Regression Fit Diagnostics Using freqparcoord

Norm Matloff[1*], **Yingkang Xie**[1]

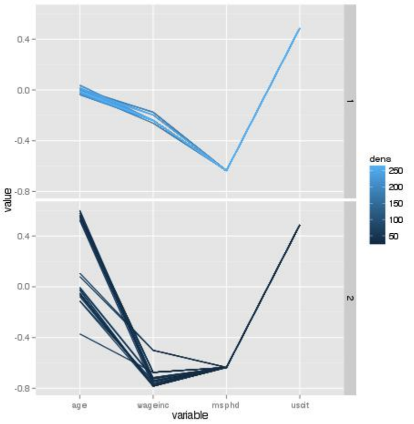
1. University of California, Davis
*Contact author: matloff@cs.ucdavis.edu

Keywords: regression diagnostics, freqparcoord, parallel coordinates

The **freqparcoord** package, available on CRAN, takes a new approach to the parallel coordinates visualization method for multivariate data. Parallel coordinates (Unwin, 2006) is an exploratory method aimed at visualizing interrelations among variables, especially within groups. But it becomes difficult or impossible to use when the number of data points becomes even moderately large, which causes the ”black screen problem,” uninterpretable, dense clutter. This problem is solved in **freqparcoord** by plotting only a few ”typical” lines in the graph, meaning the ones with the highest estimated multivariate density.

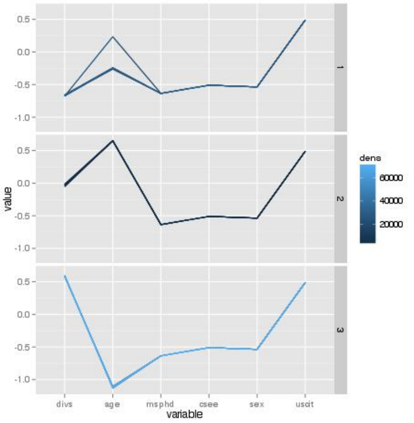
As an example, here is a **freqparcoord** plot of data from the 2000 Census data, for engineers and programmers in Silicon Valley, showing the 25 most typical data points for each gender. Compared to men (upper panel),, we see a much greater range of age among women, with lower wages, but with both genders typically being U.S. citizens with at most a bachelor’s degree.

In the present work, we apply **freqparcoord** to assessing the fit of parametric regression models. The first axis is the “divergences,” the differences beween the parametric and nonparametric estimates of the population regression function, while the other axes are the predictor variables. Note that the divergences are NOT the parametric model residuals, e.g. differences between fitted model values and response (“Y”) values.



The question addressed is, ”In what regions is the parametric fit poorer?” To answer that, the divergences are grouped into upper and lower tails; the default finds the data points that have divergences in the lower and upper 40%, then plots both groups, as well as the middle.

As an example, we fit a linear regression model, predicting wages from age, MS/PhD, CSEE, gender and U.S. citizenship in the Census data. There is a definite trend of overpredicting the young. Moreover, the text output (not shown) finds that the nonparametric R^2 is more than 10% higher than the (adjusted) one from `lm` (though both are low). This suggests adding a quadratic term in age to the model, which then indeed raises the R^2 value to a level similar to the nonparametric one. On the other hand, the graph does not suggest adding any interaction terms.



Any parametric regression model may be used. For instance, in data on graduate school admissions, we fit a logistic model predicting admission from grades, GRE scores and rank. The plot suggested a quadratic effect for grades and possible interactions.

References

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