Endogenous explanatory variables

Violation of the orthogonality condition that $\mathrm{E}(x_i u_i) = 0$ has serious consequences for the OLS estimator

This is one of the key assumptions needed to establish consistency

When one or more of the explanatory variables is correlated with the error term u_i, we have both $\mathrm{E}(u_i|x_i) \neq 0$ and $\mathrm{E}(x_i u_i) \neq 0$, so the OLS estimator will be both biased and inconsistent

We will consider two situations where this occurs:

- a linear model with $\mathrm{E}(x_i u_i) = 0$ is the correct specification, but one or more of the explanatory variables is measured with error

- a linear model with $\mathrm{E}(x_i u_i) = 0$ is the correct specification, but one or more of the explanatory variables is not measured at all, and hence omitted from the model we can estimate

These are simply two *examples* of cases where we encounter **endogeneity** (or 'simultaneity'), i.e. one or more of the explanatory variables is correlated with the error term

Measurement error/errors-in-variables

A common concern in applied econometrics is that relevant explanatory variables may be poorly measured

Examples - survey data on households:

- recall bias: how much time did you spend unemployed last year?

- rounding bias: how much money did you spend on food last week?

Illustrate 'attenuation bias' for the case of a single explanatory variable, measured with error

- the OLS estimator is biased towards zero if the explanatory variable is measured with error

- this bias does not disappear in large samples (OLS is inconsistent)

Note that measurement error in the dependent variable does not lead to the same bias and inconsistency problems, provided the measurement error in y_i is uncorrelated with (correctly measured) x_i

Consider the model with a single explanatory variable and no intercept

$$y_i^* = x_i^* \beta + u_i \qquad \text{for } i = 1, ..., n$$

where y_i^* and x_i^* denote the true values of these variables, which we may not observe

To simplify, suppose $\mathrm{E}(u_i) = \mathrm{E}(x_i^*) = \mathrm{E}(y_i^*) = 0$ for $i = 1, ..., n$ (original variables may be expressed as deviations from their sample means)

We focus on large sample properties, and assume that $\mathrm{E}(x_i^* u_i) = 0$ for $i = 1, ..., n$, and we have independently and identically distributed data, so that $\widehat{\beta}_{\mathrm{OLS}}$ would be a consistent estimator of β here if we observed the true values of y_i^* and x_i^*

First consider additive, mean zero measurement error in the dependent variable only

$$y_i = y_i^* + v_i \quad \leftrightarrow \quad y_i^* = y_i - v_i$$

y_i is the observed value

y_i^* is the true value

v_i is the measurement error, with $\mathrm{E}(v_i) = 0$ for $i = 1, ..., n$

The true values x_i^* are observed

Substituting this expression for y_i^* in the true model

$$(y_i - v_i) = x_i^* \beta + u_i$$

or

$$y_i = x_i^* \beta + (u_i + v_i)$$

Consistency requires x_i^* to be uncorrelated with the error term $(u_i + v_i)$

Given $\mathrm{E}(x_i^* u_i) = 0$, the additional requirement is that $\mathrm{E}(x_i^* v_i) = 0$ for $i = 1, ..., n$

That is, the measurement error in the dependent variable is uncorrelated with the explanatory variable

Now consider additive, mean zero measurement error (e_i, with $\mathrm{E}(e_i) = 0$) in the explanatory variable (only)

$$x_i = x_i^* + e_i \quad \leftrightarrow \quad x_i^* = x_i - e_i$$

Substituting for x_i^* in the true model

$$y_i^* = (x_i - e_i)\beta + u_i$$

$$\text{or} \qquad y_i^* = x_i\beta + (u_i - e_i\beta)$$

The OLS estimator of β here is biased and inconsistent

- for a given value of x_i^*, observed x_i and the measurement error e_i are positively correlated, which implies non-zero correlation between x_i and the error term in this model $(u_i - e_i\beta)$

$$y_i^* = x_i\beta + (u_i - e_i\beta)$$

For $\beta > 0$, this implies a negative correlation between x_i and $(u_i - e_i\beta)$

For $\beta < 0$, this implies a positive correlation between x_i and $(u_i - e_i\beta)$

For $\beta > 0$, the OLS estimator of β will be biased downwards

For $\beta < 0$, the OLS estimator of β will be biased upwards

In either case, the OLS estimator of β will be biased *towards zero*

- this is known as 'attenuation bias'

To analyse this further, we invoke the classical errors-in-variables assumptions (for $i = 1, ..., n$)

$\mathrm{E}(x_i^* e_i) = 0$ Measurement error is uncorrelated with the true value of x_i^*

$\mathrm{E}(u_i e_i) = 0$ Measurement error is uncorrelated with the true model error u_i

$\mathrm{Var}(e_i) = \sigma_e^2$ Measurement error is homoskedastic

$\mathrm{Var}(x_i^*) = \sigma_{x^*}^2$ Population variance of the true x_i^* exists and is finite

$$\text{Now} \qquad \widehat{\beta}_{\mathrm{OLS}} = (X'X)^{-1}X'y^* = \frac{\sum_{i=1}^{n} x_i y_i^*}{\sum_{i=1}^{n} x_i^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i^*}{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Using $x_i = x_i^* + e_i$ and $y_i^* = x_i^*\beta + u_i$ together with the above assumptions, we obtain

$$\operatorname{p\,lim}_{n\to\infty} \widehat{\beta}_{\mathrm{OLS}} = \frac{\operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} (x_i^* + e_i)(x_i^* \beta + u_i)}{\operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} (x_i^* + e_i)^2}$$

$$= \frac{\left(\operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} x_i^{*2} \right) \beta + \operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} x_i^* u_i + \left(\operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} x_i^* e_i \right) \beta + \operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} u_i e_i}{\operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} x_i^{*2} + 2 \operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} x_i^* e_i + \operatorname{p\,lim} \frac{1}{n} \sum_{i=1}^{n} e_i^2}$$

$$= \frac{\mathrm{E}(x_i^{*2})\beta + \mathrm{E}(x_i^* u_i) + \mathrm{E}(x_i^* e_i)\beta + \mathrm{E}(u_i e_i)}{\mathrm{E}(x_i^{*2}) + 2\mathrm{E}(x_i^* e_i) + \mathrm{E}(e_i^2)} = \frac{\mathrm{E}(x_i^{*2})\beta + 0 + 0 + 0}{\mathrm{E}(x_i^{*2}) + 0 + \mathrm{E}(e_i^2)}$$

$$= \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right) \beta = \frac{\beta}{1 + (\sigma_e^2 / \sigma_{x^*}^2)} \neq \beta \qquad \text{if } \sigma_e^2 > 0$$

$$\operatorname{p}\lim_{n\to\infty} \widehat{\beta}_{\text{OLS}} = \frac{\beta}{1 + (\sigma_e^2/\sigma_{x^*}^2)} < \beta \qquad \text{for } \beta > 0 \text{ and } \sigma_e^2 > 0$$

$$\operatorname{p}\lim_{n\to\infty} \widehat{\beta}_{\text{OLS}} = \frac{\beta}{1 + (\sigma_e^2/\sigma_{x^*}^2)} > \beta \qquad \text{for } \beta < 0 \text{ and } \sigma_e^2 > 0$$

The OLS estimator of β is inconsistent, with a bias towards zero that does not diminish as the sample becomes large

For given $\sigma_{x^*}^2$, the severity of this 'attenuation bias' increases with the variance of the measurement error (σ_e^2)

The magnitude of the inconsistency depends inversely on the 'signal-to-noise' ratio $(\sigma_{x^*}^2/\sigma_e^2)$

Under the classical errors-in-variables assumptions with homoskedastic measurement error, the presence of measurement error affects the estimated slope parameter, but not the linearity of the relationship between y_i^* and observed x_i

With heteroskedastic measurement error, the presence of measurement error may also introduce an incorrect indication of non-linearity in the relationship

For example, if $\beta > 0$ and $\text{Var}(e_i)$ tends to be larger for individuals with higher values of x_i^*, then estimation of a non-linear relationship between y_i^* and observed x_i could give an incorrect indication of a concave relationship (illustrate)

Multiple regression with errors in variables

$$y_i^* = x_i^{*\prime}\beta + u_i$$

$$x_i' = x_i^{*\prime} + e_i'$$

where x_i', $x_i^{*\prime}$ and e_i' are $1 \times K$ vectors

As before

$$y_i^* = x_i'\beta + (u_i - e_i'\beta)$$

In general, the OLS estimator of the $K \times 1$ parameter vector β will be biased and inconsistent, since $\mathrm{E}[x_i(u_i - e_i'\beta)] \neq 0$

If only one of the explanatory variables in x_i is measured with error, we can show that

- the OLS estimator of the coefficient on that variable is biased towards zero

- the OLS estimator of the coefficients on the other explanatory variables are also biased, in unknown directions

If several explanatory variables are measured with error, it is very difficult to sign the biases for any of the coefficients

Omitted variables

Another common concern in applied econometrics is that relevant explanatory variables may be omitted from the model

Relevant explanatory variables are often unobserved or unobservable

Example:

- survey data on individuals do not contain data on characteristics like ability or motivation

This may make it difficult to attach causal significance to estimated parameters in linear regression-type models

Illustrate omitted variable bias for the case of a single included variable and a single omitted variable

- the OLS estimator is biased if the omitted variable is relevant and correlated with the included regressor

- this bias does not disappear in large samples (OLS is inconsistent)

- the direction of the bias depends on the sign of the correlation between the included variable and the omitted variable

Consequently omitted variables - or 'unobserved heterogeneity' - presents a formidable challenge to drawing causal inferences from cross-section regressions

There is a serious danger that observed, included explanatory variables may just be proxying for unobserved, omitted factors - rather than exerting a direct, causal influence on the outcome of interest

Note that this problem is not confined to empirical research in economics

Beware of medical studies claiming that some activity will help you to live longer

These claims are often based on cross-section correlations

It is difficult to draw causal conclusions unless we are confident that the study has controlled for all potentially relevant confounding factors

We first consider a model with one included variable (x_{1i}) and one omitted variable (x_{2i})

The true process which generates the outcomes y_i is assumed to be

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + u_i \qquad \text{for } i = 1, 2, ..., n$$

satisfying $\mathrm{E}(u_i) = \mathrm{E}(x_{1i}) = \mathrm{E}(x_{2i}) = 0$ and $\mathrm{E}(x_{1i}u_i) = \mathrm{E}(x_{2i}u_i) = 0$

However the model we estimate excludes x_{2i}

$$y_i = x_{1i}\beta_1 + (u_i + x_{2i}\beta_2) \qquad \text{for } i = 1, 2, ..., n$$

Illustration suggests that the OLS estimator $\widehat{\beta}_1$ in the estimated model will be a biased and inconsistent estimator of β_1 in the true model, in cases where (omitted) x_{2i} and (included) x_{1i} are correlated, and where $\beta_2 \neq 0$

Stack across the n observations to obtain

$$y = X_1\beta_1 + (u + X_2\beta_2) \qquad \text{(all vectors are } n \times 1\text{)}$$

The OLS estimator of β_1 is

$$\widehat{\beta}_1 = (X_1'X_1)^{-1}X_1'y$$

Substituting for $y = X_1\beta_1 + X_2\beta_2 + u$ from the true model

$$\begin{aligned}
\widehat{\beta}_1 &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\
&= \beta_1 + \left[(X_1'X_1)^{-1}X_1'X_2\right]\beta_2 + (X_1'X_1)^{-1}X_1'u \\
&= \beta_1 + \widehat{\delta}\beta_2 + (X_1'X_1)^{-1}X_1'u
\end{aligned}$$

where $\widehat{\delta} = (X_1'X_1)^{-1}X_1'X_2$ is the OLS estimator of ...

...the coefficient δ in a linear projection of the omitted variable x_{2i} on the included variable x_{1i}, i.e.

$$x_{2i} = x_{1i}\delta + e_i$$

Taking probability limits, and using $\mathrm{E}(x_{1i}u_i) = 0$, we obtain

$$\mathrm{p}\lim_{n\to\infty} \widehat{\beta}_1 = \beta_1 + (\mathrm{p}\lim_{n\to\infty} \widehat{\delta})\beta_2 = \beta_1 + \delta\beta_2$$

The OLS estimator of β_1 in the model that omits x_{2i} is inconsistent unless:

- either $\delta = 0$ (the omitted variable is orthogonal to the included variable)

- or $\beta_2 = 0$ (the omitted variable is not a relevant explanatory variable in the true model)

$$\text{p} \lim_{n \to \infty} \widehat{\beta}_1 = \beta_1 + \delta \beta_2$$

Thus if we omit a relevant explanatory variable ($\beta_2 \neq 0$), the only case in which $\widehat{\beta}_1$ remains a consistent estimator of the true, causal parameter β_1 is the case where $\delta = 0$, i.e. where x_{1i} and x_{2i} are uncorrelated

If x_{1i} and x_{2i} are positively correlated, we have $\delta > 0$

If $\beta_2 > 0$, we expect (in large finite samples) an upward bias in $\widehat{\beta}_1$

Intuitively, the OLS estimator $\widehat{\beta}_1$ picks up an indirect relationship between x_{1i} and y_i, due to the fact that the included x_{1i} proxies for the omitted x_{2i}, as well as the direct causal effect of x_{1i} on y_i at a given level of x_{2i}, measured by β_1 in the true model (cf. illustration)

$$\text{p} \lim_{n \to \infty} \widehat{\beta}_1 = \beta_1 + \delta \beta_2$$

Conversely if x_{1i} and x_{2i} are negatively correlated ($\delta < 0$) and $\beta_2 > 0$, or if x_{1i} and x_{2i} are positively correlated ($\delta > 0$) and $\beta_2 < 0$, we expect a downward bias in the OLS estimator $\widehat{\beta}_1$

Thus if linear regression-type models omit relevant (but perhaps unmeasured) explanatory factors that are correlated with the included (measured) regressors, we cannot draw causal inferences from the pattern of partial correlations among the observed variables

Some examples

- do individuals with lots of education tend to have high earnings because education raises their productivity and wages, or because intrinsically high ability (high productivity) individuals also tend to have lots of education?

- do countries with high investment tend to have high per capita income levels because investment raises income, or because (for example) countries with good institutions tend to have both high investment and high incomes?

In the first example, causality may run from ability to both education and earnings, rather than from education to earnings

In the second example, causality may run from institutions to both investment and incomes, rather than from investment to incomes

Since it is very difficult to control adequately for individual ability or the quality of national institutions, we should be very cautious about drawing any causal inference from statistically significant coefficients reported in such cross-section regression studies

Multiple regression with omitted variables

The analysis proceeds in a similar way

$$y = X_1\beta_1 + (X_2\beta_2 + u)$$

where now X_1 is $n \times K_1$, β_1 is $K_1 \times 1$, X_2 is $n \times K_2$ and β_2 is $K_2 \times 1$ (i.e. there are K_1 included regressors and K_2 omitted variables)

As in the simpler example we can obtain

$$\text{p}\lim_{n\to\infty} \widehat{\beta}_1 = \beta_1 + (\text{p}\lim_{n\to\infty} (X_1'X_1)^{-1}X_1'X_2)\beta_2$$

Each column of the $K_1 \times K_2$ matrix $(X_1'X_1)^{-1}X_1'X_2$ is the $K_1 \times 1$ vector of OLS estimates of the coefficients in a multiple regression of the corresponding column of X_2 on all of the included variables in X_1

The general point is that it is harder to be confident about the *direction* of the biases expected in the estimated β_1 coefficients

If there is only one omitted variable ($K_2 = 1$), which is correlated with several of the included explanatory variables, we can show that the bias in each of the estimated coefficients on the included regressors will depend on their *partial* correlation with the omitted variable, not on the *simple* correlation between the included regressor and the omitted variable

i.e. direction of biases depends on the sign of coefficients in a multiple regression of the omitted variable on all the included regressors jointly

- not on the sign of coefficients in a set of simple regressions relating the omitted variable to each of the included regressors individually

Single omitted variable

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_{K-1} x_{K-1,i} + (\beta_K x_{Ki} + u_i)$$

where $\mathrm{E}(x_{ki} u_i) = 0$ for $k = 1, ..., K$ (and $x_{1i} = 1$ for all $i = 1, ..., n$)

[Relation to general model: $K_1 = K - 1$, $K_2 = 1$]

Linear projection of omitted x_{Ki} on all the included variables

$$x_{Ki} = \delta_1 + \delta_2 x_{2i} + ... + \delta_{K-1} x_{K-1,i} + v_i$$

s.t. $\mathrm{E}(x_{ki} v_i) = 0$ for $k = 1, ..., K - 1$ (by definition of linear projection)

Multiply x_{Ki} by β_K and subsititute:

$$y_i = (\beta_1 + \beta_K \delta_1) + (\beta_2 + \beta_K \delta_2) x_{2i} + ... + (\beta_{K-1} + \beta_K \delta_{K-1}) x_{K-1,i} + (u_i + \beta_K v_i)$$

$$y_i = (\beta_1 + \beta_K \delta_1) + (\beta_2 + \beta_K \delta_2) x_{2i} + ... + (\beta_{K-1} + \beta_K \delta_{K-1}) x_{K-1,i} + (u_i + \beta_K v_i)$$

Now since $\mathrm{E}[x_{ki}(u_i + \beta_K v_i)] = 0$ for $k = 1, ..., K-1$, we have

$$\mathrm{p} \lim_{n \to \infty} \widehat{\beta}_k = \beta_k + \beta_K \delta_k \qquad \text{for } k = 1, ..., K-1$$

$$(\text{Or equivalently } \mathrm{p} \lim_{n \to \infty} \widehat{\beta}_k = \beta_k + (\mathrm{p} \lim_{n \to \infty} \widehat{\delta}_k) \beta_K)$$

The inconsistency thus depends on the sign of the *partial* correlations (reflected in the sign of the δ_k coefficients)

- not on the sign of the simple correlations between each included x_{ki} and the omitted variable x_{Ki}

Multiple omitted variables

If there are several omitted variables, it is very difficult to predict the direction of the biases

But the OLS estimator $\widehat{\beta}_1$ is a biased and inconsistent estimator of β_1 in the true model, except in the (very) special case where *all* of the omitted variables are orthogonal to *all* of the included variables

Simultaneity bias

Note that both omitted variables and measurement error (in the explanatory variable(s)) result in *correlation* between the included explanatory variable(s) and the error term in the estimated model

These are both examples of the more general phenomenon of 'endogeneity' or 'simultaneity' - sources of correlation between the explanatory variable(s) and the error term, such that the OLS estimator is biased and inconsistent

This can also arise naturally in situations where the dependent variable and at least one of the explanatory variables are *chosen jointly* as part of the same decision problem

Example:

- firms choosing inputs and output jointly in models of production, where the error term includes unobserved (total factor) productivity

- high productivity firms are likely to be larger, using more inputs, as well as producing more output from given inputs

- expect OLS estimates of coefficients on the inputs to be biased and inconsistent in production functions (Marschak & Andrews, *Econometrica*, 1944)