Collinearity

Rolf Sundberg

Volume 1, pp 365–366

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Collinearity

Collinearity refers in a strict sense to the presence of exact linear relationships within a set of variables, typically a set of explanatory (predictor) variables used in a regression-type model. In common statistical language it also allows *near-collinearity*, when variables are close to linearly related, that is when their correlation matrix **C** is near-singular. In other words, data are ill-conditioned. The 'collinearity problem' refers to the fact that in a multiple regression with collinearity, least squares regression coefficients are unstable (sensitive to small changes in input data) or not even uniquely defined, have highly inflated variances, and are impossible to interpret individually. In any generalized linear modeling, an analogous collinearity problem will appear.

A frequent cause of collinearity is that the explanatory variables represent the composition of a (chemical, say) sample, so they sum to or near 100%. In experimental or observational studies, confounded variables are also collinear. Another frequent situation is that an (automatic) measuring instrument registers a large set of variables, for example a spectrum of 100 or 1000 wavelengths in a spectrophotometer to be calibrated, whereas the number of samples available is much more limited, for cost or time reasons. Then there must be exact collinearities. However, even if the number of samples had been sufficiently large to avoid such collinearities, we should have expected a large number of near-collinearities in the spectral data, simply because the sources of variation represented would not be so many. The 'chemical rank' of the situation would be smaller than the number of variables. Quantitative structure–property relationships (QSPR) studies give other example types that may show a high degree of collinearity. For example, toxicity may be determined for a small set of different dioxins and modeled by a large set of physical–chemical descriptors at molecular level, in order to form a predictor that can be used on not yet toxicity-tested dioxins.

For prediction, collinearity is not necessarily harmful. Say that we study how toxicity depends on the concentrations of some compounds, which in nature tend to appear together in certain proportions. From natural samples, it will then be difficult to distinguish their respective toxicity and to see if they interact, so collinearity is a problem for understanding relationships. However, this does not necessarily make it impossible to predict well the toxicity of a new sample, provided that the compounds appear in about the same proportions. The crucial question is whether the collinearity is inherent in the type of data studied, or worse, if it is only generated by design. Anyhow, it is not advisable to use ordinary least squares for predictor construction under collinearity. There are many better methods for constructing such predictors, under the titles of regularized or shrinkage estimators, and including ridge regression, PCR and PLS regression (*see* **Shrinkage regression**).

Mathematically, exact collinearities can be eliminated by reducing the number of variables, and near-collinearities can be made to disappear by forming new, orthogonalized and rescaled variables. However, omitting a perfect confounder from the model can only increase the risk for misinterpretation of the influences on the response variable. Replacing an explanatory variable by its residual, in a regression on the others from a set of near-collinear variables, makes the new variable orthogonal to the others, but with comparatively very little variability. Owing to this lack of substantial variability, we are not likely to see much of its possible influence. Mathematical rescaling cannot change this reality, of course. To a large extent it is a matter of judgement what should be called a comparatively large or small variation in different variables and variable combinations, and this ambiguity remains when it comes to shrinkage methods for predictor construction, because these are not invariant under individual rescaling and other nonorthogonal transformations of variables. We could say that the 'near' in near-collinearity is not scale invariant.

To collect more data does not necessarily eliminate the collinearity problem. If the new data are representative of the same population as the old data, they are likely to show the same types of collinearity.

The degree of **multicollinearity** in a set of variables can be investigated by principal components analysis (PCA) of their sample correlation matrix **C**. Small eigenvalues indicate near-collinearities, and the corresponding principal components are the near constant linear combinations of the variables. As a rule of thumb, Hocking [1] suggests that if the smallest eigenvalue is less than 0.05, this is an indication of serious collinearity, and less than 0.10 indicates

moderate collinearity. The condition number of **C** (the ratio between the largest and smallest eigenvalues) is a related diagnostic, but of more numerical than statistical relevance. For the individual variables, collinearity may be diagnosed by the variance inflation factors (VIF), which are the diagonal elements of $\mathbf{C}^{-1}$, indicating how much the variance of the regression coefficient for a single variable is inflated by the presence of the other, correlated variables (*see* **Regression diagnostics**).

Not all statistical packages are good at indicating collinearity. Such diagnostics are typically only found in their linear regression programs, if at all. To mention some examples, SPSS and **SAS** (PROC REG) have options COLLIN, yielding various characteristics, whereas MINITAB and STATA go for VIF values.

Reference

[1] Hocking, R.R. (1996). *Methods and Applications of Linear Models*, Wiley, New York.

(*See also* **Linear models**)

ROLF SUNDBERG