Mean Independence Violations

Econ 671

Purdue University
To this point, we have discussed variations on the second moment assumption, \( E(\epsilon \epsilon' | X) = \sigma^2 I_n \), but have continued to assume that mean independence holds.

It is important to recognize that this is an assumption and will not always be satisfied.

In this lecture we review some popular models / problems where mean-independence is violated. Next lecture we will then discuss instrumental variables as a potential solution to the mean-independence problem.
The models / problems where mean independence fails include:

1. The omission of relevant variables that are correlated with included right-hand side variables.

2.

3.

4.
Omitted Variables

Suppose the model we *use* is:

while the “true” model that generates the data is:

Note that the first equation is not “incorrect” in any sense; the “true” model simply implies that the error $\eta_i$ can be written as:

The relevant question is: will OLS estimation of the first equation give us unbiased and consistent estimates of $\beta$?
Omitted Variables

As for unbiasedness, note:

\[
E(\hat{\theta}) = E \left[ (X'X)^{-1}X'y \right] \\
= E \left[ (X'X)^{-1}X'(X\beta + Z\gamma + \eta) \right] \\
= \beta + E \left[ (X'X)^{-1}X'Z \right] \gamma + E \left[ (X'X)^{-1}X'\eta \right]
\]

What is reasonable to assume here?
Can we make a similar assumption regarding the term:

$$E \left[ (X'X)^{-1} X'Z \right] \gamma?$$

Not unless:

1. If $Z$ and $X$ are correlated, this assumption will be violated. As a result, OLS estimation in this case will be biased. The OLS estimator will also be inconsistent in general.
Consider the following *simple regression model* with a single, mismeasured $x$ variable:

\[
\begin{align*}
y_i &= \tilde{\alpha} + x_i \tilde{\beta} + \eta_i \\
y_i &= \alpha + x_i^* \beta + u_i \\
x_i &= x_i^* + \omega_i
\end{align*}
\]

The first equation is the model we *use*. The second is the *true* regression model, while the third describes the *measurement error process*. 
In what follows, we make standard “textbook” assumptions.

Namely that the error of measurement, $\omega$ is independent of $x^*$ and that the “true” error $u$ is independent of both $x^*$ and $\omega$.

We focus on the slope coefficient. For the simple regression model this is given by:

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

We seek to determine if this rule \textit{consistently} estimates $\beta$. 
From the true model (second equation), we know:

\[ y_i - \bar{y} = (x_i^* - \bar{x}^*)\beta + u_i - \bar{u}. \]

Subbing this into the expression for \( \hat{\beta} \) gives:

\[
\hat{\beta} = \beta \frac{\sum_i (x_i - \bar{x})(x_i^* - \bar{x}^*)}{\sum_i (x_i - \bar{x})^2} + \frac{\sum_i (x_i - \bar{x})(u_i - \bar{u})}{\sum_i (x_i - \bar{x})^2}.
\]

Dividing the numerator and denominator of each expression above by \( n \), and taking \( \lim_{n \to \infty} \), we have:
Consider the term $\text{Cov}(x, u)$. We know

Thus, $\text{plim}(\hat{\beta}) = \beta \frac{\text{Cov}(x, x^*)}{\text{Var}(x)}$.

Expanding this out, we can write:
So, what have we learned from this simple exercise?

1. $\hat{\beta}$ is *inconsistent* in general, unless $\sigma^2_\omega = 0$ (which, again, assumes away the problem).

2. If $\sigma^2_\omega$ is small relative to $\sigma^2_{\chi^*}$, then the degree of inconsistency is also small.

3. This last point potentially enables the researcher to interpret the estimate as a type of “lower bound.”
For the more general case, where we use the model:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* + \cdots + \beta_k x_k^* + \eta_i \]

(and \( x_1 \) is the only variable measured with error), we can show:

\[
\text{plim}(\hat{\beta}_1) = \beta_1 \left( \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\omega^2} \right)
\]

where \( \sigma_v^2 \) is the population variance in the regression equation:

\[ x_1^* = \alpha_1 + \alpha_2 x_2^* + \cdots + \alpha_k x_k^* + \nu. \]

Therefore, we continue to have attenuation bias for the error-ridden variable.

The direction of the bias for the remaining parameters is unclear generally.
Simultaneity

A *simultaneous equations model* is one that describes the joint determination of two (or more) variables. Moreover, each variable plays a structural role in the determination of the other variable. For example,

Can you think of any examples?

1. 
2. 
3. 

Justin L. Tobias (Purdue)
Simultaneity

If the SEM is the “true” model that generates the data, should we estimate each equation separately to estimate the \( \alpha \) and \( \beta \) parameters? We continue to assume that the \( X \) variables are exogenous (determined outside the model) and thus:

\[
E(\epsilon_1|X_1, X_2) = E(\epsilon_2|X_1, X_2) = 0,
\]

while \( y_1 \) and \( y_2 \) are endogenous (determined within the model).

To address the estimation issue, it turns out that we should express each endogenous variable in reduced form. (That is, as a function of exogenous variables only).
Simultaneity

To this end, note:

so that

\[ y_2 = \left( \frac{1}{1 - \alpha_1 \alpha_2} \right) \left[ X_1 \alpha_2 \beta_1 + X_2 \beta_2 + \alpha_2 \epsilon_1 + \epsilon_2 \right]. \]

Now, consider running least squares, using just the first equation. In order for unbiasedness and consistency to hold, we must have

\[ E(\epsilon_1 | y_2, X_1) = 0. \]

Will this hold in the SEM?
Let us again consider the covariance:

where we have assumed that $\epsilon_1$ and $\epsilon_2$ are independent. This covariance is nonzero unless $\sigma_{\epsilon_1}^2 = 0$ or $\alpha_2 = 0$ (which, again, assumes away the problem).
**Endogeneity**

*Endogeneity* is similar to the case of omitted variables or a “restricted” form of simultaneity. To illustrate, we consider a model of the form:

\[ y_2 \]

This differs from omitted variables since \( y_2 \) is treated as *endogenous*. This differs from simultaneity since \( y_2 \) is already in reduced form and has no structural dependence on \( y_1 \) (e.g., wages and education).

Moreover, what is of primary importance here is that \( \epsilon_1 \) and \( \epsilon_2 \) are *correlated* - unobserved factors affecting the “production” of \( y_2 \) also affect the (conditional) production of \( y_1 \).
Endogeneity

To illustrate, note for the case of endogeneity:

It might be helpful to illustrate with the case of a normal sampling model. That is, suppose

$$
\begin{bmatrix}
\epsilon_{1i} \\
\epsilon_{2i}
\end{bmatrix} \mid x_i \sim \mathcal{N} \left( \begin{bmatrix} 0 \\
0 
\end{bmatrix} , \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\
\sigma_{12} & \sigma_2^2 
\end{bmatrix} \right).
$$

In this case, the Jacobian of the transformation from $\epsilon$ to $y$ is unity given the **triangularity** of the model. Using properties of the Normal distribution, we can then obtain:

$$
y_{1i} \mid y_{2i}, x_i \sim \mathcal{N} \left( \alpha_0 + \alpha_1 y_2 + x_1 \beta_1 + \frac{\sigma_{12}}{\sigma_2^2} [y_2 - \delta_0 - x_2 \delta_1], \sigma_1^2[1 - \rho_{12}^2] \right),
$$

where $\rho_{12}$ denotes the correlation between $\epsilon_1$ and $\epsilon_2$. 
Grouping terms together, we obtain:

\[
y_{1i} | y_{2i}, x_i \sim \mathcal{N}\left(\left[\alpha_0 - \delta_0 \frac{\sigma_{12}}{\sigma_2^2}\right] + y_2 \left[\alpha_1 + \frac{\sigma_{12}}{\sigma_2^2}\right] - x_2 \delta_1 \frac{\sigma_{12}}{\sigma_2^2} + x_1 \beta_1, \sigma_1^2[1 - \rho_{12}^2]\right)\).
\]

The above suggests what a regression of \(y_1\) on \(y_2, x_1\) and \(x_2\) will recover (consistently estimate). Specifically, we find that coefficient on \(y_2\) will converge to:

\[
\alpha_1 + \frac{\sigma_{12}}{\sigma_2^2}
\]

and the existence of an \(x_2 \notin x_1\) appears to be critical for identification purposes.