Subject: duplicate caseid
Posted by adis on Mon, 09 May 2022 06:51:30 GMT
View Forum Message <> Reply to Message

hi would you please forward me the stata command for identifying the duplicate caseid?

Is there any way that we can change the string caseid to numeric? it was important var for our analysis but we failed to change it?

please hepl us

Subject: Re: duplicate caseid
Posted by Janet-DHS on Tue, 10 May 2022 14:13:07 GMT
View Forum Message <> Reply to Message

Following is response from DHS Research & Data Analysis Director, Tom Pullum:

If you are referring back to an earlier posting on the forum, please identify the message and the survey. The ID codes caseid and hhid are strings constructed from v001, v002, and v003 (or similar variables in other files. Those variables are numeric, not strings. You can almost always do what you need to do with the numeric variables, not using caseid or hhid.

Subject: Re: duplicate caseid
Posted by adis on Tue, 10 May 2022 17:18:58 GMT
View Forum Message <> Reply to Message

Thank you Janet again,

I was trying to use the caseid for my analysis but the var is string...I was looking for a var that could replace caseid

What do you suggest me? I tried to create ID but that also didnt work.

What do you suggest me?

The second issue was there are repeated caseid's and how do we overcome the duplicate caseid? shall we drop those duplicate caseid's? that will affect the sample size ? what do you suggest me?

Subject: Re: duplicate caseid
Posted by Janet-DHS on Fri, 13 May 2022 14:59:54 GMT
View Forum Message <> Reply to Message

Following is response from DHS Research & Data Analysis Director, Tom Pullum:

What survey and what file are you using? In the KR file, for example, there is a record for every child born in the past five years. caseid is the mother's ID code. Because many women had more than one child in the past five years, there will be several records with the same value of caseid, but children of the same mother will have different values of bidx (1, 2, etc.). In the IR file, there should never be a repeat of the same caseid, although very rarely we will find a duplicate. To check for duplicates in the IR file, in Stata, enter "gen ncases=1", then "collapse (sum) ncases, by(caseid)", then "tab ncases" and "list if ncases>1, table clean".

Subject: Re: duplicate caseid
Posted by adis on Thu, 19 May 2022 13:13:58 GMT
View Forum Message <> Reply to Message

Thank you so much Janet for all the response you made so far,

I am using the KR file and have the following two questions

First Question

I just want to know how to destring the caseid which is very important variable in my analysis. I just tried the formal ways of de string command but failed to do that.

This was the stata output for the command

destring caseid, gen(Ind_ID)

caseid: contains nonnumeric characters; no generate[/color]

second question

Excluding the non immunized children from the analysis (stata command)

I appreciate swift responses
Thank you

Subject: Re: duplicate caseid
Posted by Janet-DHS on Fri, 20 May 2022 19:45:29 GMT
View Forum Message <> Reply to Message

Following is response from DHS Research & Data Analysis Director, Tom Pullum:

I have to tell you that a comment such as "I appreciate swift responses" will not accelerate our response to a forum question.

You are looking for a way to extract the different columns of caseid (or hhid) and convert them from strings to numeric. The response to a recent forum question (#24358) describes how to do this. "destring caseid, gen(Ind_ID)" will not work because embedded blanks should not be interpreted as zeroes.

Usually, caseid just combines v001 and v002 and v003, and hhid combines hv001 and hv002. You can identify cases just as easily with those components, which are numeric, as with caseid or hhid.

The usual variables in the KR file for having received the basic vaccines are h0, h2, h3, h4, h5, h6, h7, h8, h9, h9a (you should check your survey). These variables are coded 0 if the child did not receive a specific vaccine. You could do something like "drop if (h0+h2+h3+h4+h5+h6+h7+h8+h9a+h9b)==0". I recommend caution with dropping cases from the file. An alternative would be "gen condition=0" and "replace condition=1 if (h0+h2+h3+h4+h5+h6+h7+h8+h9a+h9b)==0". Then you can exclude those cases from a specific command with something like "tab A B if condition==0"

Subject: Re: duplicate caseid
Posted by adis on Fri, 20 May 2022 20:45:08 GMT
View Forum Message <> Reply to Message

Thank you so much Janet; now it seems fine for us.

Your comments are well taken and won't do that again.

Thank you so much
