Discussion of Omitted Variable Bias versus Multicollinearity

1. OVERVIEW

In the lectures on Simple Linear Regression and Multiple linear regression we spent a lot of time talking about taking repeat draws of random samples from a population, estimating a regression based on the sample and calculating a $\hat{\beta}$. In addition, we spent a lot of time talking about (i) the expected value of $\hat{\beta}$, which we would ideally like to be equal to the true population parameter β, and (ii) the variance of $\hat{\beta}$, which we would ideally like to be low (i.e. a tight distribution of $\hat{\beta}$'s). **Omitted variable bias** and **multicollinearity** are problems to the extent that they can thwart these ideals.

1.1 Expected value of $\hat{\beta}$ and Omitted Variable Bias: When talking about the expected value of $\hat{\beta}$ (E[$\hat{\beta}$]) we discussed the desirable quality of unbiasedness, which says that the mean value of $\hat{\beta}$ over many repeat random samples should be equal to the true population beta (that E[$\hat{\beta}$]=β is satisfied). Omitted variable bias affects the expected value E[$\hat{\beta}$]. In particular, if you exclude (omit) a variable (z) from your regression model that is correlated with both your explanatory variable of interest (x) and your outcome variable (y) then the expected value of $\hat{\beta}$ will be biased (E[$\hat{\beta}$]$\neq$ β). We call this problem "omitted variable bias" (OVB).

1.2 Variance of $\hat{\beta}$ and Multicollinearity: When talking about the variance of $\hat{\beta}$ we discussed the desirable quality of having low variance (i.e. a tight/narrow distribution of $\hat{\beta}$'s), which means that the estimated β's (the $\hat{\beta}$'s) over many repeat random samples will be tightly centered. Further, low variance for the random variable $\hat{\beta}$ corresponds to having a small standard error for $\hat{\beta}$. Multicollinearity affects the var($\hat{\beta}$), (also written as $\sigma_{\hat{\beta}}{}^2$). In particular, if you include a variable (z) in your regression model that is correlated with the explanatory variable(s) of interest (x) then this acts to increase the variance of $\hat{\beta}$ (where $\hat{\beta}$ is the regression coefficient on x) whenever the variable z explains little of the variation in the outcome variable. And, of course, a larger variance of $\hat{\beta}$ corresponds to large standard error for $\hat{\beta}$. We call this problem "multicollinearity". The problem results from the fact that in multiple linear regression we only use residual variation in the explanatory variables to estimate the regression coefficients. If there is very little residual variation in our explanatory variable of interest, then this is equivalent to having only a little total sample variation in our explanatory variable. Recall, that the total sample variation in our explanatory variable of interest (SST) lies in the denominator for the variance of $\hat{\beta}$—so when this denominator goes down, then our standard error goes up. However, it is worth pointing out that if z explains a great deal of the variation in the outcome variable then this reduces the sum of squared residuals (SSR). As the SSR lies in the numerator for the formula of the variance, this acts to reduce the variance of $\hat{\beta}$.

2. EXAMPLES

2.1 Omitted Variable Bias Example: Once again, $\hat{\beta}$ will be biased if we exclude (omit) a variable (z) that is correlated with both the explanatory variable of interest (x) and the outcome variable (y). The second page of **Handout #7b** provides a practical demonstration of what can happen to $\hat{\beta}$ when you exclude a variable z correlated with both x and y; I re-produce the results here with some additional commentary.

From Handout #7b:

Omitting an important variable correlated with the other independent variables:
Omitted variable bias

$$\begin{array}{lllllll}
\widehat{\log(\text{wage})} = & 1.17 & +\ .106\ \text{educ} & +.011\ \text{exp} & -\ .26\ \text{female} & +\ .012\ \text{profocc} & R^2 = .28 \\
 & (.08) & (.005) & (.0009) & (.02) & (.03) & n = 2000
\end{array}$$

$$\begin{array}{lllllll}
\widehat{\log(\text{wage})} = & 2.57 & + & +.011\ \text{exp} & -\ .26\ \text{female} & +\ .358\ \text{profocc} & R^2 = .16 \\
 & (.03) & & (.0009) & (.02) & (.03) & n = 2000
\end{array}$$

Additional Commentary on Handout #7: The difference between these regression models is that the second model excludes ‘educ’. As indicated, in the second equation we have excluded (omitted) the variable ‘educ’ which is an important variable in that it determines the outcome (i.e. education affects log(wages)); ‘educ’ is also correlated with other explanatory variables, in particular, the indicator for whether your employment type is a professional occupation (‘profocc’). The correlation between ‘educ’ and ‘profocc’ is 0.4276 (positive) as indicated in the correlation matrix below. As a consequence, in the second equation the regression coefficient on ‘profocc’ is measuring the effect of both having higher education and having a professional occupation; that is, our estimator for the regression coefficient on ‘profocc’ exhibits a bias relative to the first equation. Consistent with demonstrations from class, the bias present in the estimator for the regression coefficient on ‘profocc’ in the second equation (0.358) is positive relative to the first equation. Said differently, due to the exclusion of ‘educ’ the estimator for the regression coefficient on ‘profocc’ in the second equation is positively/upward biased relative to the first equation.

```
. correlate lwage educ exp female profocc nonwhite
(obs=2000)

             |    lwage     educ    exper   female  profocc nonwhite
-------------+------------------------------------------------------
       lwage |   1.0000
        educ |   0.4097   1.0000
       exper |   0.2358   0.0010   1.0000
      female |  -0.1935   0.0489   0.0210   1.0000
     profocc |   0.2181   0.4276  -0.0383   0.1077   1.0000
    nonwhite |  -0.0379  -0.0051  -0.0200   0.0368  -0.0143   1.0000
```

2.2 Multicollinearity Example: As we add variables to our regression model that are correlated with the explanatory variable(s) of interest, then the standard errors for the $\hat{\beta}$’s on the explanatory variable(s) of interest will tend to increase, particularly when the added variables do not explain variation in the outcome variable (i.e. when the added variables do not reduce the sum of squared residuals). **Handout #8** provides a practical demonstration of what happens to the standard errors for your $\hat{\beta}$’s when you include a variable that is highly correlated with the explanatory variables already in the model, but does not explain much variation in *y*; I re-produce the relevant results from Handout #8 here with some additional commentary.

From Handout #8:

(1) None

```
. reg lwage educ exper female

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  3,  1996) =  247.50
       Model |   182.35726     3  60.7857535           Prob > F      =  0.0000
    Residual |  490.219607  1996  .245601005           R-squared     =  0.2711
-------------+------------------------------           Adj R-squared =  0.2700
       Total |  672.576867  1999  .336456662           Root MSE      =  .49558

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1167441   .0053157    21.96   0.000     .1063191     .127169
       exper |   .0109089    .000869    12.55   0.000     .0092046    .0126132
      female |  -.2543189   .0222067   -11.45   0.000    -.2978696   -.2107682
       _cons |   1.055792   .0757381    13.94   0.000     .9072576    1.204326
------------------------------------------------------------------------------
```

(2) Almost collinear

```
. reg lwage educ exper female age

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  4,  1995) =  185.69
       Model |  182.468262     4  45.6170655           Prob > F      =  0.0000
    Residual |  490.108605  1995  .245668474           R-squared     =  0.2713
-------------+------------------------------           Adj R-squared =  0.2698
       Total |  672.576867  1999  .336456662           Root MSE      =  .49565

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1692465   .1115687     1.52   0.129    -.0495568    .3880498
       exper |   .0633711   .1113346     0.57   0.569    -.1549732    .2817154
      female |  -.2545469   .0222135   -11.46   0.000     -.298111   -.2109827
         age |  -.0524796   .1113744    -0.47   0.638     -.270902    .1659428
       _cons |   1.370917   .6728026     2.04   0.042     .0514472    2.690386
------------------------------------------------------------------------------
```

Commentary on example from Handout #8: The difference between these regression models is that the second model includes the variable ‘age’. In the first set of regression results we see relatively small standard errors for the $\hat{\beta}$’s on ‘educ’ and ‘exper’ as indicated by numbers reported in the column under ‘Std. Err.’. In the second set of regression results we see that standard errors for these two $\hat{\beta}$’s are considerably larger. Why does this happen? If the variable added to the regression equation (e.g. age) is highly correlated with variables already in the model (e.g. educ and exper) and does not explain *lwage*, then the standard errors for the associated $\hat{\beta}s$ will get very large. This is what is meant by multicollinearity. We are concerned about multicollinearity because large standard errors for the $\hat{\beta}$'s produce large confidence intervals and makes it likely that you will fail to reject the null hypothesis even if the magnitude of the estimated regression coefficient ($\hat{\beta}$) is much different than the null hypothesis. **Note:** The heading "None" on the first set of regression results reflects the fact that the standard errors are small, which suggests that the variables in the model are not close to being collinear (i.e. they are not close to being perfectly correlated). The heading "Almost collinear" for the second set of regression results reflects the fact that the standard errors are very large, which suggest that some of the variables are very highly correlated (i.e. they are almost collinear/perfectly correlated).

3. SUMMARY OF OVB & MULTICOLLINEARITY

Consider the SLR model:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + u \qquad eq.1$$

Now suppose we run the MLR model by adding a control variable *z*:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 z + e \qquad eq.2$$

I use the tilde above the regression coefficients in eq. 2 to distinguish them from the OLS estimator in eq. 1, which have the carrot hats. The $\tilde{\beta}s$ still represent the OLS estimator, it's just a different OLS estimator than that presented in eq. 1.

1) **If *z* is related to both *x* and *y***, then we want to include *z* if we have data on it because by including *z* we reduce bias due to its omission. The extent to which *z* is correlated with *x* will also affect how much multicollinearity will act to increase the standard errors on the regression coefficient for *x*. In this scenario, if OVB is a real concern, then we need to add *z* and then live with the consequence of multicollinearity and larger standard errors. Although, it is important to note that the standard errors still might get smaller with the inclusion of *z*. Why? Because *z* is also related to *y*, adding it in eq. 2 will reduce the SSR which acts to reduce the variance, and therefore, the standard errors of the beta estimators.
2) **If *z* is unrelated to *x* but related to *y***, then we want to include *z* if we have data on it because its inclusion will reduce the SSR which acts to reduce the standard errors of the regression coefficients in the model. So even if a variable doesn't reduce bias, there can be an advantage to including it in the multiple linear regression model.
3) **If *z* is related to *x* but unrelated to *y*,** then this is multicollinearity at its worst so we want to exclude *z*. Adding *z* to the regression won't reduce bias and won't reduce the sum of squared ressiduals, but it will reduce the residual variation in *x*; that is, the variance inflation factor will scale the denominator towards zero which blows up the variance of the regression coefficient on *x*. Note: The variance inflation factor (VIF) is:

i. $VIF = {}^{1}\!\big/_{(1-R_j^2)}$

4) **If *z* is unrelated to both *x* and *y*,** then it should not matter much whether you include it or exclude it. Although if you add a lot of variables unrelated to both *x* and *y*, then you will start to eat into your degress of freedom, which also enters into the variance equation for $\hat{\beta}$.

Detecting OVB & Multicollinearity:

It can sometimes be hard to detect OVB and multicollinearity. We cannot detect OVB if we don’t have data measures on the omitted variable, one can only argue that your regression equation and its estimators for the regression coefficients are likely vulnerable to omitted variable bias. If you do have measures of some additional variable, then you can assess the importance of including it in your regression estimation.

- If you add a variable *z* to the regression equation and the estimated regression coefficient on the *x* of interest changes a lot, then this suggests that *z* is related to *x* and *y* so should be included to avoid OVB. This is the case regardless of what happens to your standard error on the regression coefficient of interest. This corresponds to summary point (1) above.
- However, if you add a variable *z* to the regression equation and the estimated regression coefficient on the *x* of interest does not change, then this suggests that *z* is unrelated to *x* or unrelated to *y* or both. Thus, excluding *z* does not introduce omitted variable bias. Given this, let’s consider what happens to the standard error for the regression coefficient on *x* when we add *z*.
    - If including *z* increases the standard error for the regression coefficient on *x* then this suggests that *z* and *x* are related, including *z* is the type of multicollinearity we want to avoid. Under the given that excluding *z* does not introduce OVB, let’s exclude *z* to avoid larger standard errors.
    - If including *z* decreases the standard error for the regression coefficient on *x* then this suggests that *z* and *y* are related. Multicollinearity is not a concern, and in fact, adding *z* to our estimation reduces the standard error for the regression coefficient on *x*. Even though excluding *z* does not generate OVB, let’s include *z* to reduce the standard errors.
    - If including *z* does not affect the standard error on *x* then this suggests that *z* is unrelated to *x* and unrelated to *y*. Multicollinearity is not a concern. Under the given OVB is not a problem by omitting *z*. We often exclude these type of variables from our regression, though we sometimes include them to appease our audience or reviewer (i.e. assure them that we aren’t introducing OVB by excluding a given variable).