

---

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.

---