PPOL502-01, Spring 2016
Course Notes #8: Factors Affecting Standard Errors of Coefficients, Multicollinearity

---

I. STANDARD ERROR COMPONENTS

In Course Notes #7 we introduced assumption MLR6 and the following formula for the standard error of a regression coefficient:

$$\sqrt{s.e._{\hat{\beta}_j}^2} = \sqrt{Var(\hat{\beta}_j)} = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}}$$

where $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k-1}$

$$SST_j = \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2$$

R_j^2 = the R-squared from the regression of X_j—where j indicates the X variable of interest—on **all the *other Xs*** in the original model of interest (i.e., all the Xs in the model $Y = \beta_0 + \beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots\beta_kX_k+u$)

Now we'll examine each component of this formula more carefully. In particular, we'll discuss what factors affect the magnitude of each piece of the formula and the corresponding impact on the standard error.

II. NUMERATOR OF SE: MEAN SQUARE ERROR

- Let's look at the numerator of the SE formula first:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k-1}$$

This piece represents the Mean Square Error (MSE). The larger the MSE, the larger the SE. In general, the larger the sum of the residuals, or alternatively the lower the model R-squared, the larger the MSE and the larger the standard error for the regression coefficient.

Since larger SEs on regression coefficients are not desirable, what can be done to reduce the MSE? Sometimes including additional independent variables will help.

Adding Regressors To <u>Reduce</u> The Error Variance

- According to omitted variable logic, we know the kinds of variables we want to include in a regression model--those that are correlated with the X variables of interest and have a partial affect on Y.

- What about variables (Z_k) that are related to *Y*, but are *not* related to the *X*s in the model? Leaving such Z_k out of the model will not bias the coefficients of the other variables in the model, so we're o.k. there.

- But including such Z_k in the model will have an added benefit: it will contribute to the explanatory power of the model by reducing the error variance (i.e., $\hat{\sigma}_2$)

- Why is this good? Because it reduces the standard error on all the remaining variables. Remember the formula?….

$$s.e._{\hat{\beta}_j}^2 = Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j(1-R_j^2)} \qquad \text{where} \quad \hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k-1}$$

- So, there are advantages to include variables in a regression that are correlated with *Y*, but not with any of the other independent variables.

NOTE:, Recall from earlier Course Notes our discussion of adding *irrelevant variables* (i.e., variables that are unrelated to *Y,* but may be related to the *Xs* in the model*)*:

- Should we include an explanatory variable X_k in the regression that we predict to have no partial effect on *Y* (i.e., $\beta_k = 0$)?

  -- There is no harm done in terms of biased coefficients on the other variables in the model if we include such irrelevant variables.

  -- However, in terms of efficiency—standard errors on the other variables in the model—is likely a cost if we include these irrelevant variables: in particular, the MSE will increase (the numerator of the standard error) <u>due to increasing *k,*</u> and at the same time the denominator of the standard error is likely to increase even more, due to likely intercorrelations of the added *X* with the *Xs* already in the model (see Section IV on multicollinearity below).

  In brief, the standard errors on all other variables in the model are likely to increase, so it is not a good idea to include irrelevant variables (no partial effect on Y) in the model.

III. DENOMINATOR OF SE (Part 1): SST_j

- Now let's look at the first part of the denominator of the SE formula:

$$SST_j = \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2$$

Since SST_j is part of the denominator, the larger the SST_j the smaller the overall SE coefficient.

So what factors drive the magnitude of the SST for a particular variable (j) in the model? There are two factors:

1. *The spread of the X values around the overall mean (for that particular X).*

   For example, assume that the variable AGE is the X variable of interest and the average age in the sample is 40. Which of the following samples would have a larger SST_j (and smaller SE)?

   --A sample where half the persons were age 10 and the other half were age 70, or

   --A sample where half the persons were age 30 and the other half were age 50?

2. *The number of observations (sample size)*

   SST_j can only increase with the sample size, so to get the smallest SE possible (all else equal) the more observations the better!

Which of these two factors would generally be more under the control of the researcher?

IV. DENOMINATOR OF SE (Part 2): SST_j / MULTICOLLINEARITY

- The second part of the denominator of the SE formula is: 1- R_j^2

It is important to understand that R_j^2 = the R-squared from the regression of X_j—where j indicates the X variable of interest—on **all the *other Xs*** in the original model of interest (i.e., all the Xs in the model $Y = \beta_0 + \beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots\beta_kX_k+u$). This is NOT the same thing as the overall R-squared for the model!

Remember that R_j^2 ranges from 0 to 1. As R_j^2 increases, 1- R_j^2 decreases.

Question: What happens to the SE as R_j^2 increases?
{Hint: remember that this is part of the *denominator* of the SE formula).

The answer to this question is a good segue to discuss *multicollinearity*.

- What is **multicollinearity**?: It is high (<u>but not perfect</u>) correlation between explanatory variables ***X***. (Recall, if there were perfect collinearity, Assumption MLR 4 would be violated and the model wouldn't run.)

- What multicollinearity does: Increases the standard errors on the variables that are highly correlated. Why? Think about how the second part of the denominator of the SE formula below is affected (see note above):

$$s.e._{\hat{\beta}_j}^2 = Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}$$

- What multicollinearity doesn't do: Does *not* bias the coefficient estimates themselves (i.e., $\hat{\beta}_j$ is still an unbiased estimator of β_j)

- Whether to worry about multicollinearity / what to do about it: "Get more data" is usually the call to arms, where "more data" means more observations. What parts of the formula above are affected by the "get more data" mantra?

  Question: Should you drop variables from the model that are highly correlated with each other?

- Handout from Goldberger….