Lecture 5

Endogeneity

Nurgul Tilenbaeva

American University - Central Asia

24.10.2022; 27.10.2022

Contents

1. What is Endogeneity?
2. Sources of Endogeneity:
    - Specification Error
    - Measurement Error
    - Omitted Variables
    - Reversed Causality
    - Sample Selection

What is Endogeneity?

Assumption MLR.4. Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$E(u|x_1, x_2, ..., x_k) = 0.$

All factors in the unobserved error term must be uncorrelated with the explanatory variables.

If x_j is uncorrelated with u (i.e. when MLR.4. holds): **exogenous explanatory variables**
If x_j is correlated with u: **endogenous explanatory variables**

Question: Why is Assumption MLR.4 so important?

What is Endogeneity?

Endogeneity

- Failure of Assumption MLR.4;
- Situation in which one or more of the explanatory variables are correlated with the error term.

Sources of Endogeneity

- Specification Error
- Measurement Error
- Omitted Variables
- Reversed Causality
- Sample Selection

Sources of Endogeneity: **Specification Error**

Usually, we assume the linear relationship between y and x_j:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

However, a more general specification of the theoretical model could be the following:

$y = f(x_1, x_2)$

where

$f(.)$ is a generic function.

Linear specification is often taken as a simple approximation of more sophisticated or perhaps unknown functional forms.

Sources of Endogeneity: **Specification Error**

If we impose a linear specification, one simple way of rewriting the general model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u + \underbrace{[f(x_1, x_2) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2)]}_{specification\ \ error}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u + \epsilon$$

$\Rightarrow$ The error term ϵ is correlated with $x_1, x_2 \Rightarrow$ **endogeneity**

Sources of Endogeneity: **Specification Error**

Examples of Specification Error:

- If we forget to include the quadratic term in the model;
- If we use the level of the variable when the log of the variable is what actually shows up in the population model, or vice versa;
- If y is some non-linear function of x_j (logistic), but we use the linear function instead.

Sources of Endogeneity: **Specification Error**

RESET as a General Test for Functional Form Misspecification:

The idea behind **regression specification error test (RESET)** is fairly simple. If the original model

$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$

satisfies MLR.4, then no nonlinear functions of the independent variables should be significant when added to the equation.

To implement RESET, we must decide how many functions of the fitted values to include in an expanded regression. There is no right answer to this question, but the squared, cubed and fourth power terms have proven to be useful in most applications.

Sources of Endogeneity: **Specification Error**

RESET as a General Test for Functional Form Misspecification:

Let $\hat{y}$ denote the OLS fitted values from estimating the original model. Consider the expanded equation

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \delta_3 \hat{y}^4 + error$$

We use this equation to test whether the original equation has missed important nonlinearities.

The null hypothesis is that the original equation is correctly specified. Thus, RESET is the F statistic for testing $H_0 : \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ in the expanded model. A significant F statistic suggests some sort of functional form problem.

The distribution of the F statistic is approximately $F_{q,n-k-1}$ in large samples under the null hypothesis (and the Gauss-Markov assumptions).

Sources of Endogeneity: **Specification Error**

Tests against Nonnested Alternatives

We want to test the model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

against the model

$y = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + u$

and vice versa.

Sources of Endogeneity: **Specification Error**

Tests against Nonnested Alternatives

The first approach by **Mizon and Richard (1986)** is to construct a comprehensive model that contains each model as a special case and then to test the restrictions that led to each of the models. In this example, the comprehensive model is:

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 log(x_1) + \gamma_4 log(x_2) + u$$

We can first test $H_0 : \gamma_3 = 0, \gamma_4 = 0$ as a test of the first model.
We can also test $H_0 : \gamma_1 = 0, \gamma_2 = 0$ as a test of the second model.

Sources of Endogeneity: **Specification Error**

Tests against Nonnested Alternatives

The second approach by Davidson and MacKinnon (1981) is based on the idea that if the first model is true, then the fitted values from the second model should be insignificant in the first model. Thus, to test the first model, we first estimate the second model by OLS to obtain the fitted values. Call these $\hat{\hat{y}}$. Then, the **Davidson-MacKinnon test** is based on the t statistic on $\hat{\hat{y}}$ in the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + error$$

A significant t statistic (against a two-sided alternative) is a rejection of the first model. Repeat the same procedure to test the second model.

Sources of Endogeneity: **Specification Error**

Tests against Nonnested Alternatives: Problems

1. A clear winner need not emerge. Both models could be rejected or neither model could be rejected.
2. In the latter case, we can use the adjusted R-squared to choose between them.
3. Rejecting the first model using, say, the Davidson-MacKinnon test does not mean that the second model is the correct model. The first model can be rejected for a variety of functional form misspecifications.

Sources of Endogeneity: **Measurement Error**

Suppose, you observe one or more of your variables with error.
For example, y may be measured with error so that what we really observe is not y but a noisy version of it, call it y^*:

$$y^* = y + \epsilon$$

where ϵ is a random error of measurement.
In this case, the model that we can effectively bring to the data needs to be specified in terms of y^*, which is observable, rather than y, which we cannot observe. So, let's assume that the

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Now:

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u + \underbrace{\epsilon}_{measurement\ \ error}$$

Sources of Endogeneity: **Measurement Error**

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u + \underbrace{\epsilon}_{measurement\ \ error}$$

- If ϵ is correlated with $x_1, x_2 \Rightarrow$ **endogeneity**;
- If ϵ is uncorrelated with $x_1, x_2 \Rightarrow$ **no endogeneity**;
  - Examples: error arises from misreporting, errors in the coding of the variables, etc.

Other implication of measurement error: estimators with larger standard errors.

Sources of Endogeneity: **Measurement Error**

Now suppose that x_1 is measured with error, so that what we observe is a noisy version of it, call it x_1^*

$$x_1^* = x_1 + \eta$$

Again, the operational version of the model has to be written in terms of the observable variables, so we get

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + u \\ y &= \beta_0 + \beta_1 (x_1^* - \eta) + u \\ y &= \beta_0 + \beta_1 x_1^* + u + \underbrace{[-\beta_1 \eta]}_{measurement\ \ error} \end{aligned}$$

Sources of Endogeneity: **Measurement Error**

$$y = \beta_0 + \beta_1 x_1^* + u + \underbrace{[-\beta_1 \eta]}_{measurement\ \ error}$$

Assume that the error of measurement is *classicial*, i.e. zero mean and uncorrelated with the true variable:

$Cov(x_1, \eta) = 0$
$Cov(u, \eta) = 0$

Then, $Cov(x_1^*, \eta) = Cov[(x_1 + \eta), \eta] = \sigma_\eta^2 \Rightarrow$ **endogeneity**

where σ_η^2 is the variance of the error of measurement.

- What we obtain is an under-estimate of the true parameter.
- The bias arising from this type of measurement is called **attenuation bias**.

Sources of Endogeneity: **Measurement Error**

Examples of Measurement Error:

- There may be errors in coding the data;
- There may be errors made by survey respondents in answering questions (income level).

Sources of Endogeneity: **Omitted Variables**

Suppose, there is a variable x_2, which is indeed a determinant of y (i.e. $\beta_2 \neq 0$ but that we have omitted it from our model, because:

- we wanted to simplify our analysis;
- we forgot about it;
- it was not available in our data.

So the true model would be

$$y = \beta_0 + \beta_1 x_1 + \underbrace{[\beta_2 x_2]}_{error\ \ term}$$

If x_2 is correlated with x_1 ⇒ **endogeneity**

Sources of Endogeneity: **Omitted Variables**

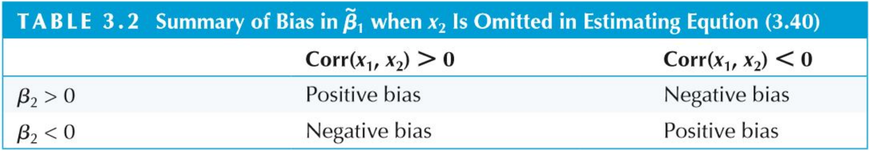

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ when x_2 Is Omitted in Estimating Eqution (3.40)

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- If $E(\tilde{\beta}_1) > \beta_1$, then we say that $\tilde{\beta}_1$ has an **upward bias**;
- If $E(\tilde{\beta}_1) < \beta_1$, then we say that $\tilde{\beta}_1$ has a **downward bias**;
- **Biased towards zero** refers to cases where $E(\tilde{\beta}_1)$ is closer to zero than is β_1.

Sources of Endogeneity: **Omitted Variables**

Example: Hourly wage equation

$log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$

satisfies Assumptions MLR.1 through MLR.4.

However, the data set does not contain data on ability, so we estimate β_1 from the simple regression

$log(\tilde{wage}) = 0.584 + 0.083 educ$
$n = 526, R^2 = 0.186$

$\Rightarrow$ Most likely, 0.083 is greater than the true β_1.

Sources of Endogeneity: **Omitted Variables**

Rules of thumb for including variables in the model:

- Include only those variables that are likely to be correlated with our variables of interest (and that are also likely to influence the outcome y);
- Any other variable, even if it explains a lot of the variation in y, is not necessary and can be safely omitted;
- The only advantage of adding explanatory variables that are uncorrelated to the variables of interest, is *efficiency*;
- If there is a lot of unexplained variation, the standard errors of the estimated coefficients will be larger.

Sources of Endogeneity: **Omitted Variables**

Using Proxy Variables for Unobserved Explanatory Variables:

$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$

However, *abil* is not observed. How can we solve the omitted variables bias?

One possibility is to obtain a **proxy variable** for the omitted variable. It is something that is related to the unobserved variable that we would like to control for in our analysis.

In the wage equation, one possibility is to use the intelligence quotient, or IQ, as a proxy for ability. This *does not* require IQ to be the same thing as ability; what we need is for IQ to be correlated with ability.

Sources of Endogeneity: **Omitted Variables**

Using Proxy Variables for Unobserved Explanatory Variables:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$

y, x_1, x_2 are observed.
x_3^* is unobserved, but we have a proxy variable x_3:

$x_3^* = \delta_0 + \delta_3 x_3 + v_3$

We run the regression of

y on x_1, x_2, x_3.

We call this the **plug-in solution to the omitted variables problem** because x_3 is just plugged in for x_3^* before we run OLS.

Sources of Endogeneity: **Omitted Variables**

Using Proxy Variables for Unobserved Explanatory Variables:

Assumptions needed for the plug-in solution to provide consistent estimators of β_1 and β_2:

1. u is uncorrelated with x_1, x_2 and x_3^*.
2. u is uncorrelated with x_3, i.e. it is x_3^* that directly affects y, not x_3.
3. v_3 is uncorrelated with x_1, x_2 and x_3, i.e. x_3^* has zero correlation with x_1 and x_2 once x_3 is partialled out. In our example, the average level of ability only changes with *IQ*, not with *educ* and *exper*.

Sources of Endogeneity: **Omitted Variables**

Using Proxy Variables for Unobserved Explanatory Variables:

Now plug in the equation for x_3^* into the main equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\delta_0 + \delta_3 x_3 + v_3) + u$$
$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3$$

Call the composite error $e = u + \beta_3 v_3$. Then,

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e,$$

where $\alpha_0 = (\beta_0 + \beta_3\delta_0)$ is the new intercept and $\alpha_3 = \beta_3\delta_3$ is the slope parameter on the proxy variable x_3.

$\Rightarrow$ we get unbiased (or at least consistent) estimators of $\alpha_0, \beta_1, \beta_2, \alpha_3$.

Sources of Endogeneity: **Omitted Variables**

Example: IQ as a proxy for ability

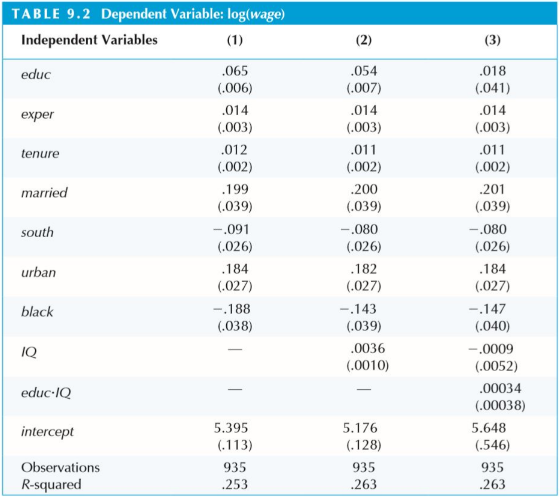

TABLE 9.2 Dependent Variable: log(*wage*)

Independent Variables	(1)	(2)	(3)
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	−.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	−.188 (.038)	−.143 (.039)	−.147 (.040)
IQ	—	.0036 (.0010)	−.0009 (.0052)
educ·*IQ*	—	—	.00034 (.00038)
intercept	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observations *R*-squared	935 .253	935 .263	935 .263

Sources of Endogeneity: **Reversed Causality**

$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 z + u$

where x_1 is exogenous.

For some reason z is being itself affected by y:

$z = \beta_0 + \beta_1 h_1 + \beta_2 y + v$

Then,

$Cov(z, u) = Cov(\beta_0 + \beta_1 h_1 + \beta_2 y + v, y - \alpha_0 - \alpha_1 x_1 - \alpha_2 z) \neq 0$

$\Rightarrow$ **endogeneity**

Sources of Endogeneity: **Sample Selection**

So far we have always assumed **random sampling** (i.e. MLR.2. is satisfied). However, there are many cases when we have a **nonrandom sample** from the population.

Examples:

- Due to missing data:
    - We are studying the returns to education. What if the probability that education is missing is higher for those people with lower than average levels of education?
    - What if obtaining an IQ score is easier for those with higher IQs?

$\Rightarrow$ the sample is not representative of the population (MLR.2. is violated).

Sources of Endogeneity: **Sample Selection**

What are the consequences for OLS estimation?

- Under the Gauss-Markov assumptions (but without MLR.2.) the sample can be chosen on the basis of **independent variables** without causing any statistical problems. This is called sample selection based on the independent variables, and is an example of **exogenous sample selection**.

$\Rightarrow$ no bias or inconsistency in OLS.

Sources of Endogeneity: **Sample Selection**

Example:

$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u$

Suppose that our data set was based on a survey of people over 35 years of age, thereby leaving us with a nonrandom sample of all adults. While this is not ideal, we can still get *unbiased* and *consistent* estimators of the parameters in the population model, using the nonrandom sample.

Sources of Endogeneity: **Sample Selection**

What are the consequences for OLS estimation?

- When selection is based on the **dependent variable**, y, it is called sample selection based on the dependent variable, and is an example of **endogenous sample selection**.

$\Rightarrow$ bias and inconsistency in OLS.

Sources of Endogeneity: **Sample Selection**

Example:

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u$$

Suppose that only people with wealth below $250,000 are included in the sample. This is a nonrandom sample from the population of interest, and it is based on the dependent variable. Using this sample will result in *biased* and *inconsistent* estimators of the parameters.

Sources of Endogeneity: **Sample Selection**

Other sampling schemes lead to nonrandom samples from the population, usually intentionally. A common method of data collection is stratified sampling, in which the population is divided into nonoverlapping, exhaustive groups, or strata. Then, some groups are sampled more frequently than is dictated by their population representation, and some groups are sampled less frequently.

Example:

Some surveys purposely oversample minority groups or low-income groups.

Sources of Endogeneity: **Sample Selection**

Whether special methods are needed again depends on whether the stratification is **exogenous** (based on exogenous explanatory variables) or **endogenous** (based on the dependent variable).

Example:

- Suppose that a survey of military personnel oversampled women because the initial interest was in studying the factors that determine pay for women in the military.

  ⇒ OLS is *unbiased* and *consistent* because the stratification is with respect to an explanatory variable, namely, gender.

- If, instead, the survey oversampled lower-paid military personnel.

  ⇒ OLS using the stratified sample does not consistently estimate the parameters of the military wage equation. In such cases, special econometric methods are needed.

Sources of Endogeneity: **Sample Selection**

Other sample selection issues are more subtle.

Example:

Labor economists are often interested in estimating the effect of education on the wage offer. The idea is this: every person of working age faces an hourly wage offer, and he or she can either work at that wage or not work.

- For someone who does work, the wage offer is just the wage earned.
- For people who do not work, we usually cannot observe the wage offer.

Sources of Endogeneity: **Sample Selection**

Example (cont.):

$log(wage^{o}) = \beta_0 + \beta_1 educ + \beta_2 exper + u$

Since this wage offer equation represents the population of all working-age people, we cannot estimate it: we have data on the wage offer only for working people.

If we use a random sample on working people to estimate this equation, will we get unbiased estimators?

- Since the sample is selected based on someone's decision to work (as opposed to the size of the wage offer), this is an exogenous selection.
- However, since the decision to work might be related to unobserved factors that affect the wage offer, selection might be endogenous, and this can result in a sample selection bias in the OLS estimators.