Other examples of collinearity displays

Baseball data

This example uses another classic data set, the data on major league baseball hitters' performance and salary used in the 1988 ASA Graphics Section poster session, and obtained originally from `lib.stat.cmu.edu/datasets/`. Many analyses of these data were collected in the *ASA Proceedings of the Section on Statistical Graphics* for 1988 and critiqued by Hoaglin and Velleman (1994). We describe below novel views of these data focussed on collinearity.

The complete data set contains 322 observations on 17 numeric variables, including players' salary in 1987 (the response), years in the major leages, batting and fielding performance measures for the 1986 season, and career batting totals. Most previous analyses have identified salary as highly skewed and heteroscedastic and therefore constructed models for **logsal** = log(salary). Years in the major leagues has a nonlinear relation to salary, increasing linearly up to about seven years, then levelling off. We modeled this as the piece-wise linear function, **years7** = $\min(years, 7)$.

For the present purposes, we focus on the career performance statistics, that show the greatest degree of collinearity: **atbatc**: career times at bat, **hitsc**: career hits, **homerc**: career home runs, **runsc**: career runs scored, **rbic**: career runs batted in, **walksc**:career walks. The model predicting logsal with these seven predictors fits reasonably well ($R^2 = 0.594$), however all predictors except years7 have extremely large VIFs, ranging from 34.7 (homerc) to 156.5 (atbhatc).

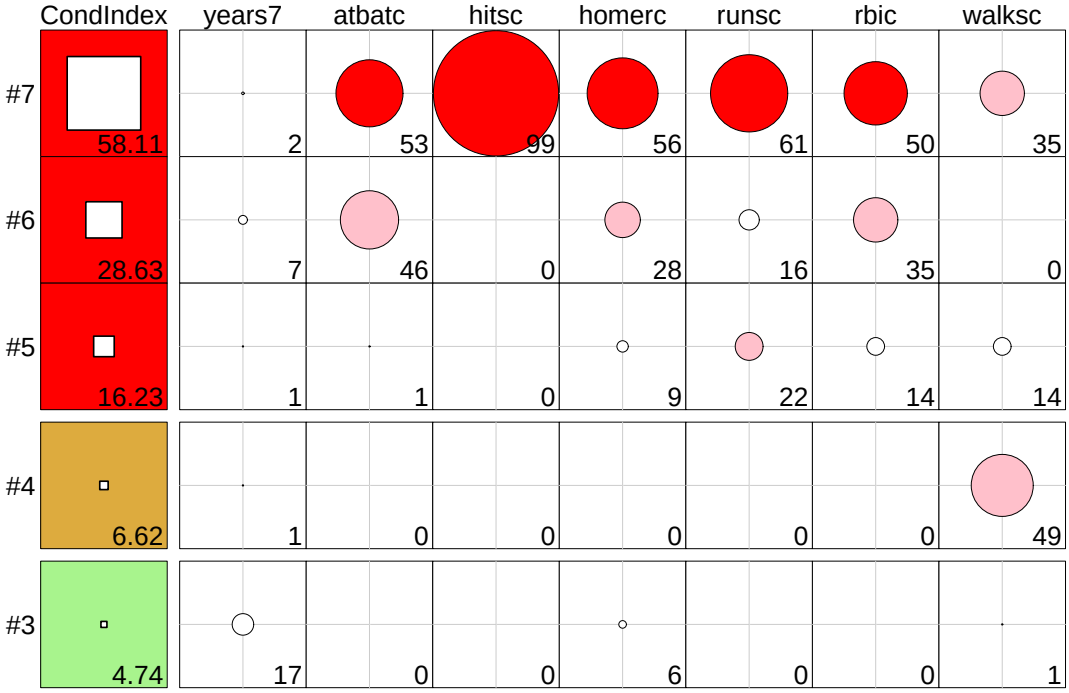

Figure 1: Tableplot of condition indices and variance proportions for the career variables in the Baseball data. In column 1, the square symbols are scaled relative to a maximum condition index of 80. In the remaining columns, variance proportions ($\times 100$) are scaled relative to a maximum of 100.

Figure 1 shows a tableplot of the five largest condition indices and associated variance proportions. There are three large condition indices, but only the largest two have substantial contributions from two or more predictors. In particular, the largest condition index (58.11) is dangerously high, and has large variance proportions associated with all predictors except years7.

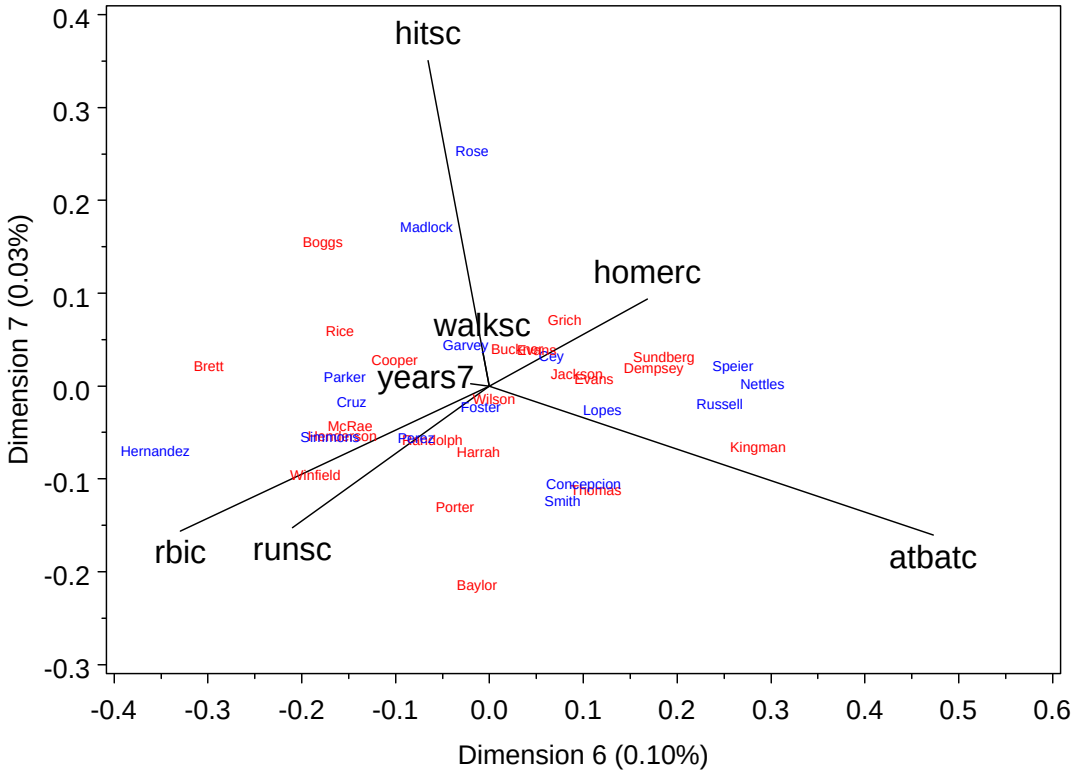

Figure 2: Collinearity biplot of the Baseball data, showing the last two dimensions. The point labels have been thinned to include only the 37 observations with the most extreme robust D^2 values in predictor space, corresponding to $\Pr(\chi^2_7) < 0.001$.

The collinearity biplot for this model (Figure 2) makes the nature of the contributions of the variables to collinearity apparent. These two dimensions account for only 0.13% of the variation in the predictor space, yet the lengths of the variable vectors reflect the degree to which each variable is redundant with the others. For example, hitsc and atbatc have R^2s of 0.997 and 0.993 against the other predictors

References

Hoaglin, D. C. and Velleman, P. F. (1994). A critical look at some analyses of major league baseball salaries. *The American Statistician*, 49, 277–285.