Subject: Using weights in regression analysis
Posted by DHS user on Wed, 20 Feb 2013 16:48:43 GMT
View Forum Message <> Reply to Message

I am planning to do regression analysis. In your manual you do not suggest to use weights for such analysis. Why is that the case, and would you advice me to use weights? And in case you do, which weight should I use since my unit of analysis is the couple?

Subject: Re: Using weights in regression analysis
Posted by Bridgette-DHS on Wed, 20 Feb 2013 16:50:09 GMT
View Forum Message <> Reply to Message

Here is a response from one of our DHS experts Tom Pullum, that should answer your question.

Future versions of the Guide to DHS Statistics will modify that recommendation. Not using weights is a minority viewpoint here at DHS. Almost all of us now advocate the use of weights. How you use them will depend somewhat on your statistical package. Most of us here use Stata.

If you do not use weights, the coefficients will be biased toward the over-sampled sub-populations.

For the HR and PR files, use hv005, for the IR, KR, and BR files, use v005, for the MR file use mv005. The CR file contains both v005 and mv005. It makes very little empirical difference which you use, but we prefer mv005 because it is adjusted for male non-response, which is typically more serious than female non-response. Some people (e.g. Stan Becker) have proposed a composite couples weight, but as I said the effect of alternatives is trivial.

If you use the AR file, the weight is hiv05, and if you form a couples file using the AR data, the weight is hiv05 for males.

For some purposes it is convenient to divide the weights by 1,000,000, but in Stata, for example, pweight is unaffected by that, and for regressions you use pweight.

You also need to adjust for the clusters (the primary sampling units) and the strata. In Stata that would be done with svyset and svy. These adjustments do not alter the coefficients but they do alter the standard errors, usually in opposite directions.

I hope this helps.

Bridgette-DHS

Subject: Re: Using weights in regression analysis
Posted by enuanand on Wed, 20 Mar 2013 03:32:47 GMT

View Forum Message <> Reply to Message

Thanks for this valuable information, But i want to know one more thing, if we are analyzing domestic violence from couple file then which weight we should use?

Subject: Re: Using weights in regression analysis
Posted by Traore on Wed, 20 Mar 2013 07:54:52 GMT
View Forum Message <> Reply to Message

I would be very glad to know how one can adjust with strata under SPSS.

Subject: Re: Using weights in regression analysis
Posted by Fabrice LOTY on Wed, 20 Mar 2013 09:34:27 GMT
View Forum Message <> Reply to Message

@enuanand: I believe the weighting stuff applies to a particular file regardless of the topic you consider.

Subject: Re: Using weights in regression analysis
Posted by Trevor-DHS on Thu, 21 Mar 2013 00:17:53 GMT
View Forum Message <> Reply to Message

You can use the Complex Samples procedures in SPSS to achieve the same as using svy in Stata. You first need to set up a Complex Sampling Plan using the CSPLAN command (I recommend creating this using the dropdown menu under Analyze, Complex Samples, Prepare for Analysis, and then pasting it into your SPSS syntax. The parameters you typically need are:
Strata: V023 - or alternatively create your strata variable from a combination of V024 and V025.
Clusters: V021 - typically this is the same as V001, but for a few surveys the Primary Sampling Unit (PSU) is different from the final cluster, and the PSU should be used.
Analysis weight: V005 - don't divide by 1000000 as SPSS expects the weight used with Complex Samples to be an integer. Your "population" size will be a million times too big in your results, but just remember to divide it by 1000000 after your analysis. If you use the weight divided by 1000000, SPSS either rounds or truncates your weight to an integer and your analysis will be wrong.
Estimator type: WR (with replacement) - DHS doesn't use replacement sampling, but to match the DHS results this option is needed.

Once you have created your Complex Samples Plan you can then use one of the Complex Samples Procedures for your analysis. I suggest using the CSDESCRIPTIVES first and reproducing the sampling errors shown in the DHS report for one indicator to ensure that you have the CSPLAN set up properly before you try using one of the other CS procedures such as CSLOGISTIC. [Note that DHS uses confidence intervals of +/-2 SEs, whereas SPSS will use +/-1.96 SE for the confidence intervals].

Subject: Re: Using weights in regression analysis
Posted by idas on Fri, 29 Mar 2013 17:59:29 GMT
View Forum Message <> Reply to Message

I plan to do regression analysis as well, but my unit of analysis is the individual. Do I still need to adjust for cluster and strata, after weighing the data?

Do you have example code either from STATA or SAS where you are doing a regression? It would be great if it included the weigh, and if it is relevant in this case, the cluster and strata.

Thank you.

Subject: Re: Using weights in regression analysis
Posted by Reduced-For(u)m on Sat, 30 Mar 2013 23:14:23 GMT
View Forum Message <> Reply to Message

From the DHS FAQs (under "using data files": http://www.measuredhs.com/faq.cfm):

***First, use the svyset command to tell Stata how your data is set up:

*generate weight
generate weight = v005/1000000

*make unique strata values by region/urban-rural (label option automatically labels the results)
egen strata = group(v024 v025), label
*check results
tab strata

*tell Stata the weight (using pweights for robust standard errors), cluster (psu), and strata:
svyset [pweight=weight], psu(v021) strata(strata)

****Now for a regression - if you prefix regress with "svy:" Stata will now know how to weight your data and compute the right standard errors

svy: reg Y X

***Quick note: computing standard errors in this way is probably not OK for a lot of regressions. Without getting off track or all statsy, a good way to think of this is that this standard error calculation is alright IF the error terms and covariates are independently and identically distributed across observations, other than as operating through the sampling procedure (the stratification and clustering prior to randomization that produces the particular sample you have). I tend to think of these standard errors as the smallest the "true" standard errors could possibly be, but I'm kind of on the conservative/stickler end of this debate, and others would surely disagree.

Subject: Re: Using weights in regression analysis
Posted by idas on Tue, 02 Apr 2013 17:15:59 GMT
View Forum Message <> Reply to Message

Thank you very much for your response.

Do you have the code available for SAS?

Subject: Re: Using weights in regression analysis
Posted by Bridgette-DHS on Thu, 11 Apr 2013 21:31:10 GMT
View Forum Message <> Reply to Message

Here is a response from one of our STATA experts Tom Pullum, that should answer your question.

At DHS we mostly use Stata, so I will answer in terms of Stata. The weighting of the data is done as part of the estimation (e.g. regression) command. There is no other sense in which you would "weight the data". For example, instead of "regress y x" you would say "regress y x [pweight=v005]".

You will get the same result if you first say "gen pwt=v005/1000000" and then "regress y x [pweight=pwt]", which some users would prefer to do, but as I said it makes no difference, because Stata always automatically normalizes the weights. Without weights, the estimates are biased toward the oversampled subpopulations and away from the undersampled subpopulations.

The adjustments for clusters and strata will affect the standard errors but not the estimates. If you want to test the significance of the coefficients, you must make those adjustments. For the clusters you expand the above statement to "regress y x [pweight=pwt], cluster(v001)". If you use strata, you must use "svyset" and "svy: regress". The svyset can specify the pweights, clusters, and strata, and then apply them with "svy: regress y x". The svyset command differs slightly across different versions of Stata, e.g. between 11 and 12, so just enter "help svyset" to get the syntax for your version. The strata variable is usually either v022 or v023. However, it is not always labeled correctly. As a general rule, the strata are all combinations of urban/rural and region (the first subnational unit). If the variable labeled "strata" is not consistent with that rule, you should ask someone at DHS to check it.

Subject: Re: Using weights in regression analysis
Posted by myigzaw on Tue, 16 Apr 2013 12:25:27 GMT
View Forum Message <> Reply to Message

The weighted numbers are less than the unweighted numbers in Ethiopian DHS data. Should we still use sample weights to do a regression analysis using DHS HIV data? Don't you think that the power is much lower in the weighted than the unweighted data to do a regression analysis for example? I wonder if weighting is still valuable to analyse HIV data further.

Subject: Re: Using weights in regression analysis
Posted by Reduced-For(u)m on Fri, 19 Apr 2013 05:04:34 GMT
View Forum Message <> Reply to Message

What do you mean by "weighted numbers"? The results you got with weighting (as compared to the results you get when you don't use the weights)?

Subject: Re: Using weights in regression analysis
Posted by Bridgette-DHS on Fri, 26 Apr 2013 14:48:14 GMT
View Forum Message <> Reply to Message

Here is a response from one of our STATA experts Tom Pullum:

"Weighting does not inflate or deflate the number of cases in your analysis. All it does is re-balance them so that under-sampled sub populations are weighted up, and over-sampled sub populations are weighted down, producing estimates of proportions, means, or coefficients that are unbiased. In any kind of regression or test using pweights in Stata, at least, the weights are calculated so that the sum of the weighted cases is exactly the same as the sum of the unweighted cases. You don't have to do anything -- this is automatic.

Standard errors and confidence intervals and statistical tests, in Stata at least, are calculated in a "robust" way with formulas that have been carefully developed. Those things will be more sensitive to whether you take the clusters and strata into account, using svyset and svy, than to whether you use weights. DHS strongly recommends that you make those adjustments."

Subject: Re: Using weights in regression analysis
Posted by mnicolson on Fri, 14 Jun 2013 20:18:11 GMT
View Forum Message <> Reply to Message

Hi, I have a couple of questions about weighting which I was hoping someone might be able to help with?

Weighting, clustering and stratification for regression

An earlier poster has responded by saying:

"From the DHS FAQs (under "using data files": http://www.measuredhs.com/faq.cfm):

***First, use the svyset command to tell Stata how your data is set up:

*generate weight
generate weight = v005/1000000

*make unique strata values by region/urban-rural (label option automatically labels the results)

egen strata = group(v024 v025), label
*check results
tab strata

*tell Stata the weight (using pweights for robust standard errors), cluster (psu), and strata:
svyset [pweight=weight], psu(v021) strata(strata)

****Now for a regression - if you prefix regress with "svy:" Stata will now know how to weight your data and compute the right standard errors

svy: reg Y X

***Quick note: computing standard errors in this way is probably not OK for a lot of regressions. Without getting off track or all statsy, a good way to think of this is that this standard error calculation is alright IF the error terms and covariates are independently and identically distributed across observations, other than as operating through the sampling procedure (the stratification and clustering prior to randomization that produces the particular sample you have). I tend to think of these standard errors as the smallest the "true" standard errors could possibly be, but I'm kind of on the conservative/stickler end of this debate, and others would surely disagree."

I have followed this and all works fine - however, I have two questions.

(1) It seems that the command given above assumes that the data has been collected using one-stage design.

The Stata Manual defines one-stage design as follows:

"A commonly used single-stage survey design uses clustered sampling across several strata, where
the clusters are sampled without replacement."

However, when I read the DHS country manuals, it suggests that samples were selected in two or more stages depending on whether the respondent comes from a rural or urban area. The Stata Manual states that we then have to use a different command, one which accounts for the multiple stages of sampling.

It gives the example:
"We have (fictional) data on American high school seniors (12th graders), and the data were collected
according to the following multistage design. In the first stage, counties were independently selected
within each state. In the second stage, schools were selected within each chosen county. Within each
chosen school, a questionnaire was filled out by every attending high school senior."

The stata command it suggests is:
svyset county [pw=sampwgt], strata(state) fpc(ncounties) || school, fpc(nschools)

Is the command -svyset [pweight=weight], psu(v021) strata(strata)- the correct way of dealing with DHS survey data? Or should I be using a command that takes into account the multiple-rounds used to collect DHS data?

I realise that there is also another factor to consider - namely, whether the clusters are sampled *with* or *without* replacement. Does DHS survey with replacement therefore making it unnecessary to account for the second-stage clusters?

(2) The comment suggests that this way of calculating standard errors ('the way' - accounting for weighting and stratification) won't be appropriate for a lot of regressions. Why is this? If it's not appropriate, does that mean an alternative way is to simply run a regression without weighting or accounting for sampling?

****

Even though I am using the sample weight, my tabulations differ from those in the country tables

I am analysing the India DHS dataset. My unit of analysis is the individual and I have appended all three DHS India datasets into one large dataset.

I am using the following command in order to attempt to replicate the total contraceptive prevalence rate given on p. 170 of the DHS-3 India country report:
tab cpr [iweight=weight] if (v025==0 | v501==1 | year==2005 & 2006)

I created the weight variable using the following command (given above)
generate weight = v005/1000000

cpr is a dummy variable that I have created from v313 (cpr=1 if any method is used; cpr=0 when no method is used)

v025==1 (means that the household type = urban)
v501==1(means that marital status = married)

My tabulation states that the CPR=45.43%
The country table states that the CPR=64%

Could the difference between my figure and the figure in the country table be due to the fact that I have appended the three datasets into one?

Or do you calculate the contraceptive prevalence rate differently from me? If so, how do you do it?

***

My apologies for the length of this post - I hope it all makes sense and I look forward to any responses

Thanks.

Hey there. These are all really good questions. I'll go through them best I can. But first off, just to be clear, I'm not a DHS employee, and have no special insight other than what I've gleaned in my working with the DHS data, discussions with other users, and my general econometrics training. So nothing I say should be taken as the voice of the DHS speaking, or even the advice of some super-expert, just another practitioner trying to figure these things out. With that out of the way...

Weighting, clustering and stratification for regression

(1)...just about everything you say here is new and interesting to me. I had never used the "svy" command before using the DHS - I always weighted and specified standard error calculations manually. What I know about using "svy" for the DHS mostly comes from the DHS FAQ and this paper: http://eprints.soton.ac.uk/8142/

Basically, I have no real insight on the proper use of "svy" to deal with complicated survey designs.

(2) This is one of those times I wish I had said less (it happens), just because leaving something fuzzy like that is probably not helpful. First things first, when it comes to standard errors/inference, we aren't really talking about weighting, we are talking about stratification and clustering. The weighting problem is really just a question of what population you want your results to be representative of (the survey population, or the national population, or the regional population or whatever). Weighting doesn't require the use of "svy".

As for statistical inference (standard errors), to me, one way of thinking of the DHS standard error assumptions is that IF the DHS had used a simple random sampling, then we could just use OLS standard errors (and weight manually with [pweight=weight]). However, in many applications, like difference-in-difference estimation or a cohort fixed-effects-type regression, even with simple random sampling, this is probably an un-conservative technique. Coming from the Labor econ world, I think two really good introductions to the problem are "How much should we trust difference-in-difference estimations" (Bertrand, Duflo, Mullainathan) and "Robust inference with clustered data" (Cameron and Miller). These papers focus on situations where there is likely to be auto-correlation and/or heteroskedasticity in error terms within "clusters" like states or counties (not to be confused with sampling clusters, but some larger grouping of people). As another example, and closer to home, when estimating cohort determinants of HAZ (like, say, effect of month of birth, or the effect of some shock in the birth cohort) I find that the "svy" technique leads to rejection rates on a placebo treatment over 25% (when it should be 5%) and up to like 70% in some cases.

One important caveat though is that all of the things I mention above are uses of the DHS for which it was probably not originally designed. I think that when it comes to things like the effect of maternal age on HAZ, then the DHS method will produce better sized standard errors. I haven't run any placebo tests on that to check implied rejection rates, but you could probably do it fairly easily.

Even though I am using the sample weight, my tabulations differ from those in the country tables

Hopefully some "-DHS" will respond to this, as I have only two (maybe, maybe not) helpful comments and one useless one.

1 - I use "pweight" instead of "iweight". My guess is that it will not make a difference, but these are probability weights (best as I can tell) and since using pweight automatically scales everything to sum to 1, it might make a difference.

2 - Still on weights...when appending multiple rounds, my understanding is that this induces a new weighting problem, as DHS weights within a survey sum to the sample size. So, by just using the given weights, you are not weighting each survey the same, you are implicitly weighting it by the sample size. I'm not sure if that is what you want or not. An alternative would be to re-scale each survey's total weight to sum to one manually, preserving probability of sampling within survey but making each survey have the same total weight (assuming population size is constant, each survey is actually "representing" the same number of women).

```
egen surveytotalweight = total(weight), by(survey)
gen new_weight = weight/surveytotalweight
```

I so far have done that AFTER I dropped all observations that wouldn't go in the regression I use or the statistics I'm tabulating.

3 - I know nothing about replicating DHS tables, so I'll go back to hoping someone "-DHS" responds.

****

I hope this has been in some way helpful. I've been struggling with the weighting thing myself, and how to deal with multiple survey rounds. Truth is, I don't think there are really "perfect" answers out there, and a lot of us are trying to figure things out on our own and doing things in different ways depending on our backgrounds. So my perspective is one that comes from dealing with problems in the Labor Econ world, and Epidemiologists or Nutritionists would have different opinions and different modelling concerns. For example, my field has basically stopped using any random effects models and switched to using an "arbitrary" or "cluster-robust" variance/covariance matrix estimation - I haven't been able to confirm, but I think that the "svy" command uses some weird random-effects-type specification of the V/C matrix. So my biases and "insights" (such as they are) come from that world, and may not be totally appropriate here. These are just my thoughts. I'd love to learn more if someone thinks I'm missing something obvious or important or just fundamentally not understanding something.

Subject: Re: Using weights in regression analysis
Posted by smgwu on Fri, 18 Oct 2013 00:19:10 GMT
View Forum Message <> Reply to Message

This was so very helpful. Thank you!

Can you tell me what I may be doing wrong here regarding the "missing standard errors because of stratum with single sampling unit"?
(unmetneed and hivtest_result are both recoded binary variables)

```
. xi: svy: logistic unmetneed i.v106 if hivtest_result ==1
i.v106            _Iv106_0-3          (naturally coded; _Iv106_0 omitted)
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        20                  Number of obs      =       358
Number of PSUs     =       196                  Population size    = 285.15532
                                                Design df          =       176
                                                F(   0,    176)    =         .
                                                Prob > F           =         .

------------------------------------------------------------------------------
             |             Linearized
   unmetneed | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iv106_1 |   1.276187          .        .       .            .           .
   _Iv106_2 |   .2371095          .        .       .            .           .
   _Iv106_3 |   .0652819          .        .       .            .           .
      _cons |   .0844227          .        .       .            .           .
------------------------------------------------------------------------------
Note: missing standard errors because of stratum with single sampling unit.
```

Any help is appreciated.

Thank you.

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.

Subject: Re: Using weights in regression analysis
Posted by soumava on Wed, 07 Feb 2018 21:49:07 GMT
View Forum Message <> Reply to Message

Hi,
I have a question regarding sampling weight. In my analysis, I have appended individual(IR) and men (MR) file, and merged it with household member (PR) file. I am confused about which weight (mv005 v005 hv005) to use when my analysis includes all eligible men and women?

Thanks,
Soumava

Subject: Re: Using weights in regression analysis
Posted by Bridgette-DHS on Thu, 08 Feb 2018 14:40:34 GMT
View Forum Message <> Reply to Message

The use of weights is described on the following page (step 7):

https://www.dhsprogram.com/data/Using-DataSets-for-Analysis. cfm#CP_JUMP_14042

Subject: Re: Using weights in regression analysis
Posted by Khaing Zar on Fri, 21 Sep 2018 01:14:01 GMT
View Forum Message <> Reply to Message

Dear DHS Program,

I am Ms.Khaing from Myanmar. I used woman individual dataset in order to study the institutional delivery. I chose woman who delivered last child prior the survey. I received 2906 sample after cleaning missing value.When I did weighted for sample. The sample size is reduced to 2740. Please let me know how can I solve this problem.In my understanding, after we had done weighted, the sample will increase or same. Thank you so much.

Best Regards,
Khaing

Subject: Re: Using weights in regression analysis
Posted by Khaing Zar on Sat, 22 Sep 2018 06:31:49 GMT
View Forum Message <> Reply to Message

Dear DHS Program,

I am Ms.Khaing from Myanmar. I used the woman individual dataset in order to study the institutional delivery. The unit of analysis in my study is women who delivered the last child prior to the survey. I received 2906 sample after cleaning missing value. When I did weight for the sample. The sample size is reduced to 2740. Please let me know how can I solve this problem. In my understanding, after we had done weighted, the sample will increase or same. Thank you so much.

Best Regards,
Khaing

Subject: Re: Using weights in regression analysis
Posted by Bridgette-DHS on Mon, 24 Sep 2018 13:16:02 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Specialist, Tom Pullum:

You do not have a problem.

It is normal for the weighted and unweighted totals to differ for a sub-population. The weighted number can be either larger or smaller than the unweighted number, and sometimes by a much larger amount than you found. The weighted and unweighted totals will be the same for the total number of women in the IR file, but not necessarily for subsamples.

Subject: Re: Using weights in regression analysis
Posted by kindu on Sat, 25 Jan 2020 17:09:39 GMT
View Forum Message <> Reply to Message

when should I apply the weighting so if it is not before analysis, um confused of it.
