# ECON 480-3
# LECTURE 5: ENDOGENEITY II

**Ivan A. Canay**
**Northwestern University**

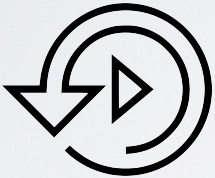
# Past & Future

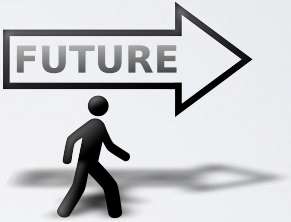
## Last Class

- Instrumental Variables
- The IV Estimator
- The 2SLS Estimator
- Properties of 2SLS
- Estimating $\mathbb{V}$

## TODAY

- Efficiency of 2SLS
- Weak IV
- LATE





# Efficiency of Two-Stage Least Squares

- Let $(Y, X, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$ and $X$ takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of $X$ is constant and equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U\,.$$

- We assume (1) $E[ZU] = 0$, (2) $E[ZX'] < \infty$, (3) $E[ZZ'] < \infty$, and (4) there is no perfect collinearity in $Z$, and (5) the rank of $E[ZX']$ is $k+1$

- Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution $P$.

- The TSLS estimator identifies $\beta$ by means of the projection matrix $\Pi = E[ZZ']^{-1}E[ZX']$. Is this a good choice?

# Efficiency of Two-Stage Least Squares

- We could solve for $\beta$ using any $(\ell+1)\times(k+1)$ dimensional matrix $\Gamma$ such that $E[\Gamma' Z X']$ has rank $k+1$.
- **Interpretation**: we could use some other linear combination of instruments, $\Gamma' Z$ instead of $\Pi' Z$.
- For any such matrix,

$$\beta = E[\Gamma' Z X']^{-1} E[\Gamma' Z Y] \,,$$

and we could have estimated $\beta$ using

$$\tilde{\beta}_n = \left( \frac{1}{n} \sum_{1 \leqslant i \leqslant n} \Gamma' Z_i X_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \leqslant i \leqslant n} \Gamma' Z_i Y_i \right) \,.$$

- Could use a consistent estimate $\hat{\Gamma}_n$ of $\Gamma$ instead.
- By arguing as before, it is possible to show under our assumptions that $\tilde{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$.

# Efficiency of Two-Stage Least Squares

- Suppose $\text{Var}[ZU] = E[ZZ'U^2] < \infty$. Then

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N(0, \tilde{\mathbb{V}}) \quad \text{as} \quad n \to \infty \quad \text{with} \quad \tilde{\mathbb{V}} = E[\Gamma'ZX']^{-1}\Gamma'\text{Var}[ZU]\Gamma E[\Gamma'ZX']^{-1\prime} .$$

- Under some assumptions: the "best" choice of $\Gamma$ is given by $\Pi$, i.e., $\tilde{\mathbb{V}} \geqslant \mathbb{V}$.

- **Show this**: assume that $E[U|Z] = 0$ and $\text{Var}[U|Z] = \sigma^2$. In addition, define $W^* = \Pi'Z$ and $W = \Gamma'Z$. To see that $\tilde{\mathbb{V}} \geqslant \mathbb{V}$, first re-write $\tilde{\mathbb{V}}$:

# Efficiency of Two-Stage Least Squares

$$\tilde{\mathbb{V}} = \sigma^2 E[WW^{*\prime}]^{-1} E[WW'] E[WW^{*\prime}]^{-1\prime} \quad \text{and similarly} \quad \mathbb{V} = \sigma^2 E[W^* W^{*\prime}]^{-1} .$$

**WTS:** $\mathbb{V} \leqslant \tilde{\mathbb{V}}$ or $\mathbb{V}^{-1} \geqslant \tilde{\mathbb{V}}^{-1}$ or $\mathbb{V}^{-1} - \tilde{\mathbb{V}}^{-1} \geqslant 0$

# QUESTIONS?

# Weak IV

- **Normal approximation**: can be poor in finite samples when the rank of $E[ZX']$ is "close" to being $< k+1$.
- **Consequence**: hypothesis tests and confidence regions based off of this approximation can behave poorly in finite samples as well.
- **Example**: To gain some insight into this phenomenon in a more elementary way, suppose

$$\begin{aligned} Y_i &= X_i\beta + U_i \\ X_i &= Z_i\pi + V_i\,, \end{aligned}$$

  where $Z_1, \ldots, Z_n$ are non-random, $(U_1, V_1), \ldots, (U_n, V_n)$ is a sequence of i.i.d. $N(0, \Sigma)$ rvs.
- Suppose $\pi \neq 0$. Consider the estimator given by

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n Z_i Y_i}{\frac{1}{n}\sum_{i=1}^n Z_i X_i}\,.$$

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i U_i}{\left(\frac{1}{n} \sum_{i=1}^{n} Z_i^2\right) \pi + \frac{1}{n} \sum_{i=1}^{n} Z_i V_i} \equiv \frac{W_1}{W_2} .$$

**The finite-sample, joint distribution of the numerator and denominator is**

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ \bar{Z_n^2}\pi \end{pmatrix}, \begin{pmatrix} \bar{Z_n^2}\sigma_U^2 & \frac{1}{\sqrt{n}}\bar{Z_n^2}\sigma_{U,V} \\ \frac{1}{\sqrt{n}}\bar{Z_n^2}\sigma_{U,V} & \frac{1}{n}\bar{Z_n^2}\sigma_V^2 \end{pmatrix} \right) ,$$

where

$$\bar{Z_n^2} = \frac{1}{n} \sum_{i=1}^{n} Z_i^2 .$$

This joint distribution completely determines the finite-sample distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$.

In particular, it is the **ratio of two (correlated) normal random variables**.

# Weak IV: the problem

- If $\bar{Z_n^2} \to \bar{Z^2} > 0$ as $n \to \infty$, then

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i U_i}{\bar{Z_n^2}\pi + \frac{1}{n} \sum_{i=1}^n Z_i V_i} \xrightarrow{d} N\left(0, \frac{\sigma_U^2}{\pi^2 \bar{Z^2}}\right) .$$

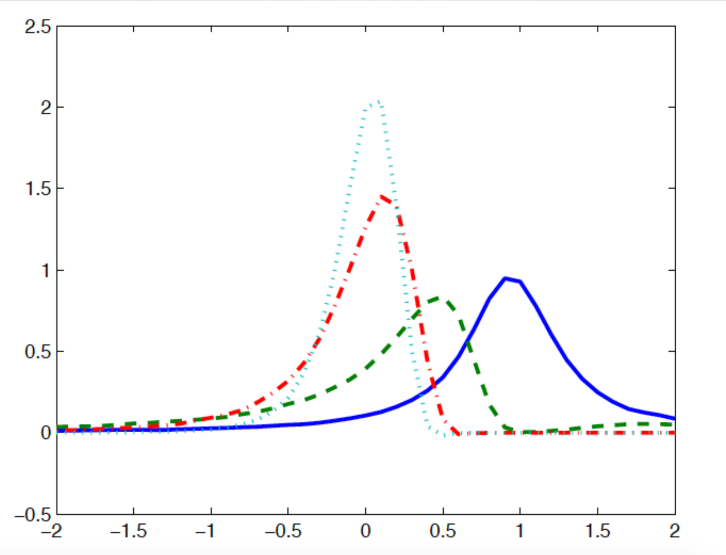
- This approximation effectively treats the denominator like a constant equal to its mean

- **Good Approx.**: when the mean is "large" relative to the sd, i.e.,

$$\bar{Z_n^2}\pi \gg \frac{1}{\sqrt{n}}\sqrt{\bar{Z_n^2}}\sigma_V \iff \pi \gg \frac{1}{\sqrt{n}} \frac{\sigma_V}{\sqrt{\bar{Z_n^2}}} .$$

- **Poor Approx.**: when $\pi$ is "small", the approximation may be quite poor in finite-samples. Note in particular that $\pi \neq 0$ is not sufficient for the approximation to be good in finite-samples.

# Weak IV: finite sample distribution



# Weak IV: a way around it

- Consider $H_0 : \beta = c$ versus $H_1 : \beta \neq c$ at level $\alpha$.
- **Under** $H_0$: one can compute $U_i = Y_i - X_i'\beta$ and $Z_i U_i = Z_i(Y_i - X_i'\beta)$.
- Since $E[ZU] = 0$, we can simply test whether this is true using $Z_1 U_1, \ldots, Z_n U_n$.
- **Formaly**: Assume Var$[ZU]$ is invertible and define $W_i(c) = Z_i(Y_i - X_i'c)$. When $\beta = c$, we have that

$$\sqrt{n}\bar{W}_n(c) = \frac{1}{\sqrt{n}} \sum_{1 \leqslant i \leqslant n} W_i(c) \xrightarrow{d} N(0, \Sigma(c)) \ ,$$

where $\Sigma(c) = \text{Var}[W(c)]$. If we define

$$\hat{\Sigma}_n(c) = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} (W_i(c) - \bar{W}_n(c))(W_i(c) - \bar{W}_n(c))'$$

and use arguments given earlier, we see that under $H_0$

$$T_n = n\bar{W}_n'(c)\hat{\Sigma}_n^{-1}(c)\bar{W}_n(c) \xrightarrow{d} \chi^2_{\ell+1} \ .$$

We can test $H_0$ by comparing $T_n$ with $c_{\ell+1,1-\alpha}$, the $1-\alpha$ quantile of the $\chi^2_{\ell+1}$ distribution.

# Weak IV: Discussion

- **Anderson-Rubin**: a closely related variant of this idea leads to the *Anderson-Rubin* test, in which one tests whether all of the coefficients in a regression of $Y_i - X_i'c$ on $Z_i$ are zero.
- **Anderson-Rubin**: has good power properties when the model is exactly identified, but may be less desirable when the model is over-identified.
- Other methods for the case in which the model is over-identified and/or one is only interested in some feature of $\beta$ (e.g., one of the slope parameters) have been proposed and are the subject of current research as well.
- The literature on weak IV is large and it is mostly based on test inversion.
- **Two Step Approach Alternative**: Step 1: investigate whether the rank of $E[ZX']$ is "close" to being $< k+1$ or not. Step 2: use these "more complicated" methods if they failed to reject this null hypothesis. This two-step method will also behave poorly finite-samples and **should not be used**.

# QUESTIONS?

# Interpretation under Heterogeneity

- Despite possible inefficiencies, TSLS remains popular.
- **Possible reason**: interpretation in the presence of heterogeneous effects of $X$ on $Y$.
- Recall that in the model
$$Y = X'\beta + U \,,$$
the effect of a change in $X$ (say, from $X = x$ to $X = x'$) is the *same* for everybody.
- What if the effect of a change in $X$ on $Y$ is different for different people.
- **To capture this**: allow for $\beta$ to be random. When $\beta$ is random, we may absorb $U$ into the intercept and simply write
$$Y = X'\beta \,.$$
- **Notation**: with a random sample where variables are indexed by $i$, we would write $Y_i = X_i'\beta_i$, which makes it explicit that every individual has a unique effect $\beta_i$.

# Notation

- Assume $k = 1$ and write $D$ in place of $X_1$, which is assumed to take values in $\{0, 1\}$. Then,

$$Y = \beta_0 + \beta_1 D \, .$$

- We interpret $\beta_0$ as $Y(0)$ and $\beta_1$ as $Y(1) - Y(0)$, where $Y(1)$ and $Y(0)$ are *potential* or *counterfactual outcomes*. Using this notation, we may rewrite the equation as

$$Y = DY(1) + (1 - D)Y(0) \, .$$

- $Y(0)$ value of the outcome that would have been observed if (possibly counter-to-fact) $D$ were $0$;
  $Y(1)$ value of the outcome that would have been observed if (possibly counter-to-fact) $D$ were $1$.

- The variable $D$ is typically called the *treatment* and $Y(1) - Y(0)$ is called the *treatment effect*. The quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*.

# Random Assignment

If $D$ were **randomly assigned** (e.g., by the flip of a coin), then

$$(Y(0), Y(1)) \perp\!\!\!\perp D \ .$$

In this case, under mild assumptions, the slope coefficient from OLS regression of $Y$ on a constant and $D$ yields a consistent estimate of the average treatment effect.

# Selection

- **Selection**: In general, we expect $D$ to depend on $(Y(1), Y(0))$
- **OLS does not** yield a consistent estimate of the average treatment effect.
- To proceed further, we therefore assume, as usual, that there is an instrument $Z$. Let $Z \in \{0, 1\}$.
- Consider the slope coefficient from TSLS/IV regression of $Y$ on $D$ with $Z$ as an instrument,

$$\frac{\text{Cov}[Y, Z]}{\text{Cov}[D, Z]} = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]},$$

  where the equality follows by multiplying and dividing by Var$[Z]$ and using earlier results.
- **Goal**: to express this quantity in terms of the treatment effect $Y(1) - Y(0)$ somehow.

# Potential Treatments

- Towards our goal, it is useful to also introduce the following equation for $D$:

$$\begin{aligned} D &= ZD(1) + (1-Z)D(0) \\ &= D(0) + (D(1) - D(0))Z \\ &= \pi_0 + \pi_1 Z \,, \end{aligned}$$

  where $\pi_0 = D(0)$, $\pi_1 = D(1) - D(0)$, and $D(1)$ and $D(0)$ are *potential* or *counterfactual treatments*

- We impose the following versions of instrument **exogeneity** and instrument **relevance**, respectively:

$$(Y(1), Y(0), D(1), D(0)) \perp\!\!\!\perp Z$$

  and

$$P\{D(1) \neq D(0)\} = P\{\pi_1 \neq 0\} > 0 \,.$$

- We further assume the following **monotonicity** condition:

$$P\{D(1) \geqslant D(0)\} = P\{\pi_1 \geqslant 0\} = 1 \,.$$

# The TSLS Estimand

## Definition (Local Average Treatment Effect)

The TSLS/IV estimand equals

$$\frac{\text{Cov}[Y,Z]}{\text{Cov}[D,Z]} = E[\underbrace{Y(1)-Y(0)}_{\text{TE}} \,|\, \underbrace{D(1)>D(0)}_{\text{local}}] \equiv \text{LATE}$$

This is called the local average treatment effects.

Average treatment effect among the subpopulation of people for whom a change in the value of the instrument switched them from being non-treated to treated: the so-called **compliers**.

# Role of Monotonicity

- **Monotonicity**: while the instrument may have no effect on some people, all those who are affected are affected in the same way. Without monotonicity, we would have

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] = {} & E[Y(1) - Y(0)|D(1) > D(0)]P\{D(1) > D(0)\} \\ & - E[Y(1) - Y(0)|D(1) < D(0)]P\{D(1) < D(0)\} \,. \end{aligned}$$

- Treatment effects may be positive for everyone (i.e., $Y(1) - Y(0) > 0$) yet the reduced form is zero because effects on **compliers** are canceled out by effects on **defiers**, i.e., those individuals for which the instrument pushes them out of treatment ($D(1) = 0$ and $D(0) = 1$).

- This doesn't come up in a constant effect model where $\beta = Y(1) - Y(0)$ is constant, as in such case

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] &= \beta\{P\{D(1) > D(0)\} - P\{D(1) < D(0)\}\} \\ &= \beta E[D(1) - D(0)] \,, \end{aligned}$$

  and so a zero reduced-form effect means either the first stage is zero or $\beta = 0$.

# Monotonicity in Latent Index (Roy) Models

## Roy Model

$$D = I\{\nu(Z) - V > 0\} = I\{\text{ Utility of choosing 1} > \text{Utility of choosing 0 }\}$$

- **Monotonicity**: Equivalent to a Roy model with separable utility (easy to interpret)

- **Roy Model**: individual choices are determined by a threshold crossing rule involving observed and unobserved components of the utility. Take $\nu(Z) = \gamma_0 + \gamma_1 Z$ so that

$$D = \begin{cases} 1 & \text{if } \gamma_0 + \gamma_1 Z - V > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where $\gamma_1 > 0$ (wlog) and $V$ is an unobserved heterogeneity assumed to be independent of $Z$.

- This latent index model characterizes potential treatment assignments as

$$D(0) = I\{\gamma_0 > V\} \text{ and } D(1) = I\{\gamma_0 + \gamma_1 > V\}\,.$$

- Monotonicity assumption is automatically satisfied since $\gamma_1 > 0$ (symmetric argument for $\gamma_1 < 0$).

# Monotonicity with one-sided compliance

- **Randomized trial with non-compliance**: the treatment assignment as an "offer of treatment" $Z$ (the instrument) and the actual treatment $D$ determines whether the subject actually had the treatment.
- Assume no one in the control group has access to the treatment: $D(0) = 0$ while $D(1) \in \{0, 1\}$
- Monotonicity automatically holds: $D(1) \geqslant D(0)$
- Since $D(1)$ is a choice, a comparison between those actually treated ($D = 1$) and the control ($D = 0$) group is misleading. Two alternatives are frequently used.
- **Intention to Treat Effect**: a comparison between those who where *offered* treatment ($Z = 1$) and the control ($Z = 0$) group.
- **LATE=ATT**: IV using $Z$ as an instrumental variable for $D$, which leads to LATE. Since $D(0) = 0$, LATE returns the effect of *treatment on the treated*, i.e., $E[Y(1) - Y(0)|D = 1]$.

# THE END!