Print

Visualizing Collinearity

You've seen that when variables are collinear, one of the variables provides nearly the same information as the other. Consider an example.

X_1 and X_2 are collinear, so they follow a reasonably straight line. When the model includes both variables, neither one might be significant. However, when the model includes only one of them, either variable might be significant. This means that collinearity can hide significant effects, and this is a good reason to deal with collinearity before using any automated model selection tool.

Second, collinearity increases the variance of the parameter estimates, which make them unstable. In turn, this increases the prediction error of the model. Let's explore this instability. Remember that you're modeling the relationship of the three variables using a plane of predicted response values. Where should the prediction plane lie?

This represents a best-fit plane through the data. (The plane rests mostly over all data points.)

Now think of a table. The plane that you're trying to build is like the tabletop. The observations guide the angle of the tabletop relative to the floor, like the table legs. What happens if the legs line up along the center of the table?

The tabletop is unstable. In this model, collinearity between the two predictors, X_1 and X_2, means that their data points don't spread out enough in the X space to provide stable support for the plane. Instead, the points cluster around the center, which makes the plane unstable. In fact, if you remove only one data point, the model's instability drastically changes the plane of predicted response values.

This is the resulting plane. (The plane changed angle drastically.) The slopes changed in magnitude and one of the slopes even changed its sign. Even if you only move a data point, the plane can shift considerably. So, collinear predictors provide redundant information and destabilize the model.

How do you solve this problem? In this case, you can delete either X_1 or X_2 from the model, because they both measure essentially the same thing.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Close