# Multicollinearity: effects and fixes

What is multicollinearity?

Multicollinearity occurs when two or more independent variables are correlated in a regression model. This means that one variable can predict another. Could be something like height and weight. This often presents problems in fitting models. When doing regression analysis, one of the main goals is to isolate the relationship between each independent variable and the dependent variable. Multicollinearity can cause problems with trusting the regression coefficient, as it represents the average change in the dependent variable in relation to each change of 1 unit in an independent variable, when all other independent variables are held constant. Multicollinearity is a problem here, as it would mean that changes in one variable are associated with changes to another. There are two different kinds of multicollinearity: Structural multicollinearity and data multicollinearity. Structural multicollinearity occurs when one variable is used to create another. Like squaring a variable X, then X and X^2 would be correlated. Data collinearity is when the multicollinearity occurs within the data itself, and is therefore not a product of the model(Frost, 2017). Multicollinearity generally happens due to highly observational data, creating new variables, poorly designed experiments, insufficient data and inclusion of identical variables(Bhandari, 2020).

Why is multicollinearity a problem?

Some of the problems of multicollinearity include coefficient estimates varying wildly, depending on the choice of variables for the model, meaning that they are very sensitive to even small changes in the model, resulting in instability and misinterpretation (Bhandari, 2020). Other problems multicollinearity can present are reduction of precision of the estimated coefficients, this can result in not being able to trust the p-value as well not being able to identify which variables are statistically significant. In addition it is hard to see which effect each variable has (Frost, 2017).

Checking multicollinearity

Multicollinearity can be tested using Variance inflation Factor (VIF), VIF tests the correlation between the independent variables and the strength of that correlation. It is predicted by regressing a variable against every other variable, and explains how well the variable is explained by other variables (Bhandari, 2020). Variance inflation factor is calculated as 1 divided by the tolerance. The tolerance is the percent of variance in the variable that cannot be accounted for by other variables and is calculated as 1 minus R squared (*Collinearity Diagnostics*, n.d.). The VIF score starts at 1 and has no upper limit, a value of 1 indicates that there is no correlation between variables and therefore no collinearity. A score between 1 and 5 suggests a moderate collinearity, but not enough to take corrective action. Above 5 is serious and warrants corrective measures, as the severity of the potential problems increase (Frost, 2017).

Fixing collinearity

Sometimes collinearity doesn’t have to be fixed. If p-values and knowing exactly how each variable affects the others and the outcome doesn’t matter, and the only goal is prediction, fixing collinearity should not be a concern. However, if these matter, fixing collinearity can be solved in different ways. Structural collinearity can be fixed by standardizing all variables or removing an identical predictor. If the collinearity is in the data, it is important to find out which variables are collinear, as the other variables can be interpreted and used, or the variables that are collinear can be removed. Other solutions include: Linearly combining variables (such as adding them together), partial least squares regression in which principal component analysis is utilized, other regressions such as LASSO and Ridge, where multicollinearity is not an issue (Frost, 2017).