

Multicollinearity

Collinearity in a regression framework refers to the situation where one of the columns of the design matrix (i.e. one of the predictors in the model) is linearly dependent on other columns. The linear dependency may be caused by one predictor being the exact duplicate of another predictor, or by one predictor being equal to a linear combination of other predictors (columns in the design matrix).

If the model

$$Y = X\beta + \varepsilon$$

is to be fitted, the solution generally would be of the form

$$b = (X'X)^{-1} X'Y$$

Obtaining the solution thus depends on obtaining the inverse of the matrix product $X'X$. If this product is singular (and do not have a unique inverse) an infinity of solutions exists.

However, the terms collinearity and multicollinearity are also commonly used to refer to situations where the above "almost" happens- in other words, when a predictor (or more than one predictor) is almost the duplicate or linear combinations of other predictors. As such, these terms can usually refer either to exact dependency or to near dependency in the design matrix.

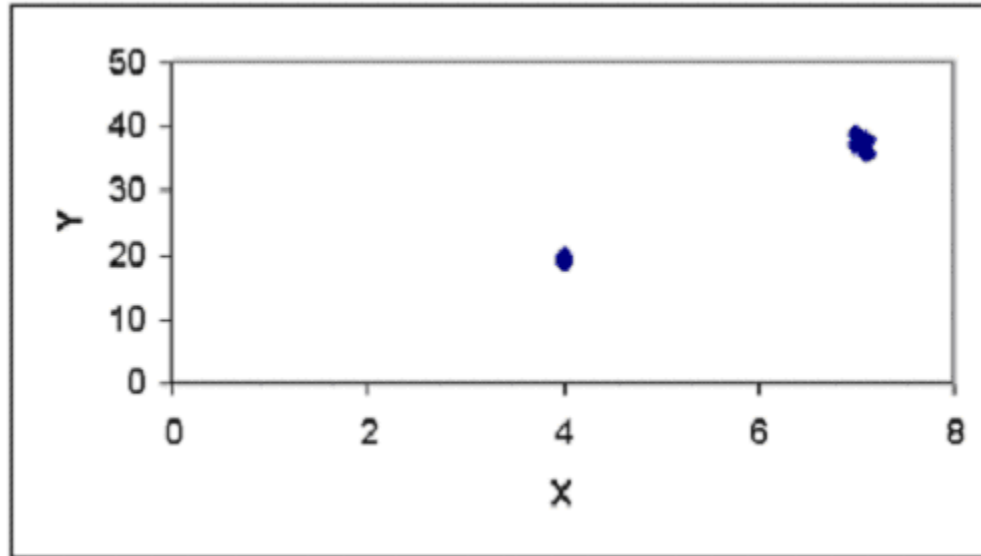
A simple example illustrating this is given in Chapter 16 of the 3rd edition of *Applied Regression Analysis* by Draper & Smith (Wiley, 1998).

Consider the two variables X and Y:

X: 4 4 7 7 7.1 7.1

Y: 19 20 37 39 36 38

When these data are plotted, the following scatterplot is obtained:



The data are clustered in two distinct groups - the X-values 7 and 7.1 being virtually indistinguishable. Where one would usually be able to fit a quadratic model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

With these data, however, the $X'X$ matrix is close to singular. Although it is possible to fit a straight line to these data, the data are completely inadequate for a quadratic fit.

In the context of a hierarchical linear model, the existence of high correlations between variables is a well-known cause of instability in the model. The correlations causing the problem may, however, be between two predictors at the same level, or it may be that a cross-level interaction is highly correlated either with the second-level variable or with the first-level variable, or both.

One way of dealing with this problem is by centering predictors entered into the model. In particular, centering of level-1 predictors around the respective group means may lower some of the correlations among the variables involved. When group mean centering is used, the correlations between second-level variables and both first-level variables and cross-level interactions are equal to zero, so that only the correlations between cross-level interactions and level-1 variables remain as a potential source of estimation problems.

The impact of high correlations on the numerical stability is also a function of the total amount of information in the actual data set used. Note, however, that centering impacts the interpretation of results and should be used with caution.

In HLM, most reports by users concerning error messages noting collinearity/multicollinearity are caused by

- Near collinearity between a predictor with little or no variation and the intercept term, which is represented in the design matrix by a column of 1's.
- Fitting quadratic growth curves to a very short series of points so that a situation similar to that described above is the result.

The best place to look should HLM print an error message warning about collinearity / multicollinearity in the random part of the model is in the Tau matrix given in the output file. Check all off-diagonal elements for correlations close to 1 or -1. Also check the diagonal elements of the Tau matrix for any elements close to zero, as this may indicate that there is no indication of random variation in this slope.