Subject: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Lizzynaija on Tue, 16 Jun 2015 14:54:00 GMT

View Forum Message <> Reply to Message

Dear DHS Researchers,

I am analyzing contextual determinants of neonatal mortality using the Nigeria 2013 DHS. I am currently trying to aggregate individual-level statistics to create community-level variables using the Stata collapse command.
I would be grateful if I could get some guidance on how to tackle the following challenges:

1)According to the Stata manual,"collapse" will, by default, use all my observations to calculate the summary statistics; if I want to exclude missing observations for variables, I am to specify the "cw" option. However, when I included this option, Stata returned an error message: "no observations" and I am not sure how to get around this.

2) I want to collapse by PSU (v001) so as to get the community means for my variables. However, I am running into problems with keeping my caseid variable in the collapsed dataset. I need to the caseid variable to stay in the dataset so that I can merge the community level means back onto the original IR dataset. However, when I put it in the by() portion of "collapse", it causes the dataset to collapse by the caseid, and not the PSU.

Below is the code I have been working with:

```
#delimit;
 collapse(mean) commresid=wherelives commregion=region meancommeduc=comm_educlvl
communemp=unemployed
 commpoverty=poverty commpovlevel=povlevel commwealth=v190 commanc=ancvisits
commpostnatal=postchk commsba=birthassist
 commdelivery=birthplace commfemeduc=femeduc commeneduc=meneduc commfemjob=femjob
commenjob=menjob
 commworkprev=workprevyr commfirstmarr=agefirstunion commfirstbirth=matagefirstbirth
commallkids=parity
 commidealkids=idealkidnum commsons=xxsonsalive commgirls=xxgirlsalive
commdecision=all_decision
 commviolence=violence commcontrol=control, by(v001 caseid) cw;
#delimit cr
```

Please what are the steps I should take to get this to work properly?

Thank you very much,
Elizabeth

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Bridgette-DHS on Wed, 17 Jun 2015 13:26:38 GMT
View Forum Message <> Reply to Message

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

If you collapse by v001, you cannot include caseid in the "by" part of the collapse command. You should replace "by(v001 caseid)" with "by(v001)". The collapsed file will have one record per cluster.

caseid is a combination of v001, v002, and v003. They are numeric variables but caseid is a string with embedded blanks.

To merge the collapsed data back onto the individual records in the IR files, you only need to sort both files on v001. However, when I do this I sort the IR file on v001 v002 v003, even though it's not really required. Since your cluster-level file does not contain v002 and v003, they are irrelevant for the merge.

So I recommend lines such as the following:

```
[your sort command]
sort v001
save temp.dta, replace
use IRdata.dta, clear
sort v001 v002 v003
merge v001 using temp.dta
keep if _merge==3
```

Like many Stata users, I prefer the old version of the merge command, but the newer one will also work.

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Lizzynaija on Fri, 19 Jun 2015 02:15:26 GMT
View Forum Message <> Reply to Message

Thank you very much - I have been able to use this successfully!

- Elizabeth

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Lizzynaija on Wed, 24 Jun 2015 00:46:10 GMT
View Forum Message <> Reply to Message

Dear DHS Researchers,

I have another question regarding the use of Stata collapse to aggregate individual-level statistics to create community-level variables.

Bearing in mind the complex survey structure of the DHS, is it necessary to include the pweight option in the collapse command in order to obtain weighted aggregate/summary statistics?

Thank you,
Elizabeth

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Bridgette-DHS on Thu, 25 Jun 2015 15:53:53 GMT

View Forum Message <> Reply to Message

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

If you construct a cluster-level variable using the collapse command, it is not necessary to use weights at all, because everyone in the same cluster has the same weight. To confirm this, you could collapse WITH weights and then collapse WITHOUT weights, and compare the two sets of numbers. They should be exactly the same.

However, if you want to collapse for a larger aggregate, such as a district or region, which includes more than one cluster, you definitely should use weights as part of the collapse.

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Lizzynaija on Thu, 16 Jul 2015 14:16:56 GMT

View Forum Message <> Reply to Message

Thank you for your advice!

I am not working with districts or regions, so I think using the collapse command as it is should be fine for my analysis.

Best regards,
Elizabeth

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Krishna on Mon, 06 Feb 2017 07:38:47 GMT

View Forum Message <> Reply to Message

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Bridgette-DHS on Mon, 06 Feb 2017 14:13:42 GMT

View Forum Message <> Reply to Message

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

I assume you are talking about cluster-level variables that are constructed on the basis of the data on the households or individuals within the cluster. In Stata, cluster-level variables can be constructed either with the "egen" set of commands or with a combination of "collapse" and "merge". I do not know how to do this in SPSS and hope someone else can help.

It would help if you could be more specific about what you want to do, and for which survey.

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Hassen on Sun, 06 May 2018 01:30:08 GMT

View Forum Message <> Reply to Message

Dears all, How to construct community level (cluster level) variables from individual level variables?
Best regards,Hassen

Subject: Re: Keeping caseid in and keeping missing observations out when using Stata "collapse"

Posted by Bridgette-DHS on Mon, 07 May 2018 12:47:00 GMT

View Forum Message <> Reply to Message

Please specify , which file (IR or PR etc.) and which variable(s) you want to construct at the community level.