approach focuses on cluster-level analysis (v001) with the goal of producing modeled map surfaces similar to the geospatial map surfaces DHS creates for certain indicators.

My current process is as follows:

1. I compute cluster-level proportions for variables (e.g., proportion of women uneducated, proportion from "poor" wealth index) using the weighted mean approach as per DHS guidelines, and sum the total number of children in the cluster eligible (total_children) and the number of "successes" (num_sick_children):

```
ddply(dhs_dataset, ~v001, summarise, mean = weighted.mean(x = my_variable, w = v005))
```

2. For model building, I'm considering a GLM structure like this:

```
glm(cbind(num_sick_children, total_children - num_sick_children) ~ prop_poor + prop_illiterate,
    data = my_dhs_data, family = "binomial")
```

My questions are:

1. In building cluster-level models, how should I utilize the sampling weights (v005)?
   - Should I use the unique v005 per cluster?
   - Should I use the total v005 in the cluster?
   - Or should cluster-level models not use the v005 weighting variable at all?

2. I've noted a quote from a DHS expert in post #9779 in response to #9772, and a related #6672:
   "If you calculate a cluster-level mean, proportion, standard deviation, etc., it will be the same whether or not you use weights. However, for analyses that include the clusters as units, you do need to save the total weight for the cluster."

   How does this apply to my GLM approach? Should I be incorporating cluster weights, and if so, how?

Thank you in advance for your guidance.