# Econometrics: Specification Errors

Burcu Eke

UC3M

# Introduction

- So far, we have focused on the models that satisfy the assumptions of linear regression model (LRM) and therefore have nice properties and interpretations.

- The model specification of the LRM includes choosing
  - Choosing the independent variables, and thus the omitted variables
  - Choosing the functional form

- Some very important questions are:
  - What would happen if we used the LRM when the assumptions are not met, i.e., when it is not appropriate?
  - What are the properties of the OLS estimators under a specification error?

# Introduction

- Specification errors in which we are interested:
  1. Inclusion of irrelevant variables.
  2. Omission of relevant variables.
  3. Measurement errors in variables.

# Omitting a Relevant Variable

- Recall the full and reduced models we have talked about when we first introduced multiple linear regression.

- Now, lets revisit this issue from another perspective:
  - Let's consider the multiple linear regression model:
    $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
  - For some reason such as unavailability of the data on $X_2$, we construct a regression model without the variable $X_2$ variable
    $Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1$

# Omitting a Relevant Variable

- In econometrics, this issue is known as "omitting a relevant variable", **if** $\beta_2 \neq 0$, and this is a type of misspecification
  - The big question then is *what is the effect of omitting a relevant variable?*
  - The answer is given in the equation below.

$$\gamma_1 = \beta_1 + \beta_2 \frac{C(X_1, X_2)}{V(X_1)} \tag{1}$$

  - The equation (1) is called the **rule of omitted variable**, which shows that the slope of the reduced model is a linear combination of $\beta_1$ and $\beta_2$ (the two slopes of the full model)

# Omitting a Relevant Variable

- Then, by omitting $X_2$, its effect becomes part of the error term in the reduced model:
  $\varepsilon_1 = \varepsilon + \beta_2 X_2$

- This implies:

$$\begin{aligned} E[\varepsilon_1|X_1] &= E\left[\varepsilon + \beta_2 X_2|X_1\right] \\ &= E\left[\varepsilon|X_1\right] + E\left[\beta_2 X_2|X_1\right] \\ &= E\left[E\left[\varepsilon|X_1, X_2\right]|X_1\right] + \beta_2 E\left[X_2|X_1\right] \\ &= \beta_2 E\left[X_2|X_1\right] \end{aligned}$$

- Hence,

$$E[Y|X_1] = \gamma_0 + \gamma_1 X_1 + \beta_2 E\left[X_2|X_1\right]$$

# Omitting a Relevant Variable

- In general, whenever we have $E[\varepsilon|X_1, \ldots, X_k] \neq 0$ for multiple linear model
  $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$
  It means that the model is **misspecified**

# Omitting a Relevant Variable: The Properties of OLS Estimators

- Now, let's focus on the properties of the OLS estimators under this specification error.

- Consider the simplest multiple linear regression, and a simple regression using only one of the variables
  - Let the MLR model be $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, with the estimated model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
  - Let the SLR model be $Y = \gamma_0 + \gamma_1 X_1 + \varepsilon$, with the estimated model $\hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 X_1$

# Omitting a Relevant Variable: The Properties of OLS Estimators

- Then we know that $\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}_1$, where $\hat{\delta}_1$ is the OLS estimator of the slope of $L\left(X_2|X_1\right) = \delta_0 + \delta_1 X_1$

- We also know that
  - $E[\hat{\beta}_1] = \beta_1$ and $p\lim \hat{\beta}_1 = \beta_1$
  - $E[\hat{\gamma}_1] = \gamma_1$ and $p\lim \hat{\gamma}_1 = \gamma_1$

# Omitting a Relevant Variable: The Properties of OLS Estimators

- Using all these information, we have

$$E\left[\hat{\gamma}_1|X_1,X_2\right]=E\left[\hat{\beta}_1|X_1,X_2\right]+E\left[\hat{\beta}_2\hat{\delta}_1|X_1,X_2\right]=\beta_1+\beta_2\hat{\delta}_1$$

- This implies

$$\begin{aligned} E[\hat{\gamma}_1] &= \beta_1+\beta_2 E[\hat{\delta}_1] \\ p\lim(\hat{\gamma}_1) &= \beta_1+\beta_2\delta_1 \end{aligned}$$

- In general:
  - $\hat{\gamma}_1$ won't be appropriate if we want to make inference about $\beta_1$
  - Furthermore, it is easy to show that $V\left(\hat{\gamma}_1\right)\leq V\left(\hat{\beta}_1\right)$

# Omitting a Relevant Variable: The Properties of OLS Estimators

- Whenever $X_2$ is a "relevant variable" (that is, $\beta_2 \neq 0$),$\hat{\gamma}_1 \Rightarrow$ **inconsistent** and **biased** estimator of $\beta_1 \Rightarrow$ bias will not disappear no matter how big the sample size is

- $V(\hat{\gamma}_1) = \dfrac{\sigma^2}{\sum_i x_{1i}}$ and $V\left(\hat{\beta}_1\right) = \dfrac{\sigma^2}{\sum_i x_{1i}(1 - R_1^2)}$, that is $V(\hat{\gamma}_1) \leq V\left(\hat{\beta}_1\right)$, but $\hat{V}(\hat{\gamma}_1)$ is a biased estimator for the variance of $\hat{\beta}_1$

- That is, the "omission of relevant variables" in the analysis generates inconsistency and bias in estimating the effects of variables, though a reduction in the variance of the estimator.

- How about hypothesis test? Are they valid?

# Omitting a Relevant Variable: The Properties of OLS Estimators

- In other words, the coefficient of $X_1$ in the regression that incorrectly omits $X_2$:
  - does not capture the ceteris paribus effect of $X_1$ on $Y$ (sice when $X_1$ changes, so does $X_2$)
  - captures the effect on $Y$ coming from a change in $X_1$ plus the effect of $X_1$ on $X_2$ (which DOES have an effect on $Y$)

- We can summarize the bias in estimating $\beta_1$ when $X_2$ is incorrectly omitted as:

| | $C(X_1, X_2) > 0$ | $C(X_1, X_2) < 0$ |
|---|---|---|
| $\beta_2 > 0$ | $+$ | $-$ |
| $\beta_2 < 0$ | $-$ | $+$ |

# Including an Irrelevant Variable

- $X_2$ is an "irrelevant" variable, that is, $\beta_2 = 0$
- Now, then, the true model is $Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1$ and we estimate
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, thinking that $X_2$ is relevant
- In this case, $\varepsilon = \varepsilon_1 - \beta_2 X_2 = \varepsilon_1$ since $\beta_2 = 0$ in the true model
- This implies:

$$\begin{aligned} E[\varepsilon|X_1, X_2] &= E\left[\varepsilon_1 - \beta_2 X_2|X_1, X_2\right] \\ &= E\left[\varepsilon_1|X_1\right] \end{aligned}$$

# Including an Irrelevant Variable: The Properties of OLS Estimators

- Then, $\hat{\beta}_1$ is a **consistent** and **unbiased** estimator of $\gamma_1$ and it has less variance than $V(\hat{\beta}_1) \geq V(\hat{\gamma}_1)$, which is also unbiased and consistent.

- That is, the "inclusion of irrelevant variables" in the analysis, does not affect the consistency of the estimated effect of the variables.

- Intuition: The true population value of the coefficient of an irrelevant variable is 0, so by including this variable, the coefficient estimators for the other variables are not affected in the limit.

# Including an Irrelevant Variable: The Properties of OLS Estimators

- However, estimated $\beta$'s will be generally **inefficient**, that is, their variances will be greater than those of the true model
  - Intuition: The higher the correlation between the irrelevant and relevant variables, the greater the variance of the estimated coefficient for the relevant variables.
  - This means that when irrelevant variables are included, we do not have the problem of inconsistency of the OLS estimator (and hence including an irrelevant variable is a less serious problem)
  - Nevertheless,the inefficiency problem can generate a serious consequences when testing hypotheses of type $H_0 : \beta_j = 0$ due to the lost of power, so we might infer that they are not relevant variables when they truly are (type II error)

# Including an Irrelevant Variable: The Properties of OLS Estimators

- In practice, Is it possible to know which model is the appropriate one?
  - Strictly: No.
  - In general the best approach is to include only the variables that, based on economic theory, affects the dependent variable, and are not accounted for other variables in the model
  - Then we can gather evidence for or against the "relevance" or "irrelevance" of one or more variables through the testing of hypotheses.

# Example 1: The impact of tobacco consumption on cancer

- Assume that we information on existence of cancer in two groups of people: smokers and nonsmokers
- Additionally, let's assume that smokers are more likely to engage in more physical activities which reduces the likelihood of cancer. But we don't observe these activities
- Therefore the impact of smoking on cancer will be overestimated because the tobacco consumption may decrease the level of physical activity.

# Example 1: The impact of tobacco consumption on cancer

- Formally, $C_i = \beta_0 + \beta_1 F_i + \beta 2_E J_i + \varepsilon$, where,
  - $C_i$ as the measure of cancer for individual $i$
  - $F_i$ is a dummy variable that takes a value of 1 is the individual $i$ is a smoker and 0 otherwise
  - $EJ_i$ is a measure of physical activity, i.e., exercise
  - Let the true values be $\beta_1 > 0,\ \beta_2 < 0$

# Example 1: The impact of tobacco consumption on cancer

- Additionally, $C_i = \delta_0 + \delta_1 F_i + \varepsilon_1$, with $\delta_1 < 0$
- Therefore, when we run the simple regression of $C_i$ on $F_i$, we will have
  - $C_i = \gamma_0 + \gamma_1 F_i + \varepsilon_2$,
  - Then we have $\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$

# Example 2: The impact of tobacco on wages

- Additionally to the impact on health outcomes, does smoking have economic consequences?

- Smokers might face lower wages than non-smokers:
  - If they were less productive due to "cigarette breaks";
  - If smoking has an impact on health outcomes, smokers would be more likely ask for sick-leaves;
  - If the firm would discriminate against smokers; and so on.

# Example 2: The impact of tobacco on wages

- We have representative data for individuals 30 years old for the US. Levine, Gustafson and Velenchik (1997) estimated a wage equation using the following variables:
  - $Y = \ln(\text{wage})$
  - $F =$ a dummy variable that takes a value of 1 for smokers and 0, otherwise.
  - $ED =$ Years of education

- We must take into consideration that smokers have, on average, less education than non-smokers (thus, education is negatively correlated with smoking)

# Example 2: The impact of tobacco on wages

- Two specifications considered are:
  - Omitting education: $\hat{Y}_i = -0.176F_i$ with $s_{\hat{\beta}_1} = 0.021$
  - Including education: $\hat{Y}_i = -0.080F_i + 0.070ED_i$ with $s_{\hat{\beta}_1} = 0.021$ and $s_{\hat{\beta}_2} = 0.004$
- By not including education in the regression we overestimated the impact of smoking.

# Measurement Error

- So far, we have assumed that $Y$ and $X_j$'s are accurate, i.e., there are no measurement errors
- Sometimes we have data that has measurement errors, or that there are no data available on the variable of interest.
- For example, according to the life cycle models, the consumption depends on the permanent income which we cannot measure without error, or we have data on the reported annual income, but not the real annual income.

# Measurement Error

- **Measurement error** occurs when we cannot accurately measure the magnitude of the variable of interest. We are interested in the effects of such errors on our OLS estimators.

- There are two types of measurement error:
  1. Measurement error in the dependent variable, $Y$
  2. Measurement error in the independent variable, $X$.

# Measurement Error in Y

- Consider the following model:

$$Y^* = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

- $E[\varepsilon|X] = 0 \Rightarrow E[Y^*|X] = L(Y^*|X) = \beta_0 + \beta_1 X$
- $E(\varepsilon) = 0,\ C(X, \varepsilon) = 0$ and $\beta_0 = E[Y^*] - \beta_1 E[X]$, $\beta_1 = \dfrac{C(X, Y^*)}{V(X)}$

# Measurement Error in Y

- Assume that the Model (2), $Y^*$ has measurement error, so that $Y = Y^* + v_0$, where $v_0$ is the measurement error in $Y^*$
- Then this model becomes $Y = \beta_0 + \beta_1 X + \varepsilon + v_0$:

$$Y = \beta_0 + \beta_1 X + \underbrace{(\varepsilon + v_0)}_{u} \tag{3}$$

- $u$ is the composite error term

# Measurement Error in Y

- Recall that $\hat{\beta}_1 = \frac{\frac{1}{n}\sum_i x_i y_i}{\frac{1}{n}\sum_i x_i^2}$, $y_i = Y_i - \bar{Y}$, and $x_i = X_i - \bar{X}$

$$
\begin{aligned}
p\lim_{n\to\infty} \hat{\beta}_1 &= \frac{p\lim_{n\to\infty} \frac{1}{n}\sum_i x_i y_i}{p\lim_{n\to\infty} \frac{1}{n}\sum_i x_i^2} = \frac{C(X,Y)}{V(X)} \\
&= \frac{C(X, Y^* + v_0)}{V(X)} = \frac{C(X,Y^*) + C(X,v_0)}{V(X)} \\
&= \beta_1 + \frac{C(X,v_0)}{V(X)} \Rightarrow \text{ consistent if } C(X,v_0) = 0
\end{aligned}
$$

# Measurement Error in Y

- For $\hat{\beta}_0$

$$\begin{aligned} p\lim_{n\to\infty} \hat{\beta}_0 &= p\lim_{n\to\infty} \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) \\ &= E[Y^* + v_0] - E[X] p\lim_{n\to\infty} \hat{\beta}_1 \end{aligned}$$

- Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent if:
  - $C(X, v_0) = 0$ and
  - $E[v_0] = 0$.
- The variances are now larger under measurement error
  - $V(Y^*|X) = V(\varepsilon|X) = \sigma^2$
  - $V(Y|X) = V(Y^* + v_0|X) = \sigma^2 + \sigma^2_{v_0}$, assuming that $C(\varepsilon, v_0|X) = 0$

# Measurement Error in X

- Consider the following model:

$$Y = \beta_0 + \beta_1 X^* + \varepsilon \tag{4}$$

- $E[\varepsilon|X^*] = 0 \Rightarrow E[Y|X^*] = L(Y|X^*) = \beta_0 + \beta_1 X^*$
- $E(\varepsilon) = 0,\ C(X^*, \varepsilon) = 0$ and $\beta_0 = E[Y] - \beta_1 E[X^*]$, $\beta_1 = \dfrac{C(X^*, Y)}{V(X^*)}$

# Measurement Error in X

- Assume that the Model (4), $X^*$ has measurement error, so that $X = X^* + v_1$, where $v_1$ is the measurement error in $X^*$
- Then this model becomes $Y = \beta_0 + \beta_1(X - v_1) + \varepsilon$:

$$Y = \beta_0 + \beta_1 X + \underbrace{(\varepsilon - \beta_1 v_1)}_{u_1} \tag{5}$$

- $E[v_1] = 0$, and $C(X, \varepsilon) = 0$: $\varepsilon$ is correlated neither with $X$ nor with $X^*$, therefore nor with $v_1$.
- $E[Y|X^*, X] = E[Y|X^*] \Rightarrow$ Given $X^*$, $X$ does not contain any additional relevant information.

# Measurement Error in X

- Assume the classical ME assumptions hold, i.e.,
  - $C(X^*, v_1) = 0$ and $C(v_1, \varepsilon) = 0$

$$
\begin{aligned}
p \lim_{n\to\infty} \hat{\beta}_1 &= \frac{p \lim_{n\to\infty} \frac{1}{n} \sum_i x_i y_i}{p \lim_{n\to\infty} \frac{1}{n} \sum_i x_i^2} = \frac{C(X,Y)}{V(X)} \\
&= \frac{C(X^* - v_1, Y)}{V(X^* - v_1)} = \frac{C(X^*, Y) + \overbrace{C(Y, v_0)}^{0}}{V(X^*) + V(v_1)} \\
&= \frac{C(X^*, Y)/V(X^*)}{\left(V(X^*) + V(v_1)\right)/V(X^*)} \\
&= \frac{\beta_1}{1 + V(v_1)/V(X^*)} \neq \beta_1
\end{aligned}
$$

# Measurement Error in X

- The asymptotic bias then:

$$\begin{aligned} p\lim_{n\to\infty}\left(\hat{\beta}_1 = \beta_1\right) &= \frac{\beta_1}{1 + V(v_1)/V(X^*)} - \beta_1 \\ &= -\beta_1 \frac{V(v_1)}{V(v_1) + V(X^*)} \end{aligned}$$

- What happens to the bias if $V(X^*)$ is relatively much larger than $V(v_1)$? What happens otherwise?

# Measurement Error in X

- In a multiple regression model, the overall measurement error in an explanatory variable produces inconsistency of all the estimators, i.e., all $\hat{\beta}_j$'s are inconsistent. In a multiple regression model with ME in only one of the $X_j$'s and this error is not correlated with either wrongly measured, say $X_m$ or with the remaining $X_j$'s $(m \neq j)$:
  - There is a bias in $\hat{\beta}_m$ (underestimate in absolute value), and it is inconsistent.
  - For the remaining $\hat{\beta}_j$'s $(m \neq j)$ they are generally inconsistent, although it is not easy to know the directions and magnitudes of the biases and inconsistency.
  - Only in the unlikely event that $X_m$ is orthogonal to the remaining $X_j$'s $(m \neq j)$, $\hat{\beta}_j$'s are consistent.

# Example 3: The impact of family income on college performance

- We want to see if the family income has an effect on the grade point average in the college. Since it is not clear that family income has a direct effect on academic performance, the recommended strategy would be to include this variable as a regressor and test whether its coefficient is significant.

- $\text{CAL} = \beta_0 + \beta_1 I^* + \beta_2 \text{PRE} + \beta_3 \text{SEL} + \varepsilon$, where
  - CAL = Average grade in college,
  - $I^*$ = Family income,
  - PRE = Average grade prior college entrance,
  - SEL = Average grade on the admission exam.

# Example 3: The impact of family income on college performance

- Suppose the data are obtained by surveying students. There may be errors in the declared family incomes, so $I = I^* + v_1$.
- Even if we assume that the measurement error, $v_1$, is uncorrelated neither with $I^*$ nor with the rest of the explanatory variables (PRE, SEL), the estimators obtained by using $I$ instead of the true value $I^*$ will be inconsistent.
- Specifically, we will underestimate $|\beta_1|$. Therefore, when testing the significance of $\beta_1$, we will be more likely to do not reject (DNR) $H_0$
- In this example, it is difficult to determine the magnitude and direction of the bias and the inconsistency for the estimators of $\beta_2$ and $\beta_3$.