# Topic 18 Header Slide

- Topic 18. Analysis of instrumental variable estimators
  - Discussion point #1. Identification
  - Discussion point #2. Instrument validity and relevance
  - Discussion point #3. Statistical properties of instrumental variable estimators

---

| Estimator | Notation | Under Complications | Extends to |
|---|---|---|---|
| 1. | $\hat{\beta}_{OLS}$ | $A2Nonlinear.additiveError$ | $\rightarrow \hat{\beta}_{NLLS}$ |
| 2. | $\hat{\beta}_{LAD}$ | | |
| 3. | $\hat{\beta}_{Lstar}$ | | |
| 4. | $\hat{\beta}_{GMM}$ | $A3EndogenousX/ExogenousW$ | $\rightarrow \hat{\beta}_{IVE}$ |
| $5a.$ | $\hat{\beta}_{IGLS}$ | | |
| $5b.$ | $\hat{\beta}_{FGLS}$ | | |
| 6. | $\hat{\beta}_{MLE}$ | $Multiple$ | $\rightarrow \hat{\beta}_{FIMLE}$ |
| 7. | $\hat{\beta}_{IVE}$ | $A2Nonlinear.AdditiveError$ | $\rightarrow \hat{\beta}_{NLIVE}$ |

# Basic Idea of Instrumental Variables

Consider $A2linear$ with two sets of regressors, $X^G$ and $X^B$:

$$y = X^G \beta^G + X^B \beta^B + \epsilon^{true}$$

or, in observation-by-observation form:

$$y_s = x_s^{G'} \beta^G + x_s^{B'} \beta^B + x_s^{B'} \beta^B + \epsilon_s^{true} = x_s' \beta + \epsilon_s^{true}$$

Given $A1$ and $A2$, the Sampling Error Vector of OLS is:

$$SEV(\hat{\beta}_{ols}) = \left( \sum_s B_s \right)^{-1} \cdot \sum_s a_s^{ols} = \left( \sum_s x_s x_s' \right)^{-1} \cdot \sum_s x_s \epsilon_s$$

But:

$$Ea_s^{ols} = E \begin{bmatrix} x_s^G \epsilon_s \\ x_s^B \epsilon_s \end{bmatrix} = \begin{bmatrix} Ex_s^G \epsilon_s \\ Ex_s^B \epsilon_s \end{bmatrix} = \begin{bmatrix} = 0 \\ \neq 0 \end{bmatrix}$$

because we are told that the 'good' variables satisfy the weak exogeneity condition $Ex^G \epsilon_s = 0$, while the 'bad' variables do not (since they are *endogenous* w.r.t. to the error).
Therefore, OLS will be inconsistent for all the $\beta$s since in general $X^G$ and $X^B$ are correlated.

# Reverse Engineering the Instrumental Variables Estimator

Suppose we can find a data matrix $W$ of the same dimension as the original $X$ and of full rank $k$. We then define:

$$W'\epsilon^{true} = \sum_s w_s \epsilon_s^{true} = \sum_s a_s^{ive}$$

such that:

$$Ea_s^{ive} = 0$$

at the true parameter values.
We now use the GMM idea and rely on the true *population* orthogonality conditions implied by the true model: $A2 : y = X\beta + \epsilon^{true}$ and

$$A3Rsru : W, X^G E w_s \epsilon^{true} = 0, E x_s^{G'} = 0$$

Therefore, we define the GMM=IVE by using the *sample* orthogonality conditions:

$$W'\hat{\epsilon}^{ive} = W'(y - X\hat{\beta}_{ive}) = 0$$

to mimic the population OCs. Finally, solving for $\hat{\beta}^{ive}$ we obtain:

$$\hat{\beta}_{ive} = (W'X)^{-1}W'y$$

because $W'X$ is square and invertible given that $rank(X) = rank(W) = k$.
In conclusion, the IVE will be consistent provided every column used to construct $W$ is a weakly exogenous variable w.r.t. the true error (i.e., satisfies A3Rsru).

## Overview/Discussion point #1.

- Discussion point #1. **Identification**
  - Identification refers to the *mathematical* ability to solve uniquely for all parameters. This is distinct from estimation, which aims to use *statistical* methods to learn about unknown parameters given a sample.
  - Parameter identification is a step in the theoretical analysis of the model given a specification; parameter estimation is a step in empirical analysis of the model given data.
  - We must assess whether our model is identified (so as to subsequently permit consistent estimation) in the presence of an exogenous-endogenous regressor dichotomy.
- Discussion point #2. **Instrument validity and relevance (and feasibility)**
  - Validity and relevance are the names given to our identification conditions. These conditions permit us to use IVE methods to consistently estimate parameters of interest.
- Discussion point #3. **Statistical properties of instrumental variable estimators**
  - The IVE method typically yields analytic estimators. This permits us to use the standard form of our (previously-seen) SEV in order to investigate exact and asymptotic properties.

## Analysis of the SEV – Discussion point #2.

- Recall any of the seven leading cases of endogeneity. In other words, suppose we have a violation of even the weakest form of exogeneity (i.e. even $A3Rsru$ does not hold).

- Recall analysis of $SEV(\hat{\beta}_{OLS})$. Under $A2linear$ and with the definition of OLS, we have

$$\hat{\beta}_{OLS} - \beta = \left( \frac{1}{S} \sum_{s=1}^{S} x_s x_s' \right)^{-1} \frac{1}{S} \sum_{s=1}^{S} x_s \varepsilon_s,$$

  so that by Slutsky's theorem, we obtain

$$\operatorname*{plim}_{S\to\infty} \hat{\beta}_{OLS} - \beta = \left( \operatorname*{plim}_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} x_s x_s' \right)^{-1} \operatorname*{plim}_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} x_s \varepsilon_s,$$

  which by "suitable" LLNs evaluates to,

$$\operatorname*{plim}_{S\to\infty} \hat{\beta}_{OLS} - \beta = \left(\mathbb{E}(x_s x_s')\right)^{-1} \mathbb{E}(x_s \varepsilon_s),$$

  a quantity that is only well-defined and equal to zero under the critical assumptions that:

  1. There exists a finite non-singular matrix, $B_{\infty}^{0}$, such that $\mathbb{E}(x_s x_s') = B_{\infty}^{0}$
  2. $\left(\sum_{s=1}^{S} x_s x_s'\right)^{-1}$ exists, or $A1$ holds (otherwise the first step of the proof fails)
  3. $\mathbb{E}(x_s \varepsilon_s) = \mathbb{E}(a_s^{ols}) = 0$, or $A3Rsru$ holds (otherwise the final step of the proof fails)

## Considering the loss of $A3Rsru$ – Discussion point #2.

- The loss of $A3Rsru$ is clearly fatal as far as our OLS estimator is concerned.
- But suppose we could (in principle, or in practice, and ideally both!) avail of data on $k$-dimensional vector $w_s$ (i.e. for $s = 1, ..., S$) that happened to exhibit weak exogeneity in respect of $\varepsilon_s$. Let us denote this new assumption by $A3Rsru.W : \mathbb{E}(w_s \varepsilon_s) = \mathbb{E}(a_s^{ive}) = 0$.
- How could we exploit this windfall source of exogenous variation?
- We cannot blindly replace $x_s$ with $w_s$ in the previous proof – this would constitute a major mis-specification of the original model. (Remember that no one is interested in estimating the partial effect of $w_s$ on $y_s$. What we care about is the partial effect of $x_s$ on $y_s$.)
- One could, however, develop a whole new approach (OLS simply cannot be rescued) as per the following slide. Let us first consider the algebra and then consider the intuition.

## SEV-based introduction to IVE (1 of 4) – Discussion point #2.

- Consider again our consistency proof, but this time we will try to avail of the windfall data on $w_s$ for $s = 1, ..., S$. Consider a thought-experiment in which we analyse the properties of

$$\hat{\beta}_{novel} \equiv \left( \frac{1}{S} \sum_{s=1}^{S} w_s x_s' \right)^{-1} \frac{1}{S} \sum_{s=1}^{S} w_s y_s$$

- (Ultimately, this "novel" estimator will be exactly our IV estimator, and we will drop the odd-sounding name. But let us just work through the mechanics first so that we can define what we mean by an instrument.)

# SEV-based introduction to IVE (2 of 4) – Discussion point #2.

- Consider the analysis of $SEV(\hat{\beta}_{novel})$. Under $A2linear$ and given the definition of our novel estimator, we have

$$\hat{\beta}_{novel} - \beta = \left( \frac{1}{S} \sum_{s=1}^{S} w_s x_s' \right)^{-1} \frac{1}{S} \sum_{s=1}^{S} w_s \varepsilon_s,$$

so that by Slutsky's theorem, we obtain

$$\operatorname*{plim}_{S\to\infty} \hat{\beta}_{novel} - \beta = \left( \operatorname*{plim}_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} w_s x_s' \right)^{-1} \operatorname*{plim}_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} w_s \varepsilon_s,$$

which by a suitable LLN evaluates to,

$$\operatorname*{plim}_{S\to\infty} \hat{\beta}_{novel} - \beta = \left( \mathbb{E}(w_s x_s') \right)^{-1} \mathbb{E}(w_s \varepsilon_s),$$

which is only well-defined and equal to zero under the critical assumptions that:

1. There exists a finite non-singular matrix, $B_\infty^0$, such that $\mathbb{E}(w_s x_s') = B_\infty^0$
2. $\left( \sum_{s=1}^{S} w_s x_s' \right)^{-1}$ exists (otherwise the first step of the proof fails)
3. $\mathbb{E}(w_s \varepsilon_s) = \mathbb{E}(a_s^{ive}) = 0$, or $A3Rsru.W$ holds (otherwise the final step of the proof fails)

## SEV-based introduction to IVE (3 of 4) – Discussion point #2.

- Clearly, under conditions 1–3 on the previous slide, our novel estimator is consistent. (One can arguably think of these conditions as reverse-engineered to ensure consistency!)
- All that remains is for us to give $w_s$ a scientific name; that is, an instrument.
- Indeed, an instrument is defined precisely such that conditions 1–3 hold. These defining conditions, in turn, are so crucial that they bear their own names. Let us review them:
  1. There exists a finite non-singular matrix, $B^0_\infty$, such that $\mathbb{E}(w_s x'_s) = B^0_\infty$
  2. $\left(\sum_{s=1}^{S} w_s x'_s\right)^{-1}$ exists (otherwise the first step of the proof fails)
  3. $\mathbb{E}(w_s \varepsilon_s) = \mathbb{E}(a^{ive}_s) = 0$, or $A3Rsru.W$ holds (otherwise the final step of the proof fails)
- Conventional nomenclature for the bullets above is as follows:
  1. Condition 1 is none other than "relevance". It ensures the inverse term in the final step of the consistency proof is well-defined.
  2. Condition 2 is "in-sample relevance" or "feasibility". It ensures the IV estimator exists to begin with. (For instance, it precludes perfect collinearity among the $w_s$ variables.)
  3. Condition 3 is none other than "validity". It ushers in the final consistency result.

## SEV-based introduction to IVE (4 of 4) – Discussion point #2.

We are now in a position to (i) define what an instrument means; and (ii) concurrently move the discussion forwards by setting up the notation to generalise our framework from the so-called "just-identified" setting to the "over-identified" setting (which we will analyse in detail shortly).

- A $q$-dimensional vector $w_s$ is said to be instrumental for $k$-dimensional vector of (potentially endogenous) regressors $x_s$ if $w_s$ satisfies the twin conditions of validity and relevance.
  - When $q < k$, the model is said to be "under-identified". (No solution is possible.)
  - When $q > k$, the model is said to be "over-identified". (We will study 2SLS shortly.)
  - When $q = k$, the model is said to be "just-identified". (Our current focus – i.e., IVE.)
- So long as $q = k$, we can define the IV estimator as:

$$\hat{\beta}_{IVE} \equiv (W'X)^{-1}W'y = \left(\sum_{s=1}^{S} w_s x_s'\right)^{-1} \sum_{s=1}^{S} w_s y_s, \text{ where } W \equiv \begin{pmatrix} w_1' \\ w_2' \\ \vdots \\ w_S' \end{pmatrix} \text{ is an } S \times k \text{ matrix}$$

  satisfying the needed feasibility condition that $(W'X)^{-1}$ exists.
- Note that when $q > k$, the over-identified setting, the inverse of $(W'X)$ cannot exist (at least not in a conventional sense) because it is not a square matrix.

## Asymptotics of the IV estimator – Discussion point #3.

The key results in the IVE case, given $S \times k$ instrument matrix $W$ are as follows:

(i) Under relevance, an LLN ensures that $\operatorname*{plim}_{S\to\infty} (W'X/S)$ is a finite non-singular $k \times k$ matrix.

(ii) Under validity, an LLN ensures that $\operatorname*{plim}_{S\to\infty} (W'\varepsilon/S)$ is the $k$-dimensional zero vector.

(iii) Under the feasibility condition, the SEV of the estimator, $(W'X)^{-1}W'\varepsilon$, is well-defined.

(iv) We will also require a 'well-behaved' population second moment matrix for instruments $W$ as captured by existence and non-singularity of $\operatorname*{plim}_{S\to\infty} (W'W/S)$.

(v) Under these conditions, a CLT ensures that $W'\varepsilon/\sqrt{S} \xrightarrow{d} \mathcal{N}(0, \Sigma)$ as $S$ passes to infinity, where $\Sigma = \operatorname*{plim}_{S\to\infty} (W'W/S)$ under $A4GM(iid)$ and $\Sigma = \operatorname*{plim}_{S\to\infty} (W'\Omega W/S)$ under $A4\Omega$.

(vi) It follows from (i)–(iii) and Slutsky's theorem that $SEV(\hat{\beta}_{IVE}) \xrightarrow{p} 0$ as $S \to \infty$; and it follows from (i)–(v) and Cramér's theorem that $\sqrt{S} \cdot SEV(\hat{\beta}_{IVE}) \xrightarrow{d} Z_0$ as $S \to \infty$, for $k \times 1$ vector $Z_0$, where $Z_0 \sim \mathcal{N}(0, \Sigma_Z)$ with the definition of $\Sigma_Z$ as per:

$$c^2 \left( \operatorname*{plim}_{S\to\infty} (W'X/S) \right)^{-1} \operatorname*{plim}_{S\to\infty} (W'\Omega W/S) \left( \operatorname*{plim}_{S\to\infty} (X'W/S) \right)^{-1}$$

under $A4\Omega$.

## A mix of endogenous-exogenous regressors (the model)

- Consider a situation in which we have a single endogenous regressor and several hundred exogenous regressors. The temptation might be to think that the exogenous variation, in some sense, swamps or drowns-out the endogenous variation. However, this reasoning is incorrect. We see why below.

- Suppose we have a linear regression model with data $(y, X)$ of sample size $S$. The $k$ regressors are grouped in two parts, $X_A$ and $X_B$, of dimensions $S \times k_A$ and $S \times k_B$ respectively, with $k_A + k_B = k$. Suppose that the model satisfies the following assumptions:

$$
\begin{array}{ll}
A1: & rank(X) = k < S \\
A2: & y = \quad X\beta + \varepsilon = \underbrace{X_A\beta^A}_{X^{Good}\beta^{Good}} + \underbrace{X_B\beta^B}_{X^{Bad}\beta^{Bad}} + \varepsilon = \underbrace{Z^I\beta^G}_{X^{Good}\beta^{Good}} + \underbrace{X_B\beta^B}_{X^{Bad}\beta^{Bad}} + \varepsilon \\
A3Rmi.X_A: & \mathbb{E}(\varepsilon|X_A) = \mathbb{E}(\varepsilon) \\
A3.X_B: & \varepsilon \text{ and } X_B \text{ correlated for all } s = 1, ..., S \\
A4\Omega: & \mathbb{E}(\varepsilon\varepsilon'|X) = c^2\Omega \\
A5Gaussian: & \varepsilon_s|X \sim \mathcal{N}(0, \sigma^2)
\end{array}
$$

  In other words, regressors $X_B$ are endogenous with respect to the true error.

- We are particularly interested in the true coefficients of the $X_A$ variables, $\beta_A$.

## A mix of endogenous and exogenous regressors (the problem)

- The sampling error vectors for OLS and IGLS estimators is respectively:

$$\begin{pmatrix} \hat{\beta}_A^{OLS} \\ \hat{\beta}_B^{OLS} \end{pmatrix} - \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} = \begin{pmatrix} X_A'X_A & X_A'X_B \\ X_B'X_A & X_B'X_B \end{pmatrix}^{-1} \begin{pmatrix} X_A'\varepsilon \\ X_B'\varepsilon \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_A^{IGLS} \\ \hat{\beta}_B^{IGLS} \end{pmatrix} - \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} = \begin{pmatrix} X_A'\Omega^{-1}X_A & X_A'\Omega^{-1}X_B \\ X_B'\Omega^{-1}X_A & X_B'\Omega^{-1}X_B \end{pmatrix}^{-1} \begin{pmatrix} X_A'\Omega^{-1}\varepsilon \\ X_B'\Omega^{-1}\varepsilon \end{pmatrix}$$

- Given these expressions, it is clear that both estimators will be biased and inconsistent for coefficients on both sets of regressors (i.e., sets $A$ and $B$). This is because:

  1. the endogeneity of regressors $X_B$ implies that the terms $X_B'\varepsilon$ and $X_B'\Omega^{-1}\varepsilon$ will not be zero in expectation, nor will $X_B'\varepsilon/S$ and $X_B'\Omega^{-1}\varepsilon/S$ converge to zero asymptotically;
  2. the bias/inconsistency carries over even to the estimated coefficients on the $X_A$ variables because in general neither $X_A'X_B$ nor $X_A'\Omega^{-1}X_B$ will vanish.

---

- Detailed formal algebraic explanations (i.e., based on partitioned regression formulas) of the above statements are provided in the extended notes. (These are DrRS's old teaching notes based on the associated problem set question – i.e., the same one that was ear-marked for you to submit to your class teachers. Compare his solution with your own...? Hope it helps.)

## A mix of endogenous and exogenous regressors (IVE – 1 of 2)

As a possible response to the problem of endogeneity of $X_B$, econometricians proposed the IVE-based solution, defined by:

$$\hat{\beta}_{IVE} = (W'X)^{-1}W'y, \text{ where}$$

$$SEV(\hat{\beta}_{IVE}) = \hat{\beta}_{IVE} - \beta^{true} = (W'X)^{-1}W'\varepsilon$$

Here is a recap of the key points; now from a practitioner's (rather than a theorist's) perspective:

- The matrix $W$ should be of the same dimension as $X$ and have full rank $k$. Further, $W'X$ should have full rank $k = k_A + k_B$.
- Further, $W$ should consist only of exogenous variables, implying that $X_A$ can be used.
- Since this disallows the use of the endogenous regressor $X_B$, in order to make the method feasible, we must find $k_E$ additional "instrument" variables ($k_E \geq k_B$) to construct $W$.
- As a reminder, $W$ should consist solely of valid instruments or linear combinations of such valid instruments. An instrument variable $z$ is termed "valid" if it is weakly exogenous w.r.t. to the error term – i.e., $\mathbb{E}(z_s\varepsilon_s) = 0$.
- An instrument variable $z$ is termed "relevant" if it has a high correlation with the endogenous variables of the model, in this case the endogenous regressors $X_B$.

## A mix of endogenous and exogenous regressors (IVE – 2 of 2)

- Notice the potential for over-identification alluded to in the previous slide. If $k_E$ were *strictly* larger than $k_B$, we would have to eliminate the extra instruments so as to maintain feasibility.
- However, wasting information is never a good idea. Formally, there are efficiency gains to be exploited by (linearly) combining the information across all $k_z = k_I + k_E$ available instruments.
- So what is the optimal choice, say $W^*$, in the over-identified setting? We explore this next.

## Dealing with over-identification

- Say we have available an $S \times k_z$ matrix of (valid and relevant) instruments, $Z$, where $k_z > k$.
- Then, the optimal choice would entail construction of an $S \times k$ matrix, $W^*$, that constitutes effectively a $k = (k_A + k_B) < k_z = (k_A + k_E)$ dimensional fitted (via least squares) version of all available weakly exogenous variables, given by $Z = (Z^I, Z^E) = (X_A, Z^E)$.
- Indeed, the optimal combination – one that maximises the correlations between $W^*$ and the original regressors $(X_A, X_B)$ – is obtained via the fitted/predicted values from a 'multivariate' regression, using OLS, of $S \times (k_A + k_B)$ matrix $(X_A, X_B)$ on $S \times (k_A + k_E)$ matrix $(X_A, Z^E)$.
- Let us denote these fitted/predicted values as $S \times k$ matrix $\hat{X} = (\hat{X}_A, \hat{X}_B) = (X_A, \hat{X}_B)$.
- The multivariate regression described on this slide is known as the "1st stage regression".
- To summarise we have our $S \times k$ optimal instrument matrix, $W^*$, given by:

$$W^* = \hat{X} = Z\hat{\pi} = Z(Z'Z)^{-1}Z'X,$$

  where $\hat{\pi} = (Z'Z)^{-1}Z'X$ is a $k_z \times k$ matrix of so-called 1st stage regression coefficients and $\hat{Z} = Z\hat{\pi}$ is the $S \times k$ (i.e. lower-dimensional) projection.
- The so-called optimal IVE is then obtained as

$$\hat{\beta}_{Opt.IVE} = (W^{*\prime}X)^{-1}W^{*\prime}y$$