The Stata Journal (2009)
9, Number 3, pp. 499–503

Stata tip 78: Going gray gracefully: Highlighting subsets and downplaying substrates

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

In graphics, as in life, going gray is often forced upon us, yet it is also occasionally a deliberate choice. Journals may enforce publication of your graphs in black and white whenever full-blown color is prohibited, or else prohibitively expensive. Even when allowed, color may prove problematic for various and quite different reasons, ranging from physiology and psychology to sociology and aesthetics. For example, many people are red–green color-blind, while the spectral or rainbow sequence from red to violet is not in fact perceived as a monotonic scale. Wilkinson (2005) and Ware (2008) give good introductions to the use of color in visualization. Fortner and Meyer (1997) give a more detailed discussion. Brewer (2005) gives many specific suggestions on color schemes, which are as appropriate for statistical graphics as they are for cartography.

Choosing differing shades of gray is also worth consideration for positive reasons. Expressing qualitative contrasts just with gray can be both effective and attractive. A previous Stata tip (Cox 2005) showed how distinct values on an ordered scale could be shown separately on scatterplots by markers of different gray-scale colors. This tip expands the theme with further examples. Naturally, black and white themselves qualify as extreme gray shades and may work very well. Here I will emphasize the use of intermediate shades. Let me underline that the examples here use the `sj` scheme. See [G] **schemes intro** for more information.

Consider the highlighting of subsets. You may want to show the distribution of a subset with the distribution of the complete set as context. Unwin, Theus, and Hofmann (2006); Chen, Härdle, and Unwin (2008); and Myatt and Johnson (2009) include several examples of this device for various kinds of graphs.

On a histogram with a frequency scale, this can be done by laying down the distribution of the complete set first and plotting the distribution of the subset on top. Display of the subset can never occlude the display of the complete set, because at most all the observations in any bin belong to the subset.

For example, after reading in a dataset from the U.S. National Longitudinal Study of Young Women in 1988,

```
. sysuse nlsw88
```

we may look at the wage distribution for college graduates compared with the complete set. Figure 1 shows such a histogram.


```
. twoway histogram wage, freq width(1) bcolor(gs14) blw(*.4) blcolor(black)
> || histogram wage if collgrad, yla(, ang(h)) xtitle(hourly wage (USD))
> ytitle(frequency) freq width(1) bcolor(gs6) blw(*.4) blcolor(black) legend(off)
```

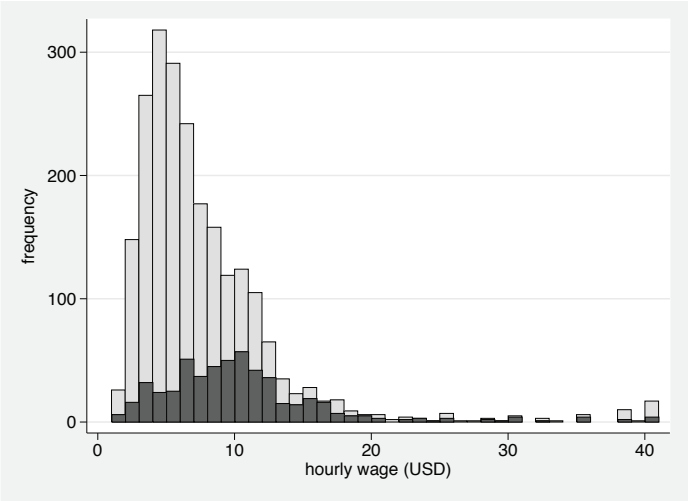

Figure 1. Wage distribution. College graduates are highlighted.

See [G] **graph twoway histogram** if you desire more detail on the command. The important detail here is spelling out that you want the same binning for comparability, as specified by `width()` and—if necessary—`start()`. Otherwise, the choices are matters of taste. For a written report, the legend is arguably dispensable, because what is being highlighted can be explained in the caption that you write within your word processor or text editor, as in this tip. For a talk, the need to make a graph self-explanatory might indicate otherwise.

The same distinction between complete set as backdrop and subset as highlight can be used in other plots. We will look at a scatterplot of wage against educational grade completed. Grade is discrete, but wage is not. We will do our own jittering of grade (only) by adding uniform noise beforehand, if only because that ensures consistency between graphs, and we plot wage on a logarithmic scale. Figure 2 shows a scatterplot with college graduates highlighted once again.

```
. generate grade2 = grade + .5 * (runiform() - .5)
. label var grade2 "`: var label grade'"
. scatter wage grade2, ms(Oh) mc(gs10) ysc(log)
> || scatter wage grade2 if collgrad, legend(off) ms(O) mc(gs2) ysc(log)
> ytitle(hourly wage (USD)) xla(0 4/18) yla(40 20 10 5 2, ang(h))
```

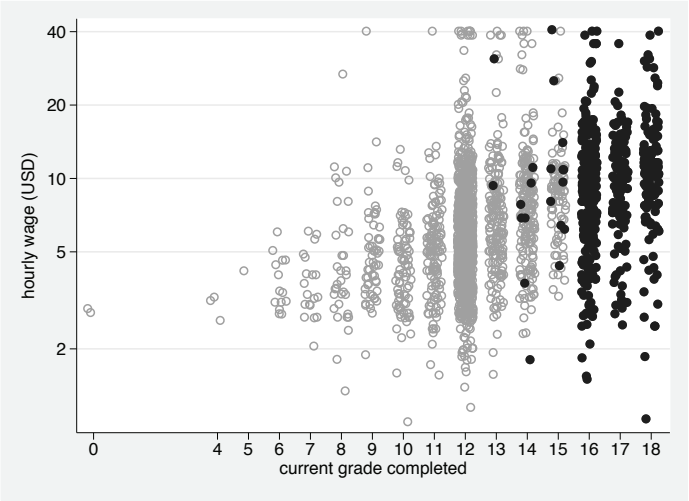

Figure 2. Wage and grade. College graduates are highlighted. Grade is jittered to give a better impression of variation.

Note that the ‘: ’ construct inserts a variable label on the fly; see [P] **macro** for more details. Hollow and filled marker symbols, such as `Oh` and `O`, are helpfully complementary.

Sometimes we want to highlight an exploratory smooth or a fitted model prediction and correspondingly downplay the substrate of the data. With this dataset, noneconomists can join economists in being unsurprised at the great variability of wage within grade. All are likely to be much more interested in the average relationship. Nevertheless, suppressing the data on a graph would often be excessive, if not dishonest. Let us first smooth on a logarithmic scale using restricted cubic splines, experimenting only with default choices; [R] **mkspline** includes details and references. Note that the smooth is calculated for the unjittered grades.

```
. gen lnwage = ln(wage)
. mkspline spline = grade, cubic
. regress lnwage spline?
. predict smooth
```

Figure 3 shows a scatterplot with overlaid smooth. We have to do a little work to get y-axis labels in dollars. See Cox (2008) for further discussion. The logic, however, is easy: we just need to spell out which axis labels we want and where to put them. With this scheme, Stata automatically makes the scatterplot lighter than black, but it seems that we can fairly go further.

```
. scatter lnwage grade2, || mspline smooth grade, xla(0 4/18)
> legend(off) ytitle(hourly wage (USD))
> yla(`= ln(40)' "40" `=ln(20)' "20" `=ln(10)' "10" `=ln(5)' "5" `=ln(2)' "2",
> ang(h))
```

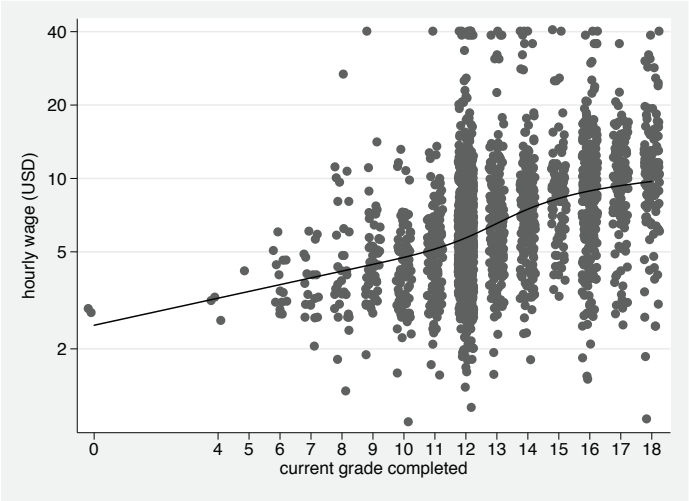

Figure 3. Restricted cubic spline smooth of wage versus grade on a logarithmic scale. Grade is jittered to give a better impression of variation in data.

Figure 4 shows the data points using a lighter gray and increases the width of the smooth. The result is likely to be closer to the researcher's message.

```
. scatter lnwage grade2, mcolor(gs10) || mspline smooth grade, lw(*3) lp(solid)
> xla(0 4/18) legend(off) ytitle(hourly wage (USD))
> yla(`= ln(40)` "40" `=ln(20)` "20" `=ln(10)` "10" `=ln(5)` "5" `=ln(2)` "2",
> ang(h))
```

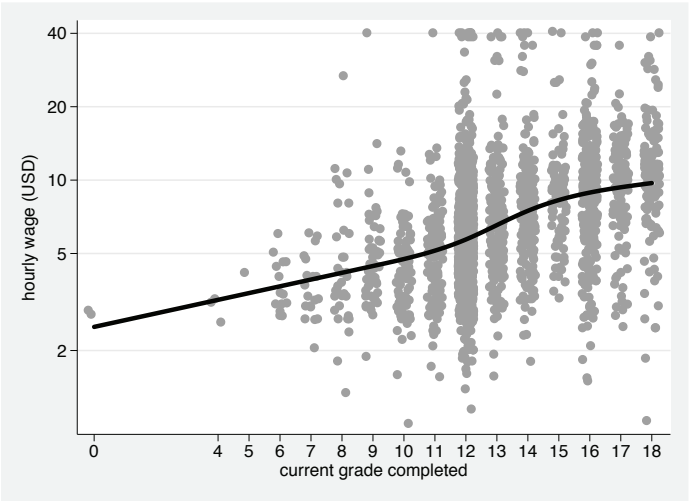

Figure 4. Restricted cubic spline smooth of wage on a logarithmic scale. Grade is jittered to give a better impression of variation in data. Note how the data are downplayed and the smooth highlighted compared with figure 3.

The ideas here can be taken in various further directions. Gray scale can be good for showing the scaffolding of the graph (axes, grids, and so forth) in a subdued but still discernible manner. Highlighting subsets can be extended to show three or more subsets. For example, to show a frequency-based histogram of three subsets, lay down the total distribution of all, followed by the total of two, followed by that of one, so that occlusion produces the desired effect. Alternatively, use `graph bar` or `graph hbar` to produce stacked or subdivided bars, introducing as much or as little histogram style as desired.

References

Brewer, C. A. 2005. *Designing Better Maps: A Guide for GIS Users*. Redlands, CA: ESRI Press.

Chen, C., W. Härdle, and A. Unwin, ed. 2008. *Handbook of Data Visualization*. Berlin: Springer.

Cox, N. J. 2005. Stata tip 27: Classifying data points on scatter plots. *Stata Journal* 5: 604–606.

———. 2008. Stata tip 59: Plotting on any transformed scale. *Stata Journal* 8: 142–145.

Fortner, B., and T. E. Meyer. 1997. *Number by Colors: A Guide to Using Color to Understand Technical Data*. New York: Springer.

Myatt, G. J., and W. P. Johnson. 2009. *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. Hoboken, NJ: Wiley.

Unwin, A., M. Theus, and H. Hofmann. 2006. *Graphics of Large Datasets: Visualizing a Million*. New York: Springer.

Ware, C. 2008. *Visual Thinking for Design*. Burlington, MA: Morgan Kaufmann.

Wilkinson, L. 2005. *The Grammar of Graphics*. 2nd ed. New York: Springer.