Understanding Imperfect Multicollinearity: The Implications for the Standard Error of a Regression Coefficient

Zach Rutledge

2/14/17

Imperfect multicollinearity between the regressors in a regression model occurs when there is a high degree of correlation between the variables in your model. If you have a simple regression model that includes a single right hand side variable X_1 and you want to add an additional variable X_2 to your model that is highly correlated with X_1, then the standard error of coefficient on the variable X_1 is likely to be larger after you add in the X_2 variable. In this paper, I provide an explanation of this phenomenon when using two regressors and then follow up with an explanation that covers the general case.

Suppose you estimate the following model:

$$Y_i = b_0 + b_1 X_i + e_i. \tag{1}$$

The formula for the variance of b1 in this simple linear regression is:

$$var[b1] = \frac{s_1^2}{\sum x_i^2}, \tag{2}$$

where $s_1^2 = ESS_1/(n-k-1)$ from the simple regression in (3), and $ESS_1 = \sum (Y_i - b_0 - b_1 X_i - e_i)^2$.

However, when two X variables are desired on the right hand side of our regression, such as when the model to be estimated is:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i, \tag{3}$$

then the formula for the variance of b1 can be calculated as follows:

$$var[b1] = \frac{s_2^2}{\sum x_i^2 (1 - R_{1,2}^2)} \tag{4}$$

Where $s_2^2 = ESS_2/(n-k-1)$ from the multiple regression with the two X variables (i.e., $ESS_2 = \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - e_i)^2$, and $R_{1,2}^2$ is the R^2 from a regression of X_1 on X_2. If X_1 and X_2 are highly multicollinear, then the $R_{1,2}^2$ from the regression of X_1 on X_2 will be large because there is a strong linear relationship between X_1 and X_2 (meaning that X_2 is a good predictor for X_1). Therefore $(1 - R_{1,2}^2)$ will be very small (remember that R^2 is always between 0 and 1). This will make the whole denominator of equation (4) very small. However, s_2^2 in the numerator of (4) will also be smaller than s_1^2 because the model includes an additional variable that provides some additional explanatory power (even if this information is not very helpful), so adding highly multicollinear variables into the model leads to a situation in which there is a "battle" between the numerator and the denominator of the formula for the variance of b_1. Because

adding an additional variable that is highly correlated with X_1 does not add much explanatory power, the s_2^2 will not be much smaller than s_1^2, so the decrease in the numerator is likely to be smaller than the decrease in the denominator, which will likely lead to a higher variance of b1. It naturally follows that this leads to the standard error of b_1 becoming larger.

In the general case, if you have more than two variables on the right hand side of your model, your model can be expressed as follows:

$$Y_i = b_0 + b_1 X_i + \cdots + b_k X_{k_i} + e_i$$

In this case the variance of b_1 can be expressed as follows:

$$var[b1] = \frac{s_k^2}{\sum x_i^2 (1 - R^2_{1,\ldots,k})} \tag{5}$$

Where $R^2_{1,\ldots,k}$ is the R^2 from a regression of X_1 on X_2 through X_k. The same rationale holds in this case as in the case with two X variables. That is: if X_2 though X_k have a high degree of correlation with X_1, you are likely to wind up with larger standard errors for b_1 than you would if you omit the other variables that are multicollinear with the X_1 variable.