

# 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

## Editor

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

## Associate Editors

Christopher Baum  
Boston College  
Rino Bellocco  
Karolinska Institutet  
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 Zurich, Switzerland  
Stephen Jenkins  
University of Essex  
Ulrich Kohler  
WZB, Berlin  
Jens Lauritsen  
Odense University Hospital

Stanley Lemeshow  
Ohio State University  
J. Scott Long  
Indiana University  
Thomas Lumley  
University of Washington, Seattle  
Roger Newson  
King's 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  
Cornell University  
Jeffrey Wooldridge  
Michigan State University

## Stata Press Production Manager

## Stata Press Copy Editors

Lisa Gilmore  
Gabe Waggoner, John Williams

**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*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

## Stata tip 27: Classifying data points on scatter plots

Nicholas J. Cox  
Durham University  
n.j.cox@durham.ac.uk

When you have scatter plots of counted or measured variables, you may often wish to classify data points according to the values of a further categorical variable. There are several ways to do this. Here we focus on the use of `separate`, gray-scale gradation, and text characters as class symbols. If different categories really do plot as distinct clusters, it should not matter too much how you show them, but knowing some Stata tricks should also help.

One starting point is that differing markers may be used on the plot whenever there are several variables plotted on the  $y$ -axis. With the `auto.dta` dataset, you can imagine

```
. sysuse auto
. gen mpg0 = mpg if foreign == 0
. gen mpg1 = mpg if foreign == 1
. scatter mpg? weight
```

Note the use of the wildcard `mpg?`, which picks up any variable names that have `mpg` followed by just one other character. Once the two variables `mpg0` and `mpg1` have been generated, different markers are automatic. This process still raises two questions. To get an acceptable graph, we need self-explanatory variable labels or at least self-explanatory text in the graph legend. Moreover, two categories are easy enough, but do we have to do this for each of say 5, 7, or 9 categories?

In fact, it would have been better to type

```
. separate mpg, by(foreign) veryshortlabel
. scatter mpg? weight
```

The command `separate` (see [D] `separate`) generates all the variables we need in one command and has a stab at labeling them intelligibly. In this case, we use the (undocumented) `veryshortlabel` option, which was implemented with graphics especially in mind. You may prefer the results of the documented `shortlabel` option. Note that the `by()` option can take true-or-false conditions, such as `price < 6000`, as well as categorical variables.

If your categorical variable consists of qualitatively different categories, you are likely to want to use qualitatively different symbols. Alternatively, if that variable is ordered or graded, the coding you use should also be ordered. One possibility is to use symbols colored in a sequence of gray scales.

Some data on landforms illustrate the point: Ian S. Evans kindly supplied measurements of 260 cirques in Wales, armchair-shaped hollows formerly occupied by small glaciers. Length tends to increase with width, approximately as a power function, but qualitative aspects of form, particularly how closely they approach a classic, well-developed shape, are also coded in a grade variable.

```

. separate length, by(grade) veryshortlabel
. scatter length? width, xsc(log) ysc(log) ms(0 ..)
> mcolor(gs1 gs4 gs7 gs10 gs13) mlcolor(black ..) msize(*1.5 ..)
> yti("': variable label length'") yla(200 500 1000 2000, ang(h))
> xla(200 500 1000 2000) legend(pos(11) ring(0) col(1))

```

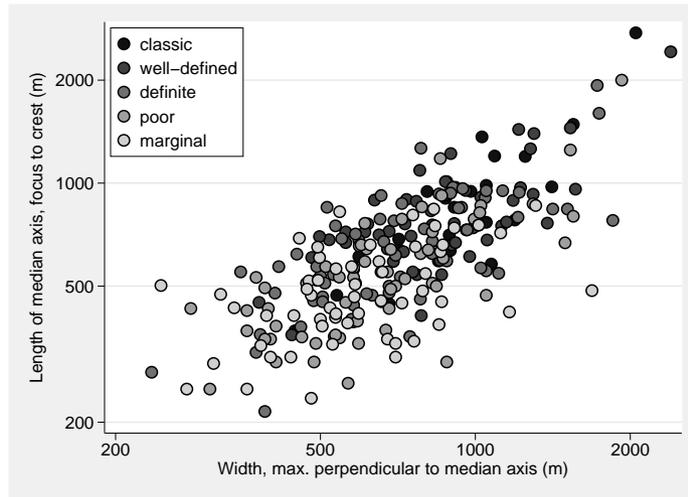


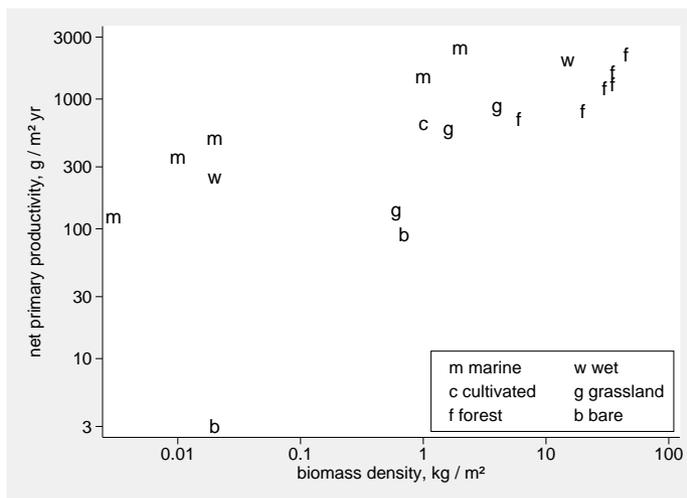
Figure 1 shows length versus width, subdivided by grade. Some practical details deserve emphasis. Gray scales near 16 (white) may be difficult to spot against a light background, including any printed page. Therefore, a dark outline color is recommended. Bigger symbols than the default are needed to do the coloring justice, but as a consequence, this approach is less likely to be useful with thousands of data points. A `by()` option showing different categories separately might work better. With the coding here, it so happens that the darkest category is plotted first and is thus liable to be overplotted by lighter categories wherever data points are dense. Some experimentation with the opposite order of plotting might be a good idea to see which works better.

An alternative that sometimes works nicely is to use ordinary text characters as different markers. One clean style is to suppress the marker symbols completely, using instead the contents of a `str1` variable as marker labels. Whittaker (1975, 224) gave data on net primary productivity and biomass density for various ecosystem types. Figure 2 shows the subdivision.

```

. scatter npp bd, xsc(log) ysc(log) ms(i) mlabpos(0) mlabsize(*1.4)
> mla(c) yla(3000 1000 300 100 30 10 3, nogrid ang(h))
> xla(0.01 "0.01" 0.1 "0.1" 1 10 100)
> legend(on ring(0) pos(5) order( - "m marine" - "w wet" - "c cultivated" -
> "g grassland" - "f forest" - "b bare"))

```



With three or four orders of magnitude variation in each variable, log scales are advisable. On those scales, there is a broad correlation whereby more biomass means higher productivity, but also considerable variation, much of which can be rationalized in terms of very different cover types. For the same biomass density, marine and other wet ecosystems have higher productivity than land ecosystems.

On the Stata side, remember `mlabpos(0)` and note that the `legend` must be set on explicitly. For different purposes, or for different tastes, what is here given as the legend might go better as text in a caption in a printed report. Behind the practice here lies general advice that lowercase letters, such as `abc`, work better than uppercase, such as `ABC`, as they are easier to distinguish from each other, and they are less likely to impart an synaesthetic sense in readers that the graph designer is shouting at them.

## References

Whittaker, R. H. 1975. *Communities and Ecosystems*. New York: Macmillan.