$Y = \beta_0 + \beta_1 X + u$ { Endogeneity: there exists some omitted variable(s) W
i) in the error term u (so W is a determinant of Y)
ii) W is correlated with X
$\therefore$ OVB in estimation of β_1

- X is endogenous : $E[u|X] \neq 0$

$\Rightarrow$ Violation of OLS assumption

$\Rightarrow$ OVB for an endogenous variable

Variation in X

Endogenous	Exogenous
$E[u|X_N] \neq 0$	$E[u|X_X] = 0$

$\Rightarrow$ Z valid instrument

① Relevance : $Cov(X, Z) \neq 0$

$Cor(X, Z) \neq 0$

② Exogenous : $E[u|Z] = 0$

① Two-stage approach

i) Regress $X = \pi_0 + \pi_1 Z + e$

$\Rightarrow \hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z$

$= f(Z)$, a linear function of Z, exogenous

$\therefore \hat{X}$ must be exogenous also

ii) Regress $Y = \beta_0 + \beta_1 \hat{X} + e$

$= \beta_0 + \beta_1 (\hat{\pi}_0 + \hat{\pi}_1 Z) + e$

$= (\beta_0 + \beta_1 \hat{\pi}_0) + \beta_1 (\hat{\pi}_1 Z) + e$

$(\beta_0 + \beta_1 \hat{\pi}_0)$: constant; $(\hat{\pi}_1 Z)$: exogenous

$\therefore \beta_1$ estimated without OVB

② Control Variable Approach

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

- X is endogenous : $\mathbb{E}[u|X] \neq 0$
- In particular, there is a (set of) variable(s) W s.t.
  - W is contained in u : it is a determinant of Y
  - $Cov(X,W) \neq 0 \iff Corr(X,W) \neq 0$ $\therefore \hat{\beta}_{1,OLS}$ estimated with OVB

Goal: Isolate the effect of X on Y independent of W (β_1)

Supposing we have data on W

The OLS assumption for valid inference in multiple regression:

$$\mathbb{E}[u|X,W] = \mathbb{E}[u|W]$$

"conditional mean assumption" : holding W constant, X is uncorrelated with the error and β_1 can be estimated without OVB

Variation in X

all variation in X NOT explained by Z must be a comprehensive source of endogenous variation

$$\mathbb{E}[u|\hat{e}] = \mathbb{E}[u|X - \hat{X}]$$
$$= \mathbb{E}[u|X - f(Z)]$$
$$= \mathbb{E}[u|X]$$
$$\mathbb{E}[u|\hat{e}, X] = \mathbb{E}[u|X]$$

conditional mean assumption is satisfied

$\Rightarrow$ Estimation of β_1 is unbiased with inclusion of W as a control variable

① Regress $X = \pi_0 + \pi_1 Z + e$

$\Rightarrow \hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z$

$= f(Z)$, must be exogenous

$\Rightarrow (X - \hat{X})$ will capture all endogenous variation X ($:= \hat{u}$)

$\hat{u}$ residual $\Rightarrow$ candidate for W

② Regress $Y = \beta_0 + \beta_1 X + \beta_2 W + u$ ($W \leftarrow \hat{u}$)

$= \beta_0 + \beta_1 X + \beta_2 (X - \hat{X}) + u$

In fact, because we're using the same information (X,Z) to identify β_1, the estimates will be exactly the same as the first approach

$\Rightarrow \hat{\beta}_{1,2S} = \hat{\beta}_{1,control} \sim$ unbiased (β_1)