Subject: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Wed, 19 Mar 2014 15:16:13 GMT
View Forum Message <> Reply to Message

Hello,

First let me start by stating that I am by no means a STATA/statistical software person, and definitely in over my head on this task but it needs to get done. Thus, I apologize for the extremely detailed question/post.

I am attempting the following:
1) Isolate variables on a specific age group(s): 15-19yrs, 15-24yrs, 20-24yrs from the 2010 DHS Malawi STATA dataset.
2) For each age group, I would like a breakdown of:
- Female and Male
- Region of residence
- Wealth quintile
3) I would then want the categories above (#2) to go with the following variables by percent:
- Education (ie. secondary school progression rate, highest grade completed)
- Literacy (ie. Percent literate)
- Employment (ie. Percent employed in last 12 months, Percent in agriculture, control cash earnings)
- Fertility (ie. Percent married, percent married by 18yrs, percent of first birth by 18yrs, Exposed to family planning messages in media, method of contraception)
- AIDS/STI (ie. Percent who had STI in last 12 months)
- Domestic Violence (ie. Percent who experience phys violence since age 15yrs, ever experienced sexual violence)
- Health Behaviors (ie. Percent who use tobacco)

Using Fertility as an example: Among 15-19yrs olds, what percent of female 915-19yrs) are married? What percent of males (15-19 yrs) are married? What is the difference (if any) among wealth quintiles (each for females and males)? What difference (if any) among regions of residence (norther, central, southern)(each for females and males).

I am working with STATA-IC and understand that I need to "extract" the variables because the dataset is too large, then merge them to create one dataset with my select variables.

Please help!!

Thank you in advance.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 19 Mar 2014 16:29:45 GMT
View Forum Message <> Reply to Message

Dear Malayaka,
I urge you to read the FRQ, FRW, and MAP files in the zip file that also contains the data to

determine the names of the variables that you need. The files with extension FRQ, FRW, and MAP can be opened in Notepad (click on the file--> right click--> open with...--> select notepad).

If you are in doubt about what the variables mean, look at the DHS Survey Questionnaire that is at the end of the Malawi Report--Appendix G (Link: http://dhsprogram.com/pubs/pdf/FR247/FR247.pdf). You can also look at the DHS recode manual for even more detail on what the variables mean.

Then, you can follow the steps I posted in an earlier post on how to extract specific variables when you only have Stata IC. Link to the post here: http://userforum.dhsprogram.com/index.php?t=msg&th=778&goto=1294&S=de16df2b6faf6307871c86871edc98b9#msg_1294

Alternatively, you may be able to create those tables in StatCompiler.

hth,
RHS

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Wed, 19 Mar 2014 19:55:52 GMT
View Forum Message <> Reply to Message

Thank you for your reply!

So far I had selected the variables I am interested in for both Women and Men STATA datasets, and saved each separately as as DTA file. Is it possible for me to merge these two (or is it necessary?). The women variables are different than that of the men, v102 and mv102 respectively.

Also, how do I make this "new" dataset specifically only for 15-19, 15-24 and 20-24 year olds? And be able to get the breakdown of my variables (education, fertility, etc) per wealth quintile as well as region?

Again, apologies for the multiple questions.

Thank you in advance.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 19 Mar 2014 20:31:21 GMT
View Forum Message <> Reply to Message

malayaka wrote on Wed, 19 March 2014 15:55Is it possible for me to merge these two (or is it

necessary?).

Depends on what you're trying to answer. If you are trying to to see agreement in response between spouses, then definitely merge these on household unique identifier. If not, keeping them separate should be fine.

malayaka wrote on Wed, 19 March 2014 15:55The women variables are different than that of the men, v102 and mv102 respectively.
As you continue to work with DHS data, you will find that DHS is very good about keeping variable naming conventions so that you can figure out whether the variable pertains to the woman (prefix v), man (prefix mv), household (prefix hv), child (prefix b), maternal questions (m), local variables (s), and so on.

malayaka wrote on Wed, 19 March 2014 15:55how do I make this "new" dataset specifically only for 15-19, 15-24 and 20-24 year olds?
You can use the Stata command -keep- and the age variable to drop observations outside of your age range.

keep if v012>35 will delete everyone who is younger than or equal to 35 from your dataset

*Important: Stata is case-sensitive. All built-in commands are in lowercase. Most user-written commands are also lowercase.

malayaka wrote on Wed, 19 March 2014 15:55and be able to get the breakdown of my variables (education, fertility, etc) per wealth quintile as well as region? The -tab- command is used to get tabulations (and cross-tabulations). syntax is tab rowvbl colvbl

tab v130 v106 gets you the tabulation between religion (row) and education (column):

. tab v130 v106

.....|.....highest.educational.level
 ...religion.|.no.educat....primary..secondary.....higher.|.. ...Total
 -----------+--+-- --------
 ...orthodox.|.....2,836......2,582........910........667.|.. ...6,995.
 ...catholic.|........77.........81..........7.........12.|.. .....177.
 .protestant.|.....1,212......1,366........209........149.|.. ...2,936.
 .....muslim.|.....3,974......1,777........266........153.|.. ...6,170.
 traditional.|........69.........22..........1..........1.|.. ......93.
 ......other.|.......107.........26..........1..........2.|.. .....136.
 ....99.|.........3..........4..........1..........0.|.. ....8.
 -----------+--+-- --------
 ......Total.|.....8,278......5,858......1,395........984.|.. ..16,515

Be aware that missing variables are coded as 99 and you need to recode it into system missing (.) for Stata to recognize it as a missing value. UCLA has an excellent resource on how to get started on Stata (Link: http://www.ats.ucla.edu/stat/stata/sk/), and I encourage you to spend some time on their site to figure out how to do what you need to do.

Stata has great documentation Stata commands. You simply need to type help and the command name you want to find out more about and a window will pop up showing you the syntax, the options for that command, and examples at the bottom of the page.

Again, StatCompiler (http://www.statcompiler.com/) might be able to get you the numbers you need without you having to write any commands, so try that first.

Good luck.

RHS

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Thu, 20 Mar 2014 16:33:50 GMT
View Forum Message <> Reply to Message

Thank you, RHS! I appreciate your help and advice -- very helpful!

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Wed, 02 Apr 2014 04:25:23 GMT
View Forum Message <> Reply to Message

Hello,

I am seeing that my STATA output does not match the Kenya 2010 DHS tables (most are off 1-4%). For example:

Percent of youth ages 20-24 who control their own cash earnings:
DHS: Females (35.2%); Males (47.9%)
STATA: Female (34.36%); Males (50%).

STATA code used:
use v005 v012 v013 v101 v102 v106 v149 v155 v190 v212 v313 v384a v384b v384c v501 v502 v511 v613 v714 v717 v731 v739 v763a v463z using "/Users/malayaka/Desktop/MWIR61FL.DTA
generate wgt = v005/1000000

The DHS website also indicated "tab var [iweight=wgt]" as part of the STATA code, but that gives me an error message stating that the variable is not found.

Am I missing something? Why am I getting different percentages when I crosstab my variables?

Please help!

Thank you in advance.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 02 Apr 2014 16:27:18 GMT
View Forum Message <> Reply to Message

Malayaka,
You need to replace "var" in tab var [iweight=wgt] with the actual variable name. Remember that Stata is case sensitive, i.e. V005 and v005 are two different variables (according to Stata).

One other thing: I wasn't sure why the DHS website told you to use iweight instead of pweight. My understanding is that v005 is the sampling weight, therefore pweight should be used instead of iweight. Note that iweight is equivalent to "weight" in SPSS, which the Stata documentation has labelled as "vague." I hope someone from DHS can clarify this.

HTH,
RHS

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Wed, 02 Apr 2014 16:32:52 GMT
View Forum Message <> Reply to Message

Hi RHS,

Thank you for your reply!

The actual variable, meaning the weight variable? So it will look like this:

use v005 v012 v013 v101 v102 v106 v149 v155 v190 v212 v313 v384a v384b v384c v501 v502 v511 v613 v714 v717 v731 v739 v763a v463z using "/Users/malayaka/Desktop/MWIR61FL.DTA
generate wgt = v005/1000000

tab v005 [iweight=wgt]

I tried running this and I still got the same slightly off results. Yes, the DHS site indicated iweight instead of pweight.

Does it matter that I "extracted" the variables instead of using the entire dataset? I only have STATA IC, so I cannot accommodate all variables.

Thank you again for your ongoing help.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 02 Apr 2014 16:39:03 GMT
View Forum Message <> Reply to Message

Try using pweight instead of iweight and see what you get. Because v005 is a sampling weight, you should use pweight (which specifies it's a sampling weight).

It does not matter that you "extracted" the variables. This is a workaround to the maximum number of variables limitation of Stata IC. Before I learned this trick from an econometrician mentor, I used to do data management in SAS (pull all the variables I need and export to Stata), and using the "use - using" command produces identical results.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 02 Apr 2014 16:41:16 GMT
View Forum Message <> Reply to Message

Note: Stata IC only has limitations on the number of columns/variables you can pull into memory at any given time. There are no restrictions on the number of observations, so extracting just the variables you need won't affect the estimates.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Wed, 02 Apr 2014 17:19:06 GMT
View Forum Message <> Reply to Message

So I ran:

use v005, v012...etc using MWIR61FL.DTA
generate wgt = v005/1000000
tab v005 [pweight=wgt]

Everything went through except for the pweight; it gave me an error "pweight not allowed."

Any ideas?

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Wed, 02 Apr 2014 17:51:52 GMT
View Forum Message <> Reply to Message

When I did a tab of the age groups (v013) I noticed that my STATA numbers reflected the DHS unweighted figures:

Women
Age 15-19 = 5,040 (DHS unweighted); 5,005 (DHS weighted). The DHS tables reflected the weighted figure of 5,005.

However, I ran the generate wgt = v005/1000000 command at the very beginning, yet I still have the unweighted figures.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 02 Apr 2014 18:17:44 GMT
View Forum Message <> Reply to Message

OK, try this. Use the -svyset- command to tell Stata to weight point estimates as follows:

svyset [pweight=wgt]

Use the -svy- prefix before your tabulations

svy: tab v013

This will give you the proportions of women in the dataset that are in a particular age group. If you want the actual numbers, you would specify "count" as an option. For cross-tabulations, decide whether you want the row or column percentages, and specify either row or column as an option.

e.g.

svy: tab v013
svy: tab v013, count /* gives counts instead of percentages */
svy: tab v013 v155, row /*gives the row percentages */
svy: tab v013 v155, col /*gives the column percentages */
svy: tab v013 v155 /*without specifying row or column, percentages are taken out of the total N */

Note that the syntax for tabulation is tab rowvbl columnvbl. Stata will give you an error message if your column variable has too many unique values. So for example if you tried to do literacy by single year age, you will get:

svy: tab v155 v012,row
too many values
r(134);

You should swap it and make v012 the row variable and v155 the column vbl (and switch the specification of row or column percentages as necessary)

svy: tab v012 v155,col

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Wed, 02 Apr 2014 18:20:49 GMT
View Forum Message <> Reply to Message

malayaka wrote on Wed, 02 April 2014 13:51
However, I ran the generate wgt = v005/1000000 command at the very beginning, yet I still have the unweighted figures.

You need to specify the weights for each command, otherwise Stata will assume SRS and not weight it. You can do this either by typing [iweight=wgt] within each command, or using the -svyset- with -svy- prefix that I wrote about in my previous post.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by malayaka on Thu, 03 Apr 2014 16:21:27 GMT
View Forum Message <> Reply to Message

Thank you, RHS for all your help so far! I really appreciate it!

I will go ahead and follow your advice. How/where in the command would I incorporate the -svy- or [iweight=wgt] in the a 3-way/4-way/etc cross tab? Or for commands where I use the -by- command.

For instance:
by mv013, sort: tabulate mv102 mv149, column

I am only looking at 15-24 year olds, so I dropped everyone over 24 yrs old. The mv013 variable is nice because it breaks up the age groups into 15-19 and 20-24. I then do crosstabs within these two age groups.

Thank you, thank you, thank you for your patience and help.

Subject: Re: HELP!: Analysis on youth-specific age group only
Posted by user-rhs on Thu, 03 Apr 2014 21:25:36 GMT
View Forum Message <> Reply to Message

svyset should be executed before you run tabulations, regressions, etc. It's a matter of preference, really. I typically do svyset at the top of my do-file, right after I create the wgt variable by dividing v005 by 100000 (or as instructed by the DHS final report).

After you run svyset, anything you run with the svy prefix will use the svyset that you specified. If you want to change the specification of svyset, you can do svyset,clear and then re-specify svyset with the new settings.

The svy prefix is done with the command you are trying to execute. So for example if you want to

cross-tabulate, do svy: tab variable1 variable2, col. I see you are trying to do the tabs by categories of the variable 'age in 5 year groups.' I don't think you can combine svy with by, so you can just tab for the subpop of interest (the different levels of mv013):

svy: tab mv102 mv149 if mv013==1,col
svy: tab mv102 mv149 if mv013==2,col
svy: tab mv102 mv149 if mv013==3,col
...
and so on.
