The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal

Editor
H. Joseph Newton
Department of Statistics
Texas A&M University
College Station, Texas 77843
979-845-8817; fax 979-845-6077
jnewton@stata-journal.com

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

Associate Editors

Christopher F. Baum
Boston College

Nathaniel Beck
New York University

Rino Bellocco
Karolinska Institutet, Sweden, and
University of Milano-Bicocca, Italy

Maarten L. Buis
Tübingen University, Germany

A. Colin Cameron
University of California–Davis

Mario A. Cleves
Univ. of Arkansas for Medical Sciences

William D. Dupont
Vanderbilt University

David Epstein
Columbia University

Allan Gregory
Queen's University

James Hardin
University of South Carolina

Ben Jann
University of Bern, Switzerland

Stephen Jenkins
London School of Economics and
Political Science

Ulrich Kohler
WZB, Berlin

Frauke Kreuter
University of Maryland–College Park

Peter A. Lachenbruch
Oregon State University

Jens Lauritsen
Odense University Hospital

Stanley Lemeshow
Ohio State University

J. Scott Long
Indiana University

Roger Newson
Imperial College, London

Austin Nichols
Urban Institute, Washington DC

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

Stata Press Editorial Manager Lisa Gilmore
Stata Press Copy Editor Deirdre Skaggs

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

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 websites, fileservers, 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, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 2, pp. 345–346

Stata tip 109: How to combine variables with missing values

Peter A. Lachenbruch
Oregon State University (retired)
Corvallis, OR
peter.lachenbruch@oregonstate.edu

A common problem in data management is combining two or more variables with missing values to get a single variable with as many nonmissing values as possible. Typically, this problem arises when what should be the same variable has been named differently in different datasets. Stata's treatment of missing values means that the combination needs a little care, although there are several quite easy solutions.

To make this situation concrete, consider a recent experience of mine. Two datasets were created. A variable `resp_ill` was created in one dataset, and the corresponding variable `moxResp` was created in the other. After appending the data with the `append` command, there were missing values on `resp_ill` if the data came from the second dataset, and there were missing values on `moxResp` if the data came from the first dataset. The total number of observations is 782. Here is a short listing of some of the observations:

```
. list if inrange(_n, 1, 5) | inrange(_n, 340, 344), separator(0)
```

	resp_ill	moxResp
1.	no	.
2.	no	.
3.	no	.
4.	no	.
5.	no	.
340.	.	no
341.	.	no
342.	.	no
343.	.	no
344.	.	yes

The variables are indicators of whether the patient had a respiratory illness. They have been labeled as "no" for 0 and "yes" for 1.

It is easy to see or to say that the variables should be renamed (with the `rename` command) for consistency. I could have done that in one of the original datasets and then repeated the `append` command. In practice, it is often quicker to fix the problem in the current dataset. Sometimes, maintaining dataset integrity is sufficiently important to rule out the `rename` solution altogether.

Experimentation will confirm that adding the two variables is not the answer, because the missing values are propagated: for Stata, nonmissing plus missing is still missing. Doing that with a `replace` of one variable rather than a `generate` would be

© 2012 StataCorp LP

an even worse idea because any original nonmissing values would be lost in the variable that was replaced.

What we want is to ignore the missings, and there are several ways to do that. One explicit solution is

```
. generate newind = min(resp_ill, moxResp)
```

The pairwise minimum will be whichever nonmissing value is present. What is less intuitive is that the function `max()` will work in the same way because it follows the same principle of ignoring missings to the extent possible, just as, say, `summarize` returns the maximum nonmissing value and not missings. So even though Stata has a general rule that missings count higher than nonmissings, in the case of the `max()` function, that rule is trumped by the principle that missings are ignored as far as possible.

Another solution is

```
. generate newind = cond(missing(resp_ill), moxResp, resp_ill)
```

In this particular case, it was known that the missings in one variable corresponded to the nonmissings in the other. In other circumstances, it would be prudent to check whether there were observations with nonmissing values on both variables, say, by typing

```
. list if var1 < . & var2 < .
```

Another solution that you may know is to use `egen`'s `rowtotal()` function, which by default ignores missings in forming a row total and so gives a single nonmissing result for one missing and one nonmissing argument. Using `egen` can be convenient, especially if you have more than two variables to combine, but it does impose an extra overhead because there is more code for Stata to interpret. For one-off tasks and small- or moderate-size datasets, you would have to strain to detect that overhead, but for repeated data management and large datasets, using `min()` or `max()` is preferable.