LECTURE 4: ENDOGENEITY

Ivan A. Canay
Northwestern University
Past & Future

**So Far**
- Three Interpretations of $\beta$
- Solving and estimating sub-vectors of $\beta$
- Properties of LS
- Estimating $\nabla$
- Classical Problems that lead to $E[XU] \neq 0$

**Today**
- Instrumental Variables
- The IV Estimator
- The 2SLS Estimator
- Properties of 2SLS
- Estimating $\nabla$
Let \((Y, X, U)\) be a random vector where \(Y\) and \(U\) take values in \(\mathbb{R}\) and \(X\) takes values in \(\mathbb{R}^{k+1}\). 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 \(\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbb{R}^{k+1}\) be such that

\[Y = X'\beta + U.\]

We do not assume \(E[XU] = 0\). Any \(X_j\) such that \(E[X_jU] = 0\) is said to be \textit{exogenous}; any \(X_j\) such that \(E[X_jU] \neq 0\) is said to be \textit{endogenous}. Normalizing \(\beta_0\) if necessary, we view \(X_0\) as exogenous.

\textbf{Instrument}: to overcome the difficulty associated with \(E[XU] \neq 0\), we assume that there is an additional random vector \(Z\) taking values in \(\mathbb{R}^{\ell+1}\) with \(\ell + 1 \geq k + 1\) such that \(E[ZU] = 0\).

Any exogenous component of \(X\) is contained in \(Z\) (the so-called \textit{included instruments}). In particular, we assume the first component of \(Z\) is constant equal to one, i.e., \(Z = (Z_0, Z_1, \ldots, Z_\ell)'\) with \(Z_0 = 1\).

We also assume that \(E[ZX'] < \infty\), \(E[ZZ'] < \infty\) and that there is no perfect collinearity in \(Z\).
Instrumental Variables

- We assume 1) $E[ZU] = 0$, 2) $E[ZX'] < \infty$, 3) $E[ZZ'] < \infty$, and 4) there is no perfect collinearity in $Z$.

- The requirement that $E[ZU] = 0$ is termed *instrument exogeneity*.

- We further assume 5) the rank of $E[ZX']$ is $k+1$. This is termed *instrument relevance* or *rank condition*.

- A necessary condition for 5) to be true is $\ell \geq k$. This is referred to as the *order condition*.

- Using that $U = Y - X'\beta$ and $E[ZU] = 0$, we see that $\beta$ solves the system of equations

\[
E[ZY] = E[ZX']\beta .
\]

- Since $\ell + 1 \geq k + 1$, this may be an over-determined system of equations.
Suppose there is no perfect collinearity in $Z$ and let $\Pi$ be such that $BLP(X|Z) = \Pi'Z$. 

$E[ZX']$ has rank $k + 1$ if and only if $\Pi$ has rank $k + 1$. Moreover, the matrix $\Pi'E[ZX']$ is invertible.
Solving for $\beta$

$\beta$ solves: $E[ZY] = E[ZX'] \beta$ or $\Pi' E[ZY] = \Pi' E[ZX'] \beta$

Using the previous lemma and $\Pi = E[ZZ']^{-1} E[ZX']$, we can derive three formulae for $\beta$. 
Interpreting the rank condition

**Interpretation:** Consider the case where $k = \ell$ and only $X_k$ is endogenous. Let $Z_j = X_j$ for all $0 \leq j \leq k - 1$. In this case,

$$
\Pi' = \begin{pmatrix}
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0 \\
\pi_0 & \pi_1 & \cdots & \pi_{\ell-1} & \pi_{\ell}
\end{pmatrix}.
$$

The rank condition therefore requires $\pi_{\ell} \neq 0$: the instrument $Z_{\ell}$ must be “correlated with $X_k$ after controlling for $X_0, X_1, \ldots, X_{k-1}$.”
Partition of $\beta$: endogenous components

- Partition $X$ into $X_1$ and $X_2$, where $X_2$ is exogenous. Partition $Z$ into $Z_1$ and $Z_2$ and $\beta$ into $\beta_1$ and $\beta_2$ analogously.

- Note that $Z_2 = X_2$ are included instruments and $Z_1$ are excluded instruments. Then,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U.$$ 

- We can conveniently re-write this by projecting (BLP) on $Z_2 = X_2$. Consider the case $k = \ell$

$$\text{BLP}(Y|Z_2) = \text{BLP}(X_1|Z_2)'\beta_1 + X_2'\beta_2.$$ 

- Define $Y^* = Y - \text{BLP}(Y|Z_2)$ and $X_1^* = X_1 - \text{BLP}(X_1|Z_2)$ so that

$$E[Z_1 Y^*] = E[Z_1 X_1^*]'\beta_1 + E[Z_1 U]$$ 

- It follows that

$$\beta_1 = E[Z_1 X_1^*]'^{-1}E[Z_1 Y^*].$$
QUESTIONS?
**Estimating \( \beta \): The IV Estimator**

- **Just identified case:** \( k = \ell \). Denote by \( P \) the marginal distribution of \((Y, X, Z)\).

- 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 \( P \).

- By analogy with \( \beta = E[ZX']^{-1}E[ZY] \), the natural estimator of \( \beta \) is simply

\[
\hat{\beta}_n = \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_iX'_i \right)^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_iY_i \right) .
\]

- This estimator is called the *instrumental variables (IV)* estimator of \( \beta \). Note that \( \hat{\beta}_n \) satisfies

\[
\frac{1}{n} \sum_{1 \leq i \leq n} Z_i(Y_i - X'_i\hat{\beta}_n) = 0 .
\]

In particular, \( \hat{U}_i = Y_i - X'_i\hat{\beta}_n \) satisfies

\[
\frac{1}{n} \sum_{1 \leq i \leq n} Z_i\hat{U}_i = 0 .
\]
Insight on the IV estimator: assume $X_0 = 1$ and $X_1 \in \mathbb{R}$. An interesting interpretation of the IV estimator of $\beta_1$ is obtained by multiplying and dividing by $\frac{1}{n} \sum_{i=1}^{n} (Z_{1,i} - \bar{Z}_{1,n})^2$, i.e.,

$$\hat{\beta}_{1,n} = \frac{\frac{1}{n} \sum_{i=1}^{n} (Z_{1,i} - \bar{Z}_{1,n}) Y_i}{\frac{1}{n} \sum_{i=1}^{n} (Z_{1,i} - \bar{Z}_{1,n}) X_{1,i}}$$

$$= \frac{\frac{1}{n} \sum_{i=1}^{n} (Z_{1,i} - \bar{Z}_{1,n}) Y_i}{\frac{1}{n} \sum_{i=1}^{n} (Z_{1,i} - \bar{Z}_{1,n})^2}$$
This estimator may be expressed more compactly using matrix notation. Define

$$Z = (Z_1, \ldots, Z_n)'$$
$$X = (X_1, \ldots, X_n)'$$
$$Y = (Y_1, \ldots, Y_n)'$$

In this notation, we have

$$\hat{\beta}_n = (Z'X)^{-1}(Z'Y).$$
The Two-Stage Least Squares (TSLS) Estimator

- **Over-identified case:** \( \ell > k \)

- The expressions we derived for \( \beta \) in this case, like

\[
\beta = E[\Pi' E[ZX']]^{-1} \Pi' E[ZY],
\]

all involved the matrix \( \Pi \), where

\[
BLP(X|Z) = \Pi' Z.
\]

- An estimate of \( \Pi \) can be obtained by OLS.

- Since \( \Pi = E[ZZ']^{-1} E[ZX'] \), a natural estimator of \( \Pi \) is

\[
\hat{\Pi}_n = \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i X_i' \right).
\]
The Two-Stage Least Squares (TSLS) Estimator

Let $X_i = \hat{\Pi}_n' Z_i + \hat{V}_i$ where $\hat{\Pi}_n = \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i X_i' \right)$.

With this estimator of $\Pi$, a natural estimator of $\beta$ is simply
The **TSLS Estimator**

- Note that \( \hat{\beta}_n \) satisfies
  
  \[
  \frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i (Y_i - X_i' \hat{\beta}_n) = 0.
  \]

- In particular, \( \hat{U}_i = Y_i - X_i' \hat{\beta}_n \) satisfies
  
  \[
  \frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i \hat{U}_i = 0.
  \]

- This implies that \( \hat{U}_i \) is orthogonal to all of the instruments equal to an exogenous regressors, but may not be orthogonal to the other regressors.

- It is termed the TSLS estimator because it may be obtained in the following way:
  
  1. regress (each component of) \( X_i \) on \( Z_i \) to obtain \( \hat{X}_i = \hat{\Pi}_n' Z_i \);

  2. regress \( Y_i \) on \( \hat{X}_i \) to obtain \( \hat{\beta}_n \). However, in order to obtain proper standard errors, it is recommended to compute the estimator in one step (see the following section).
This estimator may be expressed more compactly using matrix notation. Define

\[
\begin{align*}
Z &= (Z_1, \ldots, Z_n)' \\
X &= (X_1, \ldots, X_n)' \\
Y &= (Y_1, \ldots, Y_n)' \\
\hat{X} &= (\hat{X}_1, \ldots, \hat{X}_n)' \\
\hat{X} &= \hat{P}_Z X,
\end{align*}
\]

where

\[
P_Z = Z(Z'Z)^{-1}Z'
\]

is the projection matrix onto the column space of \( Z \). In this notation, we have

\[
\hat{\beta}_n = (\hat{X}' \hat{X})^{-1}(\hat{X}' Y) = (\hat{X}' \hat{X})^{-1}(\hat{X}' Y) = (X' P_Z X)^{-1}(X' P_Z Y).
\]
QUESTIONS?
Let \((Y, X, U)\) be a random vector where \(Y\) and \(U\) take values in \(\mathbb{R}\) and \(X\) takes values in \(\mathbb{R}^{k+1}\). 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 \(\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbb{R}^{k+1}\) 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 \(P\).

Under these assumptions the TSLS estimator is **consistent** for \(\beta\), and under the additional requirement that \(\text{Var}[ZU] < \infty\), it is **asymptotically normal** with limiting variance

\[
V = E[\Pi'ZZ'\Pi]^{-1}\Pi'\text{Var}[ZU]\Pi E[\Pi'ZZ'\Pi]^{-1}.
\]
Consistency of TSLS

\[ \hat{\beta}_n = \left( \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i X_i' \right) \right)^{-1} \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Y_i \right) \xrightarrow{P} \beta \text{ as } n \to \infty. \]
Assume that \( \text{Var}[ZU] = E[ZZ'U^2] < \infty \). Then, as \( n \to \infty \),

\[
\sqrt{n}(\hat{\beta}_n - \beta) \overset{d}{\to} N(0, V).
\]
Estimation of $V$

A natural estimator of $V$ is given by

$$
\hat{V}_n = \left( \hat{n}^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right) \hat{n} \right) \times \hat{n}^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \hat{U}_i^2 \right) \hat{n} \times \left( \hat{n}^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right) \hat{n} \right)^{-1},
$$

where $\hat{U}_i = Y_i - X_i' \hat{\beta}_n$.

- **Primary difficulty** in establishing the consistency of this estimator lies in showing that

$$
\frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \hat{U}_i^2 \xrightarrow{p} \text{Var}[ZU]
$$

as $n \to \infty$. The complication lies in the fact that we do not observe $U_i$ and therefore have to use $\hat{U}_i$.

- However, the desired result can be shown by arguing exactly as in the second part of this class.

- **Note:** $\hat{U}_i = Y_i - X_i' \hat{\beta}_n \neq Y_i - \hat{X}_i' \hat{\beta}_n$, so the standard errors from two repeated applications of OLS will be incorrect.
QUESTIONS?