Multicollinearity

The term collinearity, or multicollinearity, refers to the condition in which two or more predictors are highly correlated with one another. We touched on the issue with collinearity earlier.

In a regression context, collinearity can make it difficult to determine the effect of each predictor on the response, and can make it challenging to determine which variables to include in the model.

Collinearity can also cause other problems. The coefficients might be poorly estimated, or inflated. The coefficients might have signs that don't make sense. And the standard errors for these coefficients might be inflated. For illustration, we take a look at a new example, BodyFat. This data set includes measurements of 252 men. The goal of the study was to develop a model, based on physical measurements, to predict percent body fat. We focus on a subset of the potential predictors: Weight (in pounds), Height (in inches), and BMI (Body Mass Index). Note that we have hidden and excluded a couple of observations for this illustration. We'll return to these observations, and the full data set, in an exercise.

Weight is highly correlated with BMI, and is moderately correlated with Height. We fit a model to predict Fat% as a function of these three variables. BMI and Weight are highly significant, and Height is borderline significant. But BMI is a function of both Weight and Height. So, there is some redundant information in these predictors.

What happens if we remove BMI from the model?

Notice how the parameter estimates for Weight and Height have changed. The coefficient for Weight changed from negative to positive. The coefficient for Height changed from positive to negative. Both Weight and Height are also now highly significant. Another dramatic change is in the accuracy of the estimates. The standard errors for Weight and Height are much larger in the model containing BMI. When we fit a model, how do we know if we have a problem with collinearity? As we've seen, a scatterplot matrix can point to pairs of variables that are correlated. But collinearity (or multicollinearity) can also occur between many variables, and this might not be apparent in bivariate scatterplots. One method for detecting whether collinearity is a problem is to compute the Variance Inflation Factor or VIF. This is a measure of how much the standard error of the estimate of the coefficient is inflated due to multicollinearity. The VIF for a predictor is calculated using this formula. For a given predictor variable, a regression model is fit using that variable as the response and all the other variables as predictors. The RSquare for this model is calculated, and the VIF is computed. This is repeated for all predictors. The smallest possible value of VIF is 1.0, indicating a complete absence of collinearity.

Statisticians use the term orthogonal to refer to variables that are completely uncorrelated with one another. A VIF for a predictor of 10.0 corresponds to an RSquare value of 0.90. Likewise, a VIF of 100 corresponds to an RSquare of 0.99. This would mean that the other predictors explain 99% of the variation in the given predictor. In most cases, there will be some amount of collinearity. As a rule of thumb, a VIF of 5 or 10 indicates that the collinearity might be problematic.

In our example, the VIFs are all very high, indicating that collinearity is indeed an issue. After we remove BMI from the model, the VIFs are now very low. In some cases, collinearity can be resolved by removing a redundant term from the model. In more severe cases, simply removing a term will not address the issue. In these cases, more advanced techniques such as principal component analysis (PCA) or partial least squares (PLS) might be appropriate. Other modeling approaches, such as tree-based methods and penalized regression, are also recommended.

We introduce tree-based methods and penalized regression techniques, such as ridge regression, in the Predictive Modeling module.

Statistical Thinking for Industrial Problem Solving

Close