Econ 361: Advanced Econometrics

“Exogeneity” (and “Causality”)

Linear Regression Equation

$$\begin{aligned} Y_i &= X_{1i}'\beta_1 + X_{2i}'\beta_2 + \epsilon_i \\ Y &= \underbrace{X_1\beta_1 + X_2\beta_2}_{\text{observable}} \quad + \quad \underbrace{\epsilon}_{\text{unobservable}} \end{aligned}$$

Recall: $\epsilon \equiv Y - X_1\beta_1 - X_2\beta_2$

Suppose

$$E[\epsilon|X_1] \;=\; E[\epsilon] \quad \text{but} \quad E[\epsilon|X_2] \;=\; g(X_2) \;\neq\; E[\epsilon]$$

X_1 is linearly informative about the observable variation **only**. But X_2 is linearly informative about **both** the observable and unobservable variation. As such, there are complications estimating β_2 as the predictive power X_2 is observed as having about Y is for both variations. Rather than estimating β_2, estimating some approximation of $\beta_2 + \frac{\partial g(X_2)}{\partial X_2}$. Difficult to isolate channels

Exogeneity and "Causality"

$$Y \quad = \quad \underbrace{X_1\beta_1 + X_2\beta_2}_{\text{observable}} \quad + \quad \underbrace{\epsilon}_{\text{unobservable}}$$

- We want the variation in (X_1, X_2) to be informative about the variation in Y but **not** for the variation in Y to be informative about the variation in (X_1, X_2)
- This would be the case if, for example, the values of (X_1, X_2) were first chosen and then those fixed values were used to determine the value of Y, i.e. (X_1, X_2) helped "cause" the Y ... as in a **controlled** experiment
- In which case, (β_1, β_2) could be considered the "casual" effect of (X_1, X_2), respectively, on Y – more specifically, the causal effect of a **marginal** change in (X_1, X_2), respectively, on Y on **average**

Endogeneity Problems: Omitted Variables

$$
\begin{aligned}
Y &= X_1\beta_1 + \underbrace{X_2\beta_2}_{X_3\beta_3+X_4\beta_4} + \epsilon \\
Y &= \underbrace{X_1\beta_1 + X_3\beta_3}_{\text{now observable}} + \underbrace{(X_4\beta_4 + \epsilon)}_{\text{now unobservable}}
\end{aligned}
$$

- Let $X_2 = (X_3 X_4)$ where we observe X_3 but not X_4 ... X_4 is omitted
- Further, $E[X_4|X_1] = E[X_4]$ but $E[X_4|X_3] = h(X_3) \neq E[X_4]$
- X_1 may still be exogenous but X_3 is not as X_3 is informative about the unobservable X_4

Endogeneity Problems: Measurement Errors

$$
\begin{aligned}
Y &= X_1\beta_1 + X_2\beta_2 + \epsilon \\
Y &= \underbrace{X_1\beta_1 + \tilde{X}_2\beta_2}_{\text{now observable}} + \underbrace{(-\nu\beta_2 + \epsilon)}_{\text{now unobservable}}
\end{aligned}
$$

- Let $\tilde{X}_2 = X_2 + \nu$
- X_1 may still be exogenous but $\tilde{X}_2$ is not as $\tilde{X}_2$ is informative about the unobservable ν

Endogeneity Problems: Simultaneity

$$Y = \underbrace{X_1\beta_1 + X_2\beta_2}_{\text{observable}} + \underbrace{\epsilon}_{\text{unobservable}}$$

- If (Y, X_2) are **simultaneously** determined, then $E[X_2|Y] \neq E[X_2]$ in general
- The actual realization of Y impacts the actual realization of X_2 and, therefore X_2 is informative about even ϵ
- X_1 may still be exogenous but X_2 is not as X_2 is informative about the unobservable ϵ

Exogeneity

- Regressors X are considered “exogenous” if it is **mean independent** of the regression error: $E[\epsilon|X] = E[\epsilon]$
- Note that the above implies that $E[X'\epsilon] = 0$ when $E[\epsilon] = \vec{0}$
  $E[X'\epsilon] = E_X[\, E[X'\epsilon|X]\,] = E_X[\, X'E[\epsilon|X]\,] = E_X[\, X'E[\epsilon]\,] = \vec{0}$
  $E[\epsilon] = \vec{0}$ without much loss of generality when constant included as regressor
- So, a regressor is considered “exogenous” if $E[X'\epsilon] = 0$
- Sample analog to the above exogeneity population moment condition is $X'e = 0$ where e is the regression residual

Exogeneity: OLS

- Linearity Condition: $E[Y|X] = X\beta$

  implies $E[\epsilon|X] = \vec{0}$ and therefore $E[X'\epsilon] = \vec{0}$

- Sample analog: $X'e = \vec{0}$

  $$X'e = X'(Y - Xb_{ols}) = X'Y - X'Xb_{ols} = 0$$

  $$b_{ols} = (X'X)^{-1}X'Y$$

- Sample analog is the FOC from the OLS minimization problem

- Violation of the Linearity Condition implies that the population moment condition upon which the OLS estimator is built is wrong, hence an improper moment-based estimator of β

Exogeneity: GLS

- Linearity Condition: $E[\tilde{Y}|\tilde{X}] = \tilde{X}\beta$

  implies $E[\tilde{\epsilon}|\tilde{X}] = \vec{0}$ and therefore $E[\tilde{X}'\tilde{\epsilon}] = \vec{0}$

  Recall $(\tilde{Y}, \tilde{X})$ is the suitably transformed data

- Sample analog: $\tilde{X}'\tilde{e} = \vec{0}$

  $$\tilde{X}'\tilde{e} = \tilde{X}'(\tilde{Y} - \tilde{X}b_{gls}) = \tilde{X}'\tilde{Y} - \tilde{X}'\tilde{X}b_{gls} = 0$$

  $$b_{gls} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$$

- Sample analog is the FOC from the GLS minimization problem

- Violation of the Linearity Condition implies that the population moment condition upon which the GLS estimator is built is wrong, hence an improper moment-based estimator of β

Exogeneity: 2SLS

- Linearity Condition: $E[Y|\hat{X}] = \hat{X}\beta$

  implies $E[\epsilon|\hat{X}] = \vec{0}$ and therefore $E[\hat{X}'\epsilon] = \vec{0}$

  Recall $\hat{X}$ is the properly instrumented transformation of X

- Sample analog: $\hat{X}'e = \vec{0}$

  $\hat{X}'e = \hat{X}'(Y - \hat{X}b_{2SLS}) = \hat{X}'Y - \hat{X}'\hat{X}b_{2sls} = 0$

  $b_{2sls} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$

- Sample analog is the FOC from the 2SLS minimization problem

- Violation of the Linearity Condition implies that the population moment condition upon which the 2SLS estimator is built is wrong, hence an improper moment-based estimator of β

Exogeneity: IV

- Linearity Condition: $E[Y|Z] = E[X|Z]\beta$

  implies $E[\epsilon|Z] = \vec{0}$ and therefore $E[Z'\epsilon] = \vec{0}$

  Recall Z are proper instruments for X.

  Note that $\hat{X}$ can serve as Z too ... and even X if exogenous !!!

- Sample analog: $Z'e = \vec{0}$

  $$Z'e = Z'(Y - Xb_{IV}) = Z'Y - Z'Xb_{iv} = 0$$

  $$b_{iv} = (Z'X)^{-1}Z'Y$$

- (OLS, GLS, 2SLS) can be thought as versions of IV

- Violation of the Linearity Condition implies that the population moment condition upon which the IV estimator is built is wrong, hence an improper moment-based estimator of β