Endogeneity:

Let in a linear regression ~~model~~ setup, we have

Y: outcome/dependent variable

$X_1, \ldots, X_p$: independent or explanatory variables or regressors; $p \geq 1$

In general, our objective is to fit a ~~best~~ linear regression model of Y on $X_1 \ldots X_p$, i.e.,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i$$

by minimizing the ~~error~~ ~~sq~~ sum of squares error, $\sum_{i=1}^{n} e_i^2$, with ~~resp~~ respect to $\beta_0, \beta_1, \ldots \beta_p$.

What is endogeneity?:

In econometrics, endogeneity refers to the situations in which an explanatory variable (X_i) is correlated with the error term (e_i).

As a consequence, OLS estimators of the regression coefficients are no longer ~~consistent~~ unbiased.

However, if the correlation is not contemporaneous, OLS estimators may still be consistent. But usually it is inconsistent

~~Com~~ Some common sources of endogeneity:

1. Omitted variable
2. Measurement error
3. Simultaneity

1 Omitted variable:

Assume that the true model is

$$y_i = \alpha + \beta x_i + \gamma z_i + e_i$$

but in real life z_i is omitted from the regression model. Then the working model is

$$y_i = \alpha + \beta x_i + u_i$$

where $u_i = \gamma z_i + e_i$

If $Cov(x, z) \neq 0$ and

z seperately affects y (i.e. $\gamma \neq 0$)

then $Cov(x, u) \neq 0$

Here x variable is called <u>endogeneous</u> variable.

Measurement error:

Suppose instead of observing the actual x_i^*, we observe $x_i = x_i^* + v_i$, where v_i is the measurement error. In this case, a model is given by

$$y_i = \alpha + \beta x_i^* + e_i$$

$$= \alpha + \beta(x_i - v_i) + e_i$$

$$= \alpha + \beta x_i + (e_i - \beta v_i)$$

$$= \alpha + \beta x_i + u_i, \quad \text{where } u_i = e_i - \beta v_i$$

Since both x_i & u_i depend on v_i, hence $cor(x_i, u_i) \neq 0$, so the OLS estimation of β will be biased downward.

Qn: Are they still consistent ?? **Ans:** Yes! If the correlation is not contemporaneous, then $\hat{\beta}_{OLS}$ may still be consistent.

Note: ~~Howe~~ ~~Note~~ Note that measurement error in the dependent variable y_i, does not cause endogeneity, though it ~~may~~ does increase the variance of the error term. But usually it is inconsistent

3 Simultaneity:

Suppose

$$y_i = \beta_1 x_i + \gamma_1 z_i + u_i$$

$$z_i = \beta_2 x_i + \gamma_2 y_i + v_i$$

$$\Rightarrow \quad z_i = \frac{\beta_2 + \gamma_2 \beta_1}{1 - \gamma_1 \gamma_2} + \frac{1}{1 - \gamma_1 \gamma_2} v_i + \frac{\gamma_2}{1 - \gamma_1 \gamma_2} u_i$$

Assuming $cor(x_i, u_i) = 0$ & $cor(v_i, u_i) = 0$

$cor(z_i, u_i) \neq 0$.

Instrumental variable

Date

Omitted variable case:

Suppose we have a simple linear regression with one regressor, i.e.,

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where $Cov(x_i, \varepsilon_i) \neq 0$, $E(\varepsilon_i) = 0$, $var(\varepsilon_i) = \sigma^2$

Then ordinary least squares estimate

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$E(\hat{\beta}_{OLS}) = E\left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

[Note that for data, $x_1 \ldots x_n$ are not random]

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\alpha + \beta x_i + \varepsilon_i)}{\sum (x_i - \bar{x})^2}$$

$$= \beta + \frac{\sum (x_i - \bar{x}) \varepsilon_i}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{\beta}_{OLS} \mid x_1 \ldots x_n) = \beta + \underbrace{\frac{\sum_{i=1}^{n} (x_i - \bar{x}) E(\varepsilon_i \mid x_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}_{\neq 0}$$

How to show that $\hat{\beta}_{OLS}$ is still consistent?

Hence $\hat{\beta}_{OLS}$ is a biased estimator of β.

Reg. model with one regressor:

$$Y_i = \alpha + \beta X_i + u_i$$

where $Cov(X_i, u_i) \neq 0$

$\Rightarrow$ X_i is endogenous variable in predicting Y_i

Two conditions are required for a valid ~~ist~~ instrumental variable (Z):

(i) $Cov(Z_i, X_i) \neq 0$: called <u>instrument relevance</u>

(ii) $Cov(Z_i, u_i) = 0$: <u>instrument exogeneity</u>

Suppose for now, we have a valid instrumental variable Z, then how can we use Z to estimate β?

<u>Method - called ~~Two~~ 2-stage least squares (2SLS) method</u>:

<u>Stage 1</u>: Regress X on Z, i.e.,

$$X_i = \delta_0 + \delta_1 Z_i + v_i$$

Note that $Cov(Z_i, u_i) = 0$

$$\Rightarrow Cov(\delta_0 + \delta_1 Z_i, u_i) = 0$$

However, here we don't know δ_0 & δ_1 and we have to estimate them ~~first~~ ~~by~~ ~~minimi~~ using OLS method. Let $\hat{\delta}_0$ & $\hat{\delta}_1$ are the OLS estimates of δ_0 & δ_1

Then compute

$$\hat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i$$

<u>Stage 2:</u> Replace X_i by $\hat{X}_i$ in equation (1), i.e.,

$$Y_i = \alpha + \beta \hat{X}_i + u_i$$

Note that here $\hat{X}_i$ is exogenous because $\text{Cov}(\hat{X}_i, u_i)$

$$= \text{Cov}(\hat{\delta}_0 + \hat{\delta}_1 Z_i, u_i)$$

$$= 0.$$

Thus β can be estimated using OLS method and is given by

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n} (\hat{X}_i - \bar{\hat{X}}) Y_i}{\sum (\hat{X}_i - \bar{\hat{X}})^2}$$

→ this resulting estimator is called the 2-stage least squares estimator.

<u>Result:</u> $\hat{\beta}_{2SLS}$ is a consistent estimator of β

i.e.

$$\hat{\beta}_{2SLS} \xrightarrow{P} \beta.$$

Interpretation of the estimator $\hat{\beta}_{2SLS}$:

Note that our linear reg. model is given by

$$Y_i = \alpha + \beta X_i + u_i$$

Thus, for a valid instrumental variable Z_i, we have

$$\begin{aligned} Cov(Y_i, Z_i) &= Cov(\alpha + \beta X_i + u_i, Z_i) \\ &= Cov(\alpha, Z_i) + \beta\, Cov(X_i, Z_i) \\ &\quad + Cov(u_i, Z_i) \\ &= 0 + \beta\, Cov(X_i, Z_i) + 0 \end{aligned}$$

$$\Rightarrow \beta = \frac{Cov(Y_i, Z_i)}{Cov(X_i, Z_i)}$$

Now if we replace the population covariance by sample covariance, then we have

$$\hat{\beta}_{2SLS} = \frac{S_{YZ}}{S_{XZ}} = \frac{\sum (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum (X_i - \bar{X})(Z_i - \bar{Z})}$$

Note that sample covariances are consistent estimators of population covariances, ~~and hence~~ i.e.,

$$s_{YZ} \xrightarrow{p} cov(Y, Z) \quad \& \quad s_{XZ} \xrightarrow{p} cov(X, Z).$$

Thus,

$$\hat{\beta}_{2SLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{Cov(Y, Z)}{Cov(X, Z)} = \beta_1$$

* prepare a note for endogeneity in the more than one endogenous variable.