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.

Subject: Re: Query on Cluster-Level Modeling with DHS Data and Sampling Weights
Posted by Bridgette-DHS on Fri, 20 Sep 2024 10:41:46 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

You are using R. Below I will paste a simple example in Stata, showing what I would do. The example shows how the weights and number of cases come into play with a binary outcome and a glm model. Substantively, this would not be a good analysis of the data, but it is just intended as an example of the setup. Hope you can convert to R and hope this is helpful.

```
* Example of individual-level and cluster-level analysis with the same variables
* Kenya 2014 DHS survey

use "...KEIR81FL.DTA" , clear

* construct a binary outcome variable for 4+ children
gen     nch4plus=0
replace nch4plus=1 if v201>=4

* construct dummies for wealth quintiles
xi i.v190
rename _I* *

* Individual-level analysis
svyset v001 [pweight=v005], strata(v023) singleunit(centered)
glm nch4plus v190_* , family(binomial) link(logit) eform


* Cluster-level analysis; first switch to clusters as units
gen cases=1
collapse (first) v005 v023 (sum) nch4plus cases (mean) v190_*, by(v001)

svyset [pweight=v005], strata(v023) singleunit(centered)
glm nch4plus v190_* , family(binomial cases) link(logit) eform
```

Subject: Re: Query on Cluster-Level Modeling with DHS Data and Sampling Weights
Posted by sayianka on Sun, 22 Sep 2024 04:33:48 GMT
View Forum Message <> Reply to Message

Thank you for your detailed explanation.
I have successfully recreated the R version of the Stata code.

I needed clarification on this line:

collapse (first) v005 v023 (sum) nch4plus cases (mean) v190_*, by(v001)

I've noted that we use simple means for the variables in this line; Is this meant to be so, or we should weigh the means ?

I've also taken note of your comment: "Substantively, this would not be a good analysis of the data.", as this sentiment has been echoed elsewhere in this forum as well.
The only reason I am attempting this cluster-level analysis is to perform an analysis similar to what the DHS team does for geospatial covariates and modeled map surfaces. According to their report, this analysis must be conducted at the cluster level.

Thank you for your assistance.

Subject: Re: Query on Cluster-Level Modeling with DHS Data and Sampling Weights
Posted by Bridgette-DHS on Mon, 23 Sep 2024 18:55:27 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

When I said "this" would not be a good analysis of the data, I was referring to my own example. I was not judging what you are doing!

In the glm command that I suggest, the outcome is the numerator of a proportion (which is the mean of a 0/1 variable) and the option "family(binomial cases)" specifies the denominator (as "cases"). This is equivalent to a model in which the cluster-level outcome is a proportion and it is weighted by the number of cases in the denominator. A fitted proportion for a cluster would be the fitted frequency divided by "cases".

This part of the collapse command: "(mean) v190_*" will construct five proportions that add to one. These will be the proportions of "cases" that are in wealth quintiles 1, 2, 3, 4, 5. On the right hand side of the estimation command you could include all five of those proportions (as I did); one will be aliased because of the linear constraint (they add to 1). However, I would recommend a recode to a single proportion, such as the proportion in the bottom two quintiles, which will give just one coefficient and be easier to interpret.

You could perhaps include other covariates after the "mean" portion of the collapse command, but you could also just use your geospatial variables.

Hope this helps but let us know if you have questions specifically for the geospatial team.

Subject: Re: Query on Cluster-Level Modeling with DHS Data and Sampling Weights

Posted by Bridgette-DHS on Tue, 24 Sep 2024 14:08:47 GMT
View Forum Message <> Reply to Message

See the following from Senior DHS staff member, Tom Pullum:

A slightly modified version of yesterday's Stata program is attached. Dummies for all 5 of the wealth quintiles (whether you need them or not) are constructed by adding "noomit" to the "xi" command. Also, I show how to construct the observed and fitted proportions with 1 on the outcome variable.

File Attachments

1) multilevel_example_24Sep2024.txt, downloaded 16 times