Print

Exploring Collinearity

Collinearity, also called multicollinearity, is a potential problem in multiple regression. It occurs when two or more predictor variables are highly correlated with each other. For example, suppose your model contains three variables: the response variable, weight, and the predictors, height in inches and height in centimeters. Are the two predictors completely redundant? Yes. Height is the same variable, whether measured in inches or centimeters. They are both trying to explain the same variation in weight. Strong correlations between pairs of predictors can be detected with correlation analysis.

But what about strong correlations among sets of several predictors? For these, bivariate correlations might fall short. Imagine a high-school student receives a final grade for a course (F) that is equal to the sum of his midterm (M) and homework (H) grades. Bivariate correlations between F, M, and H might reveal only moderate correlations. However, F, M, and H together are highly collinear. If you know any two of the three, the remaining variable can be predicted perfectly. This collinearity would make the parameter estimates unstable and the standard errors large if all three were used as predictors in the model.

Because collinearity involving several predictors can be missed by correlations, we need additional tools for collinearity detection such as Variance Inflation Factors.

Collinearity doesn't violate the assumptions of multiple regression. It means that there's redundant information among the predictor variables. So, why is this a problem? When multiple variables try to explain the same variation in the response, it leads to inflated standard errors and instability in the regression model. So, you want to avoid having two highly correlated predictor variables in the same model.

Close