# Specification Error

Class 6

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Gujarati textbook (Chapter 13, 5th edition).

# *Introduction: Specification Error*

# Types of Specification Errors

- Another classical assumption we did not deal with explicitly is that the regression model is *correctly* specified.
- If the model is not correctly specified, we encounter the problem of **model specification error** (or **model specification bias**).
- Consider the following model:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + e_i$$

  - $Y$: total cost of production
  - $X$: output
  - So the above model is the familiar textbook example of the cubic total cost function.

# Types of Specification Errors [cont'd]

- **Omitting a relevant variable**

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_i$$

  - The error is, in fact, $u_i = e_i + \beta_4 X_i^3$.

- **Including an irrelevant variable**

$$Y_i = \eta_1 + \eta_2 X_i + \eta_3 X_i^2 + \eta_4 X_i^3 + \eta_5 X_i^4 + v_i$$

  - The true model assumes $\eta_5 = 0$, so there exists an unnecessary variable.

# Types of Specification Errors [cont'd]

- **Wrong functional form**

$$\ln Y_i = \lambda_1 + \lambda_2 X_i + \lambda_3 X_i^2 + \lambda_4 X_i^3 + \nu_i$$

  - In the true model, $Y$ appears linearly whereas in the above it appears log-linearly.

- **Measurement error**

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + \beta_3^* X_i^{*2} + \beta_4 X_i^{*3} + e_i^*$$

  - where $Y_i^* = Y_i + \varepsilon_i$ and $X_i^* = X_i + \epsilon_i$
  - $\varepsilon_i$ and $\epsilon_i$ represent the measurement errors.
  - Instead of using the true $Y_i$ and $X_i$, we use their proxies $Y_i^*$ and $X_i^*$ which may contain errors of measurement.

# *Consequences of Specification Errors*

# Underfitted Model

**Omitting a Relevant Variable (Underfitted Model)**

- Suppose the true model is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$$

- But for some reason we fit the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$$

# Underfitted Model [cont'd]

$$\hat{\alpha}_2 = \frac{\sum x_{2i} y_i}{\sum x_{2i}^2}$$

- OLS estimator of $\alpha_2$ where $x_{2i} = X_{2i} - \bar{X}_2$ and $y_i = Y_i - \bar{Y}$

$$\begin{aligned}
\hat{\alpha}_2 &= \frac{\sum x_{2i} Y_i}{\sum x_{2i}^2} = \frac{\sum x_{2i} \left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i\right)}{\sum x_{2i}^2} \\
&= \beta_1 \frac{\sum x_{2i}}{\sum x_{2i}^2} + \beta_2 \frac{\sum x_{2i} X_{2i}}{\sum x_{2i}^2} + \beta_3 \frac{\sum x_{2i} X_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i} e_i}{\sum x_{2i}^2} \\
&= \beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i} e_i}{\sum x_{2i}^2}
\end{aligned}$$

# Underfitted Model [cont'd]

$$E\left(\hat{\alpha}_2\right) = \beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2}$$

- Note that if we run a regression $X_{3i} = b_{31} + b_{32} X_{2i} + u_i$, the estimator of the slope coefficient $\hat{b}_{32}$ is $\sum x_{2i} x_{3i} / \sum x_{2i}^2$.
- If $\hat{b}_{32} \neq 0$, $\hat{\alpha}_2$ is biased. Furthermore, it is inconsistent ($\hat{\alpha}_1$ too).
- If $\hat{b}_{32} = 0$, $\hat{\alpha}_2$ is unbiased and consistent (but $\hat{\alpha}_1$ is still biased and inconsistent).

# Underfitted Model [cont'd]

- How about efficiency?

$$Var\left(\hat{\alpha}_2\right) = \frac{\sigma^2}{\sum x_{2i}^2}$$

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum x_{2i}^2} \frac{1}{\left(1-\gamma_{23}^2\right)}$$

- $\gamma_{23}$ is the correlation coefficient between $X_{2i}$ and $X_{3i}$.
- Note that $1/\left(1-\gamma_{23}^2\right)$ is known as **VIF** (Variance Inflation Factor).
- $Var\left(\hat{\beta}_2\right) > Var\left(\hat{\alpha}_2\right)$: Although $\hat{\alpha}_2$ is biased, its variance is smaller! $\rightarrow$ *Trade-off*
  - Then, the OLS estimator in the underfitted model can be better than the OLS estimator in the true model?
  - Generally, **NO** (even if MSE of $\hat{\alpha}_2$ is smaller) because $\hat{\alpha}_2$ is inconsistent when $\hat{b}_{32} \neq 0$.

# Underfitted Model [cont'd]

- Suppose that $\gamma_{23} = 0$, which implies $\hat{b}_{32} = 0$.
  - $\hat{\alpha}_2$ is unbiased and consistent.
  - Variances of $\hat{\alpha}_2$ and $\hat{\beta}_2$ are the same.
  - Is there no harm in dropping the variable $X_{3i}$ from the model even though it may be relevant theoretically?
  - **NO!** We should estimate $\sigma^2$ using the residuals, and the residuals from the underfitted model are different from the true model. It means that the estimated variance of $\hat{\alpha}_2$ is biased. $\rightarrow$ the hypothesis testing procedures are likely to be wrong.

# Overfitted Model

**Inclusion of an Irrelevant Variable (Overfitted Model)**

- Suppose the true model is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + e_i$$

- But for some reason we fit the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i$$

# Overfitted Model [cont'd]

- OLS estimators of the parameters of the overfitted model are all unbiased and consistent.
  - $E(\hat{\alpha}_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$, $E(\hat{\alpha}_3) = \beta_3 = 0$
- The variance of error $\sigma^2$ is correctly estimated. $\rightarrow$ The usual hypothesis-testing procedure remain valid.
- However, the estimated $\alpha$'s will be generally inefficient, that is, their variance will be generally larger than those of the estimated $\beta$'s of the true model.
- (unwanted result) *It is better to include irrelevant variables than to omit the relevant ones.*
- Nevertheless, the best approach is to include only regressors that directly influence the dependent variable!
  - Unnecessary variables will lead to a loss in the efficiency of the estimators and the problem of multicollinearity.

# Measurement Error in Dependent Variable

## Measurement Error in the Dependent Variable

- Consider the following model:

$$Y_i^* = \beta_1 + \beta_2 X_i + e_i \tag{1}$$

- Suppose that $Y_i^*$ is not directly measurable. Instead we may use an observable $Y_i$, such that

$$Y_i = Y_i^* + \varepsilon_i$$

where $\varepsilon_i$ denote errors of measurement in $Y_i^*$.

  - For example, $Y_i^*$ is permanent consumption expenditure and $X_i$ is current income. Since we cannot directly measure the permanent expenditure, we may use an observable current expenditure $Y_i$. But there will be gaps between the permanent expenditure and the current expenditure, which are measurement errors ($\varepsilon_i$).

# Measurement Error in Dependent Variable [cont'd]

- Therefore, we estimate

$$\begin{aligned} Y_i - \varepsilon_i &= \beta_1 + \beta_2 X_i + e_i \\ \rightarrow \quad Y_i &= \beta_1 + \beta_2 X_i + e_i + \varepsilon_i \\ \rightarrow \quad Y_i &= \beta_1 + \beta_2 X_i + v_i \end{aligned} \tag{2}$$

  where $v_i = e_i + \varepsilon_i$ is a composite error term.

- Assume that $E(e_i) = E(\varepsilon_i) = 0$, $Var(e_i) = \sigma_e^2$, $Var(\varepsilon_i) = \sigma_\varepsilon^2$, $Cov(X_i, e_i) = Cov(X_i, \varepsilon_i) = 0$, and $Cov(e_i, \varepsilon_i) = 0$. (also, there is no autocorrelation in error term and measurement error.)

- Then, $v_i$ satisfies the classical assumptions. $\rightarrow \hat{\beta}_2$ in (2) exhibits unbiasedness and consistency.

# Measurement Error in Dependent Variable [cont'd]

- For the model (1):

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma_e^2}{\sum x_i^2}$$

- For the model (2):

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma_v^2}{\sum x_i^2} = \frac{\sigma_e^2 + \sigma_\varepsilon^2}{\sum x_i^2}$$

- Obviously, the latter variance is larger than the former.
- Therefore, although the measurement error in the dependent variable still gives unbiased and consistent estimator, the variances are now larger than in the case where there are no such errors of measurement.

# Measurement Error in Explanatory Variable

**Measurement Error in the Explanatory Variable**

- Consider the following model:

$$Y_i = \beta_1 + \beta_2 X_i^* + e_i \tag{3}$$

- For example, $Y_i$ is current consumption expenditure and $X_i^*$ is permanent income. Suppose instead of observing $X_i^*$, we observe current income

$$X_i = X_i^* + \eta_i$$

where $\eta_i$ represent errors of measurement in $X_i^*$.

# Measurement Error in Explanatory Variable [cont'd]

- Therefore, instead of estimating (3), we estimate

$$\begin{aligned} & Y_i = \beta_1 + \beta_2 (X_i - \eta_i) + e_i \\ \rightarrow \quad & Y_i = \beta_1 + \beta_2 X_i + e_i - \beta_2 \eta_i \\ \rightarrow \quad & Y_i = \beta_1 + \beta_2 X_i + \zeta_i \end{aligned} \tag{4}$$

  where $\zeta_i = e_i - \beta_2 \eta_i$ is a composite error term.

- Assume that $E(e_i) = E(\eta_i) = 0$, $Var(e_i) = \sigma_e^2$, $Var(\eta_i) = \sigma_\eta^2$, $Cov(X_i, e_i) = Cov(X_i, \eta_i) = 0$, and $Cov(e_i, \eta_i) = 0$. (also, there is no autocorrelation in error term and measurement error.)

# Measurement Error in Explanatory Variable [cont'd]

$$\begin{aligned} Cov\left(X_i, \zeta_i\right) &= E\left[X_i - E\left(X_i\right)\right]\left[\zeta_i - E\left(\zeta_i\right)\right] \\ &= E\left(\eta_i\right)\left(e_i - \beta_2 \eta_i\right) \\ &= E\left(-\beta_2 \eta_i^2\right) \\ &= -\beta_2 \sigma_\eta^2 \end{aligned}$$

- Thus, $X_i$ and the composite error term $\zeta_i$ are correlated $\rightarrow$ **Endogeneity issue**
- The OLS estimators are not only biased but also inconsistent.
- Therefore, measurement errors pose a serious problem when they are present in the explanatory variables.
- One suggested remedy is the use of **instrumental variable** as we discussed.