Collinearity biplots and related displays for outliers

Figure 1 shows the biplot of the Cars data for the smallest two dimensions— what we can call the *collinearity biplot*. The projections of the variable vectors on the Dimension 5 and Dimension 6 axes are proportional to their variance proportions. The relative lengths of these variable vectors can be considered to indicate the extent to which each variable contributes to collinearity for these two near-singular dimensions.

Moreover, there is one observation: #20 (a Buick Estate wagon), that stands out as an outlier in predictor space, far from the centroid.[1] This and other high-leverage observations may be seen in other

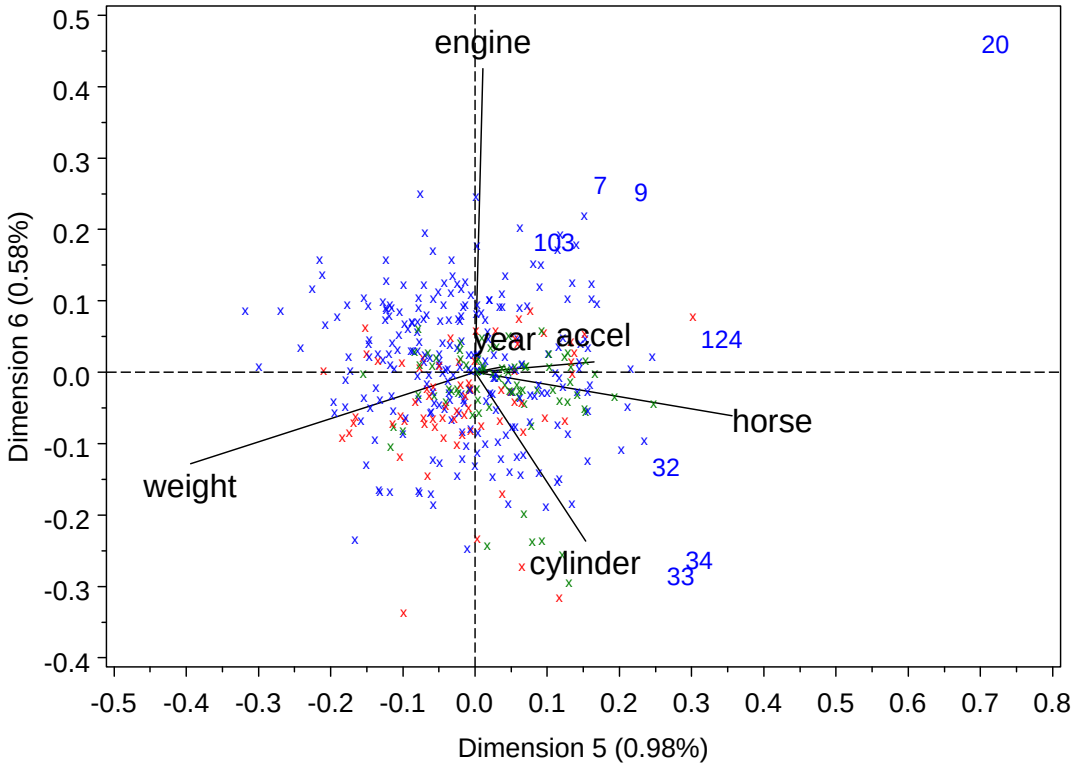

Figure 1: Collinearity biplot of the Cars data, showing the last two dimensions. The projections of the variable vectors on the coordinate axes are proportional to their variance proportions. To reduce graphic clutter, only the eight most outlying observations in predictor space (see Figure 2, left) are identified by case labels. An extreme outlier (case #20) appears in the upper right corner.

graphical displays; but it is useful to know here that they will often also be quite noticeable in what we propose here as the collinearity biplot.

For example, Figure 2 shows two related graphical displays that shed light on the outlier seen in the collinearity biplot: The left panel is a robust outlier-detection QQ plot (Friendly, 1991), plotting robustified squared Mahalanobis distances (D^2) of the observations in predictor space against the corresponding χ^2_6 quantiles.[2] The right panel is an influence bubble plot for the model predicting

[1] It turns out that this vehicle is an early-year (1970) American behemoth, with an 8-cylinder, 455 cu. in, 225 horse-power engine, and able to go from 0 to 60 mph in 10 sec. As can be seen from Figure 2 (right), its' MPG is only slightly under-predicted from the regression model.

[2] The robust method used here is simple iterative multivariate trimming (Gnanadesikan and Kettenring, 1972), whereby observations for which the χ^2_p quantiles of their D^2 having improbably small p-values are given 0 weight in a new computation of the mean vector and covariance matrix on which the calculation of D^2 is based. This panel labels the eight points

MPG from the other variables, showing studentized residual on the ordinate against leverage on the abscissa, with the area of the bubble symbol proportional to Cook's D (Cook, 1977, 1979) influence statistic. In all these displays, observation 20 is clearly discrepant, but we can also see several other observations (e.g., cases 9, 33, 34, ...) that are also distinguished in these plots.

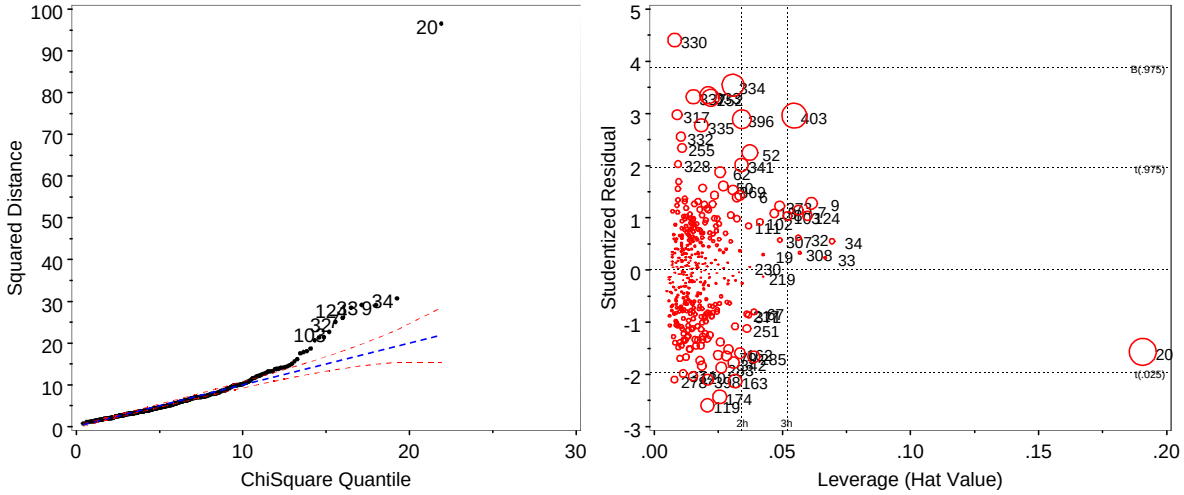

Figure 2: Left: Robust outlier plot of the Cars data. The dotted envelope around the dashed reference line of equality gives a 95% confidence band for a multivariate normal distribution. Right: Influence plot for the model. The horizontal and vertical reference lines give standard cutoffs for leverage and studentized residuals. Bubble size is proportional to Cook's D statistic, predicting MPG.

References

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365), 169–174.

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28, 81–124.

for which $\Pr(D^2) < .001$ after two such iterations.