

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

Zach Rutledge

2/14/17

Imperfect multicollinearity in a regression model occurs when there is a high degree of correlation between the regressor of interest and another regressor in the model, but the variables are not perfectly correlated (i.e., the correlation coefficient between the variables does not equal 1 or -1). When a researcher uses a regression model that contains only a single right hand side variable  $X_1$  and wants to add an additional variable  $X_2$  to the 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 she adds in the  $X_2$  variable. In this handout, I provide an explanation for this phenomenon when adding a second regressor into the model and then follow up with an explanation that covers the general case where the researcher wants to add  $k-1$  more independent variables ( $X_2, \dots, X_k$ ) into the model.

Suppose the researcher estimates the following model with her sample of data:

$$Y_i = b_0 + b_1 X_{1i} + e_i,$$

where  $Y_i$  is the dependent variable,  $b_0$  is the constant term,  $X_{1i}$  is the variable of interest, and  $e_i$  is the error term. The formula for the variance of  $b_1$  in this simple linear regression is:

$$var[b_1] = \frac{s_1^2}{\sum (x_i - \bar{x}_i)^2}, \quad (1)$$

where  $s_1^2 = ESS_1 / (n - k - 1)$  from the simple regression in (1),  $n$  represents the number of observations used to estimate the regression,  $k$  represents the number of independent variables in the model (in this case only 1), and  $ESS_1 = \sum (Y_i - b_0 - b_1 X_{1i})^2$ . However, when two independent variables ( $X_{1i}$  and  $X_{2i}$ ) are included on the right hand side of the regression, such as when the model to be estimated is:

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

then the formula for the variance of  $b_1$  can be calculated as follows:

$$var[b_1] = \frac{s_2^2}{[\sum (x_i - \bar{x}_i)^2] (1 - R_{1,2}^2)}, \quad (2)$$

where  $s_2^2 = ESS_2 / (n - k - 1)$  from the multiple regression with the two X variables (i.e.,  $ESS_2 = \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$ ), 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 (2) very small. However,  $s_2^2$  in the numerator of (2) 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 a highly multicollinear variable into the model leads to a situation in which there is a “battle” between the shrinking numerator and the shrinking 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 may be proportionately smaller than the decrease in the denominator, which will likely lead to a higher variance of  $b_1$ . It naturally follows that this leads to the standard error of  $b_1$  becoming larger. In the general case, if the researcher wants to include more than two variables on the right hand side of her model, the model can be expressed as follows:

$$Y_i = b_0 + b_1 X_{1i} + \dots + b_k X_{ki} + e_i.$$

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

$$var[b_1] = \frac{s_k^2}{[\sum (x_i - \bar{x}_i)^2](1 - R_{1,\dots,k}^2)}, \quad (3)$$

where  $s_k^2 = ESS_k / (n - k - 1)$ ,  $ESS_k = \sum (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki})^2$ ,  $R_{1,\dots,k}^2$  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  $X_2$  through  $X_k$  have a high degree of correlation with  $X_1$ , the researcher is likely to wind up with a larger standard error for  $b_1$  than she would if she omitted the other variables that are multicollinear with the  $X_1$  variable.