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 City 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
ETH Zürich, Switzerland

Stephen Jenkins
University of Essex

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 Patterson

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 web page

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®)
- 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 web sites, 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* (ISSN 1536-867X) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

The Stata Journal (2010)
10, Number 1, pp. 152–156

Stata tip 83: Merging multilingual datasets

Devra L. Golbe
Hunter College, CUNY
New York
dgolbe@hunter.cuny.edu

merge is one of Stata's most important commands. Without specific instructions to the contrary, merge holds the master data file inviolate. Its variables are neither replaced nor updated. Variable and value labels are retained. All these properties and more are well documented. What is not so well documented is how merge interacts with Stata's multiple language support (label language), added in Stata 8.1 and described in Weesie (2005). In essence, Stata users must pay careful attention to which languages are defined and current when merging files.

The language feature is useful not only for multiple "real" languages (e.g., English and French) but also for using different sets of labels for different purposes, such as short labels ("Mines") and long ("Non-Metallic and Industrial Metal Mining") in one data file. merge may generate unexpected results if attention is not paid to the language definitions and, in particular, the current language in each file. Multilingual datasets to be merged should be defined with common languages and each should have the same language set as the current language.

I illustrate using auto.dta. Starting with autotech.dta, create a new dichotomous variable, guzzler, defined as mpg < 25, label the variable and its values in English (en) and French (fr), and save to a file called tech.dta. The tabulations below display variable and value labels:

```
. label language en

. tabulate guzzler

Gas Guzzler |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |         19       25.68       25.68
        Yes |         55       74.32      100.00
------------+-----------------------------------
      Total |         74      100.00

. label language fr

. tabulate guzzler

      Verte |      Freq.     Percent        Cum.
------------+-----------------------------------
        Oui |         19       25.68       25.68
        Non |         55       74.32      100.00
------------+-----------------------------------
      Total |         74      100.00
```

© 2010 StataCorp LP dm0046

From `auto.dta`, create an English-labeled file, `origin.dta`. Now merge in `tech.dta` and tabulate `guzzler` and `foreign`:

```
. tabulate guzzler foreign

           |        Car type
     Verte |  Domestic    Foreign |     Total
-----------+----------------------+----------
       Oui |         8         11 |        19
       Non |        44         11 |        55
-----------+----------------------+----------
     Total |        52         22 |        74
```

`foreign` is labeled in English, but `guzzler` is labeled in French. How did that happen? The `label language` and `labelbook` commands can clarify:

```
. label language

Language for variable and value labels

    In this dataset, value and variable labels have been defined in only one
    language:  en
  (output omitted)
. labelbook
-------------------------------------------------------------------------------
value label origin
-------------------------------------------------------------------------------

  (output omitted)
  definition
           0   Domestic
           1   Foreign

   variables:  foreign

-------------------------------------------------------------------------------
value label yesno_en
-------------------------------------------------------------------------------

  (output omitted)
  definition
           0   No
           1   Yes

   variables:

-------------------------------------------------------------------------------
value label yesno_fr
-------------------------------------------------------------------------------

  (output omitted)
  definition
           0   Oui
           1   Non

   variables:  guzzler
  (output omitted)
```

On reflection, this is what we might have expected. The master dataset is inviolate in the sense that its language—English (and only English)—is preserved. The using dataset has two languages, but only the labels from the current language (French) are attached to the single language (English) in the master dataset.

We could, of course, redefine our languages and reattach the appropriate labels. However, if we plan to merge our master file with a multilingual dataset, a better strategy is to prepare the master file by defining the same two languages and then merge. It is crucial that the current languages are the same in both files. To illustrate, add French labels to `origin.dta`. Then consider what happens if the current languages differ. Suppose that the current language in the master file (`origin.dta`) is French, but the current language in the using file (`tech.dta`) is English. Then the labels from the current language in the using file (English) are attached in the current language of the master file (French), and the French labels (noncurrent) from the using file are not attached at all:

```
. label language fr
(fr already current language)

. tabulate guzzler foreign

       Gas |        Origine
   Guzzler |       USA      Autre |     Total
-----------+----------------------+----------
        No |         8         11 |        19
       Yes |        44         11 |        55
-----------+----------------------+----------
     Total |        52         22 |        74

. label language en

. tabulate guzzler foreign

           |       Car type
   guzzler |  Domestic    Foreign |     Total
-----------+----------------------+----------
         0 |         8         11 |        19
         1 |        44         11 |        55
-----------+----------------------+----------
     Total |        52         22 |        74
```

Although we could of course correct the labels afterward, it is easier to make sure that the files are consistent before the merge. If we set the current language to English (or French) in both files before merging, we see that the labels are properly attached:

```
. label language en
(en already current language)

. tabulate guzzler foreign

       Gas |       Car type
   Guzzler |  Domestic    Foreign |     Total
-----------+----------------------+----------
        No |         8         11 |        19
       Yes |        44         11 |        55
-----------+----------------------+----------
     Total |        52         22 |        74

.  label language fr
.  tabulate guzzler foreign

           │        Origine
     Verte │       USA      Autre │     Total
───────────┼──────────────────────┼──────────
       Oui │         8         11 │        19 
       Non │        44         11 │        55 
───────────┼──────────────────────┼──────────
     Total │        52         22 │        74 
```

Whether or not the labels are properly attached, `merge` does preserve them. But a little planning in advance will ensure that they are attached in the way you expect.

In passing, it is worth noting that the situation is more complex if there are no languages defined in the master file. In that case, issuing a `label language` statement changes the behavior of `merge`. First, let's see what happens if we ignore the language of the master file. Thus we start with the original `autotech.dta` and merge in the multilingual `origin.dta`. We see that two languages have been defined, and the labels are properly attached:

```
. webuse autotech, clear
(1978 Automobile Data)

. merge 1:1 make using origin, nogenerate
  (output omitted)

. tabulate foreign

   Car type │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
   Domestic │         52       70.27       70.27
    Foreign │         22       29.73      100.00
────────────┼───────────────────────────────────
      Total │         74      100.00

. label language fr

. tabulate foreign

    Origine │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
        USA │         52       70.27       70.27
      Autre │         22       29.73      100.00
────────────┼───────────────────────────────────
      Total │         74      100.00
```

If before merging, however, we check that the master file's only language is the default, we get different results:

```
. webuse autotech, clear
(1978 Automobile Data)

. label language

Language for variable and value labels

    In this dataset, value and variable labels have been defined in only one
    language:  default
  (output omitted)
```

```
. merge 1:1 make using origin, nogenerate
  (output omitted)
. label language
Language for variable and value labels
    In this dataset, value and variable labels have been defined in only one
    language:  default
  (output omitted)
```

In this case, only the default language is defined. Tabulation of `foreign` indicates that labels are attached from only the current language in the using file. The reason is that issuing the `label language` command sets the characteristics that define the current language and all available languages. `merge` then declines to overwrite those characteristics with characteristics from the using dataset. If they are as yet undefined, however, those characteristics are taken from the using language.

Reference

Weesie, J. 2005. Multilingual datasets. *Stata Journal* 5: 162–187.