AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

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.

heck for updates

The Stata Journal (2020)
20, Number 2, pp. 489–492 DOI: 10.1177/1536867X20931008

Stata tip 136: Between-group comparisons in a scatterplot with weighted markers

Andrew Musau
Inland School of Business & Social Sciences
INN University
Lillehammer, Norway
andrew.musau@inn.no

Scatterplots are a convenient tool to represent the relationship between two continuous variables. Often, it is necessary to classify this relationship according to values of a third categorical variable. There are several ways to do this (see, for example, Cox [2005]). Here we consider a variation of these graphs, sometimes referred to as bubbleplots, where an additional dimension of the data is represented in the size of the markers. In Stata, one can create such a graph by explicitly specifying a weight in the standard scatterplot syntax (see [G-2] **graph twoway scatter**). As an example, we will use `auto.dta`. Suppose we want to create a scatterplot of mileage and weight with markers weighted by repair record. Suppose further that we want to compare domestic (American) and foreign cars. A problem arises if we do not observe all values of the weighting variable in each of the groups defined by the categorical variable. Consider a cross-tabulation of repair record and car type.[1]

```
. sysuse auto
(1978 Automobile Data)

. tabulate rep78 foreign
    Repair │
    Record │       Car type
      1978 │  Domestic    Foreign │     Total
───────────┼──────────────────────┼──────────
         1 │         2          0 │         2 
         2 │         8          0 │         8 
         3 │        27          3 │        30 
         4 │         9          9 │        18 
         5 │         2          9 │        11 
───────────┼──────────────────────┼──────────
     Total │        48         21 │        69 
```

All five values of repair record are observed for domestic cars, whereas two values are not present for foreign cars. A comparison of scatterplots of mileage and weight with markers weighted by repair record for all cars in the dataset and for groups defined by car type results in the pair of graphs shown in figure 1.

```
. twoway (scatter mpg weight [aweight = rep78], mcolor(black)
> msymbol(smcircle_hollow) text(31 2400 "2", color(black))
> text(41 2240 "1", color(black)) text(18 2550 "4", color(black))
> text(21 2330 "3", color(black)) scheme(sj) legend(on order(1 "All cars")))
```

[1] I thank an anonymous referee for suggesting improvements to the presentation of the problem and proposed solutions.

© 2020 StataCorp LLC gr0084

```
. twoway (scatter mpg weight [aweight = rep78] if foreign==0, mcolor(gs5)
> msymbol(smcircle_hollow))(scatter mpg weight [aweight = rep78] if foreign==1,
> mcolor(gs11) msymbol(smcircle_hollow)
> legend(order(1 "American" 2 "Foreign") row(1))
> text(31 2350 "2", color(black)) text(41 2190 "1", color(black))
> text(18 2560 "4", color(black)) scheme(sj) text(21 2280 "3", color(black)))
```

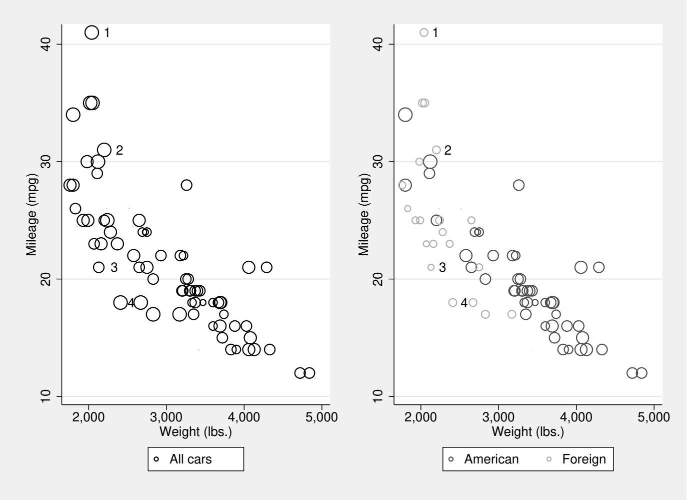

Figure 1. Scatterplot with all values of the weighting variable not present in each group

The size of the weighted markers corresponding to foreign cars is smaller on the graph on the right-hand side, as can be seen from the selection of markers numbered in figure 1. The issue is that Stata internally rescales the weights within groups, thereby precluding between-group comparisons. Note that the problem also arises if the graphs are created with the commonly used `by()` option.

```
. twoway scatter mpg weight [aweight = rep78], by(foreign, total)
> mcolor(gs5 gs11) msymbol(smcircle_hollow smcircle_hollow)
> legend(order(1 "American" 2 "Foreign") row(1))
> text(31 2350 "2", color(black)) text(41 2190 "1", color(black))
> text(18 2560 "4", color(black)) text(21 2280 "3", color(black))
> scheme(sj)
```

The top panel of figure 2 resembles figure 1, while the bottom panel illustrates the two proposed solutions. The first solution is to add "pseudo-observations" to the dataset to ensure that all values of the weighting variable are present in each group of the categorical variable. However, an ensuing concern is that these extra observations will distort the resulting graph. Fortunately, this is not the case if the added observations are missing values for the continuous variables. The command `fillin` (see [D] **fillin**) allows us to achieve this by adding observations with missing data so that all interactions of car type and repair record exist.

```
. fillin foreign rep78
```

A cross-tabulation of repair record and car type will now confirm that each group of the latter includes all values of the former. The second proposed solution, drawing on Cox (2005), is to use the command separate (see [D] **separate**). The underlying mechanics between both approaches are basically the same. For each variable, separate produces missing values in the continuous variable in all but one group, yet the weights are rescaled based on all observations.

```
. separate mpg, by(foreign)
  (output omitted)
. twoway scatter mpg? weight [aweight = rep78], mcolor(gs5 gs11)
> msymbol(smcircle_hollow smcircle_hollow)
> legend(order(1 "American" 2 "Foreign") row(1))
> text(31 2350 "2", color(black)) text(41 2190 "1", color(black))
> text(18 2560 "4", color(black)) text(21 2280 "3", color(black))
> scheme(sj)
```

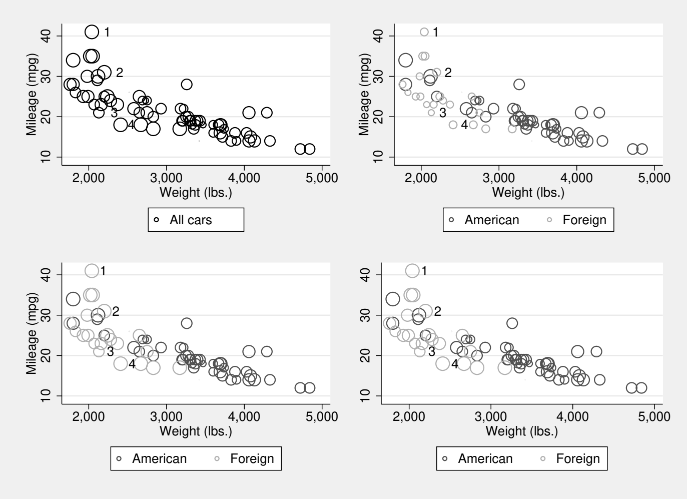

Figure 2. Scatterplots before and after implementing suggested solutions

In summary, the fillin approach adds observations to the dataset corresponding to the number of nonexistent values of the weighting variable across groups defined by the categorical variable. On the other hand, the separate approach adds variables to the dataset corresponding to the number of groups in the categorical variable. While extra observations created by fillin do not distort the graph, they can distort other analyses (for example, as illustrated by the proposed cross-tabulation) and should be deleted once the graphs are created. These are marked by the _fillin variable, which should be used to revert to the original dataset. For separate, the added variables may subsequently be deleted, but their presence in the dataset does no harm. If you use the if qualifier in the graph twoway command, you should use the separate approach because the syntax is shorter. On the other hand, if you are creating graphs by groups defined by the categorical variable, you should use the fillin approach because you can use the `by()` option. To achieve the same result using the `separate` approach, you would need to create one graph at a time and thereafter use the command `graph combine` (see [G-2] **graph combine**) to merge these graphs. Finally, in terms of efficiency, because of maximum size limits, it would appear that adding variables to the dataset is costlier than adding observations.[2] However, because only a few groups can reasonably be differentiated in the scatterplot with weighted markers, comparing the approaches practically, this difference matters little. Therefore, preference for one over the other is a matter of taste.

Reference

Cox, N. J. 2005. Stata tip 27: Classifying data points on scatter plots. *Stata Journal* 5: 604–606. https://doi.org/10.1177/1536867X0500500412.

[2] In Stata/SE 16, for example, the current maximum limit for observations is approximately 2,147 million compared with only 32,767 for variables.