

THE STATA JOURNAL

Editor

H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142; FAX 979-845-3144
jnnewton@stata-journal.com

Associate Editors

Christopher F. Baum
Boston College

Rino Bellocco
Karolinska Institutet, Sweden and
Univ. degli Studi di Milano-Bicocca, Italy

A. Colin Cameron
University of California–Davis

David Clayton
Cambridge Inst. for Medical Research

Mario A. Cleves
Univ. of Arkansas for Medical Sciences

William D. Dupont
Vanderbilt University

Charles Franklin
University of Wisconsin–Madison

Joanne M. Garrett
University of North Carolina

Allan Gregory
Queen's University

James Hardin
University of South Carolina

Ben Jann
ETH Zürich, Switzerland

Stephen Jenkins
University of Essex

Ulrich Kohler
WZB, Berlin

Stata Press Production Manager

Stata Press Copy Editor

Editor

Nicholas J. Cox
Department of Geography
Durham University
South Road
Durham City DH1 3LE UK
n.j.cox@stata-journal.com

Jens Lauritsen
Odense University Hospital

Stanley Lemeshow
Ohio State University

J. Scott Long
Indiana University

Thomas Lumley
University of Washington–Seattle

Roger Newson
Imperial College, London

Marcello Pagano
Harvard School of Public Health

Sophia Rabe-Hesketh
University of California–Berkeley

J. Patrick Royston
MRC Clinical Trials Unit, London

Philip Ryan
University of Adelaide

Mark E. Schaffer
Heriot-Watt University, Edinburgh

Jeroen Weesie
Utrecht University

Nicholas J. G. Winter
University of Virginia

Jeffrey Wooldridge
Michigan State University

Lisa Gilmore
Gabe Waggoner

Copyright Statement: The Stata Journal and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

The articles appearing in the Stata Journal may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the Stata Journal, in whole or in part, on publicly accessible web sites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

Stata tip 50: Efficient use of summarize

Nicholas J. Cox
Department of Geography
Durham University
Durham City, UK
n.j.cox@durham.ac.uk

The `summarize` command must be one of the most commonly used Stata commands. Yet, strangely, one of its options is often not used, even though it can be the best solution to a user's problem. Here I flag this neglected `meanonly` option and speculate briefly on why it is often overlooked.

If you fire up `summarize`, `meanonly`, no results appear in the Results window for you to examine. This lack is deliberate. The option leaves r-class results in memory. (If you are unclear what that means, start with the online help for `return`.) The user must access those results by typing `return list` to see what they are or by feeding one or more results to something else, such as an explicit `display`, `generate`, or `replace` statement. Accessing the saved results should be done promptly after the `summarize`, `meanonly` command has finished, because results are ephemeral and will not survive beyond the next r-class command that is issued.

The `meanonly` option leaves in memory the mean, as the name implies, in `r(mean)`. However, contrary to what you might guess, it also leaves behind the count of nonmissing values, the sum, the weighted sum, the minimum, and the maximum in appropriately named results. These results are for the last-named variable. Thus, although invoking `summarize`, `meanonly` with two or more variables is legal, doing so is utterly pointless because results for all but the last will disappear and machine time will be wasted.

Incidentally, if all you want is a count, the `count` command offers a more direct solution; see [D] `count` and Cox (2007).

The difference between `summarize`, `meanonly` and `summarize` with no options is that the latter also calculates the variance and its square root, the standard deviation. The reason for the `meanonly` option is that this last calculation can be fairly time consuming in large datasets. Thus, if you need to use only one or more of the results left behind after `summarize`, `meanonly`, then specifying the option will be sensible. Programs or do-files that will be used repeatedly and/or on large datasets are especially suitable. Budding programmers can entertain themselves by identifying StataCorp programs that passed up opportunities for using `summarize`, `meanonly`. This issue underscores an old joke that you can always speed up a program that was originally written to run slowly.

As a concrete example, one common task is cycling over a set of categories defined by one or more variables. An easy way to do this is to use `egen`, `group()` to create a variable with integer values 1 and up (and, optionally, value labels with informative text). When you do not know the number of categories in advance,

```
. summarize group, meanonly
```

produces the maximum of `group`, which is the same as the number of categories present. Thus we can feed `r(max)` to whatever code that needs it, possibly a `forvalues` loop.

A small problem remains of explaining why people often overlook this `meanonly` option. I have three guesses. First, `summarize` is one of the commands that Stata users learn early. Typically, it quickly becomes clear that `summarize` does various things and `summarize, detail` does even more. Thus, users tend to feel that they are familiar with the command and do not study its help carefully. Second, the name `meanonly` is in some ways unfortunate and misleading, because much more than the mean is produced. Perhaps a synonym such as `summarize, short` would be a good idea. (Dropping the `meanonly` name is not likely, given the number of programs and commands that would break.) Third, the explanation of `meanonly` in the manual at [R] `summarize` does not give the complete picture on this option.

Reference

Cox, N. J. 2007. Speaking Stata: Making it count. *Stata Journal* 7: 117–130.