

# ECONOMETRIC

## Topic 5: SPECIFICATION ERRORS

Julio Cáceres-Delpiano

UC3M

2009/2010

- We have seen that the OLS estimator has good properties under the assumptions of linear regression model (LRM).
- What would happen if we used the LRM when it is not suitable?
- What properties does the OLS estimator have when we make some kind of specification error?
- Specification errors in which we focus are:
  - Inclusion of irrelevant variables.
  - Omission of relevant variables.
  - Measurement errors in variables.
- keeping the linearity in parameters, the different problems of incorrect specification will be seen (summarized) as a problem of omitted variable (for example, omission of a quadratic term,  $X^2$ , etc.).

# Inclusion of irrelevant variables and omission of relevant variables

## Omitted Variable Rule

- When we saw the relationship between the simple regression model and multiple regression model, we defined the relationship between the parameters of the long and the short regression.
- Here the same thing from another perspective. Given the multiple linear regression model

$$E(Y|X_1, X_2) = L(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

# Inclusion of irrelevant variables and omission of relevant variables

## Omitted Variable Rule

- For some reason (ignorance, or unobservability of the variable, etc..) we build a regression model that does not include the explanatory variable  $X_2$ :

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1,$$

thus

$$L(Y|X_1) = \gamma_0 + \gamma_1 X_1$$

(Note that in general  $E(Y|X_1, X_2)$  and  $L(Y|X_1)$  are characterized by different parameters).

# Inclusion of irrelevant variables and omission of relevant variables

## Omitted Variable Rule

- In econometrics, it is said that "it has omitted a relevant variable  $X_2$ " (if  $\beta_2 \neq 0$ ), which defines a specification error.
  - Question: ¿What are the consequences of omitting  $X_2$  on the relationship between  $Y$  and  $X_1$ ?
  - **Answer:** Instead of  $\beta_1$  we will have  $\gamma_1$ , which is related to  $\beta_1$  for the following relationship

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

This expression is called the rule of the omitted variable, which shows that the slope of a "short regression" is a linear combination of slopes of the "long regression".

# Inclusion of irrelevant variables and omission of relevant variables

## Omitted Variable Rule

- From another point of view, by omitting  $X_2$ , its effect will be part of the error term:

$$\varepsilon_1 = \varepsilon + \beta_2 X_2$$

and therefore

$$\begin{aligned} E(\varepsilon_1 | X_1) &= E(\varepsilon + \beta_2 X_2 | X_1) \\ &= E(\varepsilon | X_1) + \beta_2 E(X_2 | X_1) \\ &= E[E(\varepsilon | X_1, X_2) | X_1] + \beta_2 E(X_2 | X_1) \\ &= \beta_2 E(X_2 | X_1) \neq 0 \end{aligned}$$

Consequently,

$$E(Y | X_1) = \gamma_0 + \gamma_1 X_1 + \beta_2 E(X_2 | X_1)$$

# Inclusion of irrelevant variables and omission of relevant variables

## Omitted Variable Rule

- In general, a multiple linear regression model  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon$ , if

$$E(\varepsilon | X_1, X_2, \dots, X_K) \neq 0$$

can be understood as the model is misspecified.

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of OLS estimates under Specification Error

- We will use the results we got when we studied the relationship between the long and short regression for both the population and the sample.
- Considering the simplest multiple linear regression model ( $k = 2$ ) and confronting it with the basic bivariate model. That is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

with  $L(Y | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$  and its estimated version

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

versus

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1,$$

with  $L(Y | X_1) = \gamma_0 + \gamma_1 X_1$  and its estimated version

$$\hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 X_1$$

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of OLS estimates under Specification Error

- We know that

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

with  $\hat{\delta}_1$  the OLS estimator for the slope of  $L(X_2 | X_1) = \delta_0 + \delta_1 X_1$ .

- Moreover:

$$E(\hat{\beta}_1) = \beta_1$$

$$p \lim \hat{\beta}_1 = \beta_1$$

and

$$E(\hat{\gamma}_1) = \gamma_1$$

$$p \lim \hat{\gamma}_1 = \gamma_1$$

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of OLS estimates under Specification Error

- Therefore:

$$E(\hat{\gamma}_1 | X_1, X_2) = E(\hat{\beta}_1 | X_1, X_2) + E(\hat{\beta}_2 \hat{\delta}_1 | X_1, X_2) = \beta_1 + \beta_2 \hat{\delta}_1$$

and

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

- 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  $V(\hat{\gamma}_1) \leq V(\hat{\beta}_1)$

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of OLS estimates under Specification Error

- Whenever  $X_2$  is a “relevant” variable (that is,  $\beta_2 \neq 0$ ),  $\hat{\gamma}_1$  will be an inconsistent (and biased) of  $\beta_1$  but will have less variance than  $\hat{\beta}_1$ .  
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.

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of OLS estimates under Specification Error

- In other words, the coefficient of  $X_1$  in the regression that (incorrectly) omits  $X_2$ :
  - does not capture the ceteris paribus effect on  $Y$  due to a change of  $X_1$  (since when  $X_1$  varies also a movement in  $X_2$  is observed, because  $X_1$  and  $X_2$  are correlated).
  - captures the impact on  $Y$  coming from a change in  $X_1$  plus the effect of  $X_1$  on  $X_2$  (which finally end up having an impact on  $Y$ ).
- We can summarize the bias in estimating  $\beta_1$  when (incorrectly) omitted  $X_2$  as:

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

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of the OLS estimator: Inclusion of irrelevant variables

- If  $X_2$  is a variable "irrelevant" (ie,  $\beta_2 = 0$ ),  $\hat{\gamma}_1$  will be a consistent estimator (and unbiased) of  $\beta_1$ , and it will have less variance than  $\hat{\beta}_1$ , which will also be 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: In the population, the coefficient of an irrelevant variable is equal to 0, so that in estimating the model incorrectly by including this variable, the coefficient estimated for the other variables are not affected in the limit.

# Inclusion of irrelevant variables and omission of relevant variables

## Properties of the OLS estimator: Inclusion of irrelevant variables

- However, it generates a loss of efficiency in the estimation (the greater the larger the number of irrelevant variables to be included).
  - **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, does not generate inconsistency of the OLS estimator (and hence it is a less serious problem than the omission of relevant variables). Nevertheless, this problem can generate a serious problem when testing hypotheses of type  $H_0 : \beta_j = 0$  due to the lost of power (that is, increase the type II error), so we might infer that they are not relevant variables when they truly are.

# Inclusion of irrelevant variables and omission of relevant variables

Properties of the OLS estimator: Inclusion of irrelevant variables

- In practice, Is it possible to know which model is the appropriate one?
  - Strictly: No.
  - What It can be done is using economic theory to guide us and gather evidence for or against the "relevance" or "irrelevance" of one or more variables through the testing of hypotheses.

# Inclusion of irrelevant variables and omission of relevant variables

Properties of the OLS estimator: Inclusion of irrelevant variables

**Example 1:** The impact of tobacco consumption on cancer.

- Let say that we have a group of smokers and a group of nonsmokers. We also have information about cancer incidence for each individual
- Additionally let assume that smokers are more likely to engage in physical activity which reduces the likelihood of cancer. Nevertheless we do not observe physical activity.
- Therefore the impact of smoking on cancer incidence will be overestimated because the consumption of tobacco decreases the level of physical activity.
- Formally,  $C_i = \beta_0 + \beta_1 F_i + \beta_2 EJ_i + \varepsilon$ ,  
with,  $C_i$  as the measure of cancer incidence for the  $i$  individual,  $F_i$  is a dummy variable that takes a value of 1 for smokers and 0 otherwise, and  $EJ_i$  is a measure of physical exercise. Therefore,  $\beta_1 > 0$ ,  $\beta_2 < 0$ .

# Inclusion of irrelevant variables and omission of relevant variables

Properties of the OLS estimator: Inclusion of irrelevant variables

- Additionally,  $EJ_i = \delta_0 + \delta_1 F_i + v_i$ , with  $\delta_1 < 0$ .
- Therefore, when we run the simple regression of  $C_i$  on  $F_i$ , we will estimate

$$C_i = \gamma_0 + \gamma_1 F_i + \varepsilon_i,$$

whose estimated slope will be

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

# Inclusion of irrelevant variables and omission of relevant variables

**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 ("cigarette breaks");
    - If smoking has an impact on health outcomes, smokers would be more likely to face sick-leaves;
    - If the firm would discriminate against smokers;
    - etc.

# Inclusion of irrelevant variables and omission of relevant variables

- We have representative data for individuals 30 years old for the US. Levine, Gustafson and Velenchik (1997)<sup>1</sup> 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 in average one years less than education than non-smokers (thus, education is negatively correlated with smoking)

---

<sup>1</sup>Levine, P., T. Gustafson y A. Velenchik (1997), "More Bad News for Smokers? The Effects of Cigarette Smoking on Wages", *Industrial and Labor Relations Review*, 50(3), 493-509.

# Inclusion of irrelevant variables and omission of relevant variables

- Two specifications are considered:
  - Omitting education

$$\hat{Y}_i = -0.176 F_i \\ (0.021)$$

# Inclusion of irrelevant variables and omission of relevant variables

- Including education

$$\hat{Y}_i = -0.080 F_i + 0.070 ED_i$$

(0.021)                      (0.004)

- By not including education in the regression we overestimated the impact of smoking.

- Sometimes we have no data on the economic variable that really interests us.
- Examples:
  - We have information about the reported annual income but we lack information about annual real income.
  - According to the life model, the consumption depends on the permanent income which differs from the disposable income.
  - The marginal tax rate might be difficult to obtain for all levels of income. Nevertheless, we could use the tax for the average income in the economy.

- **Measurement Error:**

They appear when we use an unprecise measure of an economic variable in a regression model.

- How does the measurement error affect OLS estimates?

# Measurement Errors

## Measurement Error in the dependent variable

- Let's consider the following model:

$$Y^* = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

with

$$E(\varepsilon | X) = 0 \Rightarrow E(Y^* | X) = L(Y^* | X) = \beta_0 + \beta_1 X$$

therefore,  $\beta_0$  and  $\beta_1$  verify:

$$E(\varepsilon) = 0, \quad C(X, \varepsilon) = 0 \Rightarrow$$

$$\beta_0 = E(Y^*) - \beta_1 E(X) \quad \beta_1 = C(X, Y^*) / V(X)$$

# Measurement Errors

## Measurement Error in the dependent variable

- $Y^*$  is measured with error, therefore:

$$v_0 = Y - Y^* = \text{ME} \Rightarrow Y = Y^* + v_0.$$

- If we would estimate the model by running a regression of  $Y$  on  $X$ ,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X,$$

- Would  $\hat{\beta}_0$  and  $\hat{\beta}_1$  be consistent estimators of  $\beta_0$  and  $\beta_1$ ?
- **Note:** By substituting (1) in the correctly specified model we will have:

$$Y = \beta_0 + \beta_1 X + u,$$

with  $u = (\varepsilon + v_0)$ .

# Measurement Errors

## Measurement Error in the dependent variable

- Note that:

$$\begin{aligned} p \lim \hat{\beta}_1 &= \frac{p \lim \left( \frac{1}{n} \sum_i x_i y_i \right)}{p \lim \left( \frac{1}{n} \sum_i x_i^2 \right)} = \frac{C(X, Y)}{V(X)} \\ &= \frac{C(X, Y^* + v_0)}{V(X)} = \frac{C(X, Y^*) + C(X, v_0)}{V(X)}, \end{aligned}$$

- with:  $y_i = Y_i - \bar{Y}$ ,  $x_i = X_i - \bar{X}$ ,
- and therefore:

$$p \lim \hat{\beta}_1 = \frac{C(X, Y^*)}{V(X)} = \beta_1, \text{ if } C(X, v_0) = 0.$$

- Then**, if  $C(X, v_0) = 0$  (ie., the measurement error in the independent variable is not systematically related to the explanatory variable), the OLS estimator will be consistent.

# Measurement Errors

## Measurement Error in the dependent variable

- Note that:

$$\begin{aligned} p \lim \hat{\beta}_0 &= p \lim (\bar{Y} - \hat{\beta}_1 \bar{X}) = E(Y) - p \lim \hat{\beta}_1 E(X) \\ &= E(Y^* + v_0) - p \lim \hat{\beta}_1 E(X), \end{aligned}$$

- and therefore:

$$p \lim \hat{\beta}_0 = E(Y^*) - \beta_1 E(X) = \beta_0,$$

if  $E(v_0) = 0$  and  $C(X, v_0) = 0$ .

# Measurement Errors

## Measurement Error in the dependent variable

- Conclusion :

- Estimating the model by using  $Y$  instead of  $Y^*$ , our OLS estimators will be consistent and the usual inference will be valid, when:

$$C(X, v_0) = 0$$

$$E(v_0) = 0$$

- If the second condition is not satisfied we will have an inconsistent estimator of the intercept but a consistent one for the slope.
- Nevertheless the OLS estimators in respect to the model without ME will be less efficient:

$$V(Y^* | X) = V(\varepsilon | X) = \sigma^2$$

assuming that  $C(\varepsilon, v_0 | X) = 0$ , we get:

$$V(Y | X) = V(\varepsilon + v_0 | X) = V(\varepsilon | X) + V(v_0 | X) = \sigma^2 + \sigma_{v_0}^2 > \sigma^2$$

# Measurement Errors

## Measurement error in an explanatory variable

- It is usually considered a more serious problem than the measurement error in the dependent variable.
- Let's consider the model:

$$Y = \beta_0 + \beta_1 X^* + \varepsilon, \quad (2)$$

with

$$E(\varepsilon | X^*) = 0 \Rightarrow E(Y | X^*) = L(Y | X^*) = \beta_0 + \beta_1 X^*,$$

and therefore  $\beta_0$  and  $\beta_1$  verify:

$$E(\varepsilon) = 0, \quad C(X^*, \varepsilon) = 0 \Rightarrow$$

$$\beta_0 = E(Y) - \beta_1 E(X^*), \quad \beta_1 = C(X^*, Y) / V(X^*).$$

# Measurement Errors

## Measurement error in an explanatory variable

- Let say that  $X^*$  is measure with error:

$$v_1 = X - X^* = \text{ME} \Rightarrow X = X^* + v_1$$

and it is known that:

- $E(v_1) = 0$ ,
  - $C(X, \varepsilon) = 0$  ( $\varepsilon$  is not correlated neither with  $X$  nor  $X^*$ , therefore nor with  $v_1$ ).
- In terms of the conditional mean:

$$E(Y | X, X^*) = E(Y | X^*).$$

- That is:  $X$  does not affect  $Y$  as  $X^*$  take it into account.

# Measurement Errors

## Measurement error in an explanatory variable

- What are the properties of the OLS estimators that result from running a regression **of  $Y$  on  $X$** ?
  - Are  $\hat{\beta}_0$  and  $\hat{\beta}_1$  consistent estimators of  $\beta_0$  and  $\beta_1$ ?
    - The answer will depend on the assumptions made about the ME.
- **Note:** by substituting in (2) we will have:

$$Y = \beta_0 + \beta_1 X + u,$$

with

$$u = (\varepsilon - \beta_1 v_1).$$

- Let's assume that:
  - $C(X^*, v_1) = 0$  (**Classic ME assumption -CME**)
  - $C(v_1, \varepsilon) = 0$

# Measurement Errors

## Measurement error in an explanatory variable

- Therefore we will have that:

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

and therefore:

$$\text{asymptotic bias } (\hat{\beta}_1) = p \lim (\hat{\beta}_1 - \beta_1) = -\beta_1 \frac{V(v_1)}{V(X^*) + V(v_1)}.$$

# Measurement Errors

## Measurement error in an explanatory variable

- Note that when facing ME we will underestimate (in absolute value) the slope of the variable that is measured with error.
- If  $V(X^*)$  is big in relationship to  $V(v_1)$ , the lack of consistency could be negligible.
- That is: if the variability of the measurement error relative to the variability of the original explanatory variable is small, then the effect of measurement error on the consistency of the estimator may be negligible.

# Measurement Errors

## Measurement error in an explanatory variable

- In a multiple regression model, the overall measurement error in an explanatory variable produces inconsistency of all the estimated coefficients  $\hat{\beta}$ 's. In this regard, a multiple regression model in which only one of the regressors is measured with error (and this error is not correlated with either the variable measured with error or with the rest of the explanatory variables):
  - It remains the result that the slope of the explanatory variable measured with error tends to underestimate in absolute value.
  - Estimates of slopes associated with other explanatory variables are generally inconsistent, although it is not easy to know what the directions and magnitudes of the biases of inconsistency.
  - Only in the unlikely event that the explanatory variables are orthogonal to the variable measured with error, estimates of their slopes are consistent.

# Measurement Errors

## Measurement error in an explanatory variable

**Example 3:** The impact of family income on college performance.

- We want to see if the family income has an effect on the mean scores obtained in college.
- 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 zero.

$$CAL = \beta_0 + \beta_1 I^* + \beta_2 PRE + \beta_3 SEL + \varepsilon, \quad \text{where}$$

- $CAL$  = Average grade in college,
- $I^*$  = Family income,
- $PRE$  = Average grade prior college entrance,
- $SEL$  = Average grade on the admission exam.

# Measurement Errors

## Measurement error in an explanatory variable

- Suppose the data are obtained directly by surveying students.
  - You may declare the family's income incorrectly, so  $I = I^* + v$ .
- Even assuming that the measurement error,  $v$ , is uncorrelated neither with  $I^*$  nor with the rest of the explanatory variables ( $PRE$ ,  $SEL$ ), the estimates of the parameters using  $I$  (instead  $I^*$ , which is unobserved) will be inconsistent.
  - Specifically we will underestimate  $\beta_1$ .
    - Therefore, by testing  $H_0 : \beta_1 = 0$ , will be more likely that we could not reject  $H_0$ .
  - In this example, it is difficult to determine the magnitude and direction of bias and inconsistency of the estimator of  $\beta_2$  and  $\beta_3$ .