Chapter 5: Introduction to Instrumental Variables and the Endogeneity Problem

Frontier Topics in Empirical Economics

Lecture notes based on slides by Zibin Huang
College of Business, Shanghai University of Finance and Economics

October 2025

Contents

1 Introduction: The Endogeneity Problem

In Chapter 1, we introduced the potential outcomes framework and emphasized that the central challenge of causal inference is the selection problem: individuals who receive treatment may differ systematically from those who do not, even in the absence of treatment. When we can run a randomized controlled trial (RCT), the treatment assignment is independent of potential outcomes, and a simple comparison of means identifies the average treatment effect. In practice, however, most economic variables of interest are not randomly assigned. This chapter introduces the most important tool economists have developed to address this problem: **instrumental variables (IV)**.

Consider one of the oldest and most studied questions in labor economics: what is the causal effect of schooling on wages? For individual i, suppose the relationship is linear and homogeneous (constant treatment effect):

$$Y_i = \alpha + \rho s_i + \eta_i, \tag{1}$$

where Y_i is wage, s_i is years of schooling, and η_i is an unobserved error term that captures everything else affecting wages. If schooling were randomly assigned, then s_i would be independent of η_i, and the OLS estimator of ρ would consistently estimate the average treatment effect (ATE). But schooling is usually an endogenous choice made by the individual. People who attend college tend to have higher innate ability, stronger motivation, or better family backgrounds, all of which independently affect wages. This is **selection bias**.

To make the problem concrete, suppose that ability A_i is the key unobserved confounder. We can decompose the error as:

$$\eta_i = \gamma A_i + \nu_i, \tag{2}$$

where ν_i is independent of s_i. Plugging (2) into (1):

$$Y_i = \alpha + \rho s_i + \gamma A_i + \nu_i. \tag{3}$$

If A_i is observed, we can simply control for it and OLS on (3) consistently estimates ρ. But if A_i is unobserved, we have an **omitted variable bias** problem: OLS on (1) conflates the causal effect of schooling with the correlation between schooling and ability.

What can we do? This is where instrumental variables come in.

2 Simple IV: Definition and Identification

Let us begin with the simplest setting: a single endogenous variable, a single instrument, and a constant treatment effect. We return to the model in (1):

$$Y_i = \alpha + \rho s_i + \eta_i.$$

Definition 2.1 (Instrumental Variable)**.** A variable z_i is called an **instrumental variable** for s_i if it satisfies two conditions:

1. **Exogeneity (Exclusion Restriction):** $z_i \perp\!\!\!\perp \eta_i$. The instrument is uncorrelated with the unobserved determinants of the outcome.
2. **Relevance (Existence of First Stage):** $\operatorname{Cov}(s_i, z_i) \neq 0$. The instrument is correlated with the endogenous variable.

The exclusion restriction says that z_i affects Y_i *only through* s_i—there is no direct effect of z_i on Y_i other than through its influence on the endogenous variable. The relevance condition ensures that z_i actually shifts s_i; without it, we have no "lever" to move the endogenous variable.

2.1 Identification via Covariances

How does an instrument help us identify ρ? The key insight is to compute the covariance of z_i with both sides of (1):

$$\operatorname{Cov}(z_i, Y_i) = \operatorname{Cov}(z_i, \alpha + \rho s_i + \eta_i) = \rho \operatorname{Cov}(z_i, s_i),$$

where the last equality uses the exogeneity condition $\operatorname{Cov}(z_i, \eta_i) = 0$. Solving for ρ:

IV Identification Formula

$$\rho = \frac{\operatorname{Cov}(z_i, Y_i)}{\operatorname{Cov}(z_i, s_i)} = \frac{\operatorname{Cov}(z_i, Y_i)/\operatorname{Var}(z_i)}{\operatorname{Cov}(z_i, s_i)/\operatorname{Var}(z_i)}. \tag{4}$$

The causal effect is identified by the ratio of two covariances—or equivalently, the ratio of two regression coefficients.

When the instrument z_i is binary, this formula simplifies to the **Wald estimator**:

$$\rho = \frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[s_i \mid z_i = 1] - E[s_i \mid z_i = 0]}. \tag{5}$$

The numerator is the effect of the instrument on the outcome (the "reduced form"), and the denominator is the effect of the instrument on the endogenous variable (the "first stage"). The ATE is the ratio of the two.

2.2 The Wald Estimator

To see the connection more formally, consider two regressions:

$$s_i = \alpha + \pi_1 z_i + \eta_{1i} \qquad \text{(First Stage)}, \tag{6}$$
$$Y_i = \alpha + \pi_2 z_i + \eta_{2i} \qquad \text{(Reduced Form)}. \tag{7}$$

Each of π_1 and π_2 is a simple regression coefficient, hence a ratio of covariance to variance. From (4):

$$\rho = \frac{\pi_2}{\pi_1}.$$

The sample analogue, the **Wald/IV estimator**, is:

$$\hat{\rho}_{\text{wald}} = \frac{\hat{\pi}_2^{\text{ols}}}{\hat{\pi}_1^{\text{ols}}}. \tag{8}$$

3 Two Stage Least Squares (2SLS)

A more general estimator for IV is **Two Stage Least Squares (2SLS)**. It is simple to implement and economically intuitive.

Consider the structural equation and first-stage equation:

$$Y_i = X_i'\alpha + \rho s_i + \eta_i, \tag{9}$$
$$s_i = X_i'\pi_{10} + \pi_{11} z_i + \xi_{1i}, \tag{10}$$

where X_i is a vector of exogenous control variables and z_i is the instrument.

Substitute (10) into (9):

$$\begin{aligned} Y_i &= X_i'\alpha + \rho(X_i'\pi_{10} + \pi_{11}z_i + \xi_{1i}) + \eta_i \\ &= X_i'\alpha + \rho(X_i'\pi_{10} + \pi_{11}z_i) + \xi_{2i}, \end{aligned} \tag{11}$$

where $\xi_{2i} = \rho\xi_{1i} + \eta_i$. The term $(X_i'\pi_{10} + \pi_{11}z_i)$ is the population projection of s_i onto z_i and X_i. Because $z_i \perp \eta_i$ and X_i is exogenous, this projected part is uncorrelated with ξ_{2i}. The endogeneity of s_i came from the component ξ_{1i} that is correlated with η_i; by replacing s_i with its predicted value from the first stage, we purge that problematic component.

The 2SLS procedure is exactly what its name suggests:

1. **First Stage:** Regress s_i on both z_i and X_i to obtain the predicted value:

$$\hat{s}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11}z_i.$$

2. **Second Stage:** Regress Y_i on the predicted value $\hat{s}_i$ and X_i:

$$Y_i = X_i'\alpha + \rho\hat{s}_i + \xi_{2i}'.$$

   The coefficient on $\hat{s}_i$ is the 2SLS estimate of ρ.

Several important points deserve emphasis:

1. **Include the same controls in both stages.** The control variables X_i must appear in both the first-stage and second-stage regressions.
2. **Never compute 2SLS by hand.** In Stata, use `ivregress 2sls` or `ivreg2`. If you manually run two OLS regressions, the second-stage standard errors will be wrong because they fail to account for the estimation uncertainty in $\hat{s}_i$.
3. **The first stage need not have a causal interpretation.** You can always run a regression without claiming causality. What matters is that the instrument predicts the endogenous variable. That said, in practice it is better if you have a reason to believe that z actually affects s, a strong theoretical story makes the first stage more credible.

There is a close connection between Wald and 2SLS estimators.

- The Wald estimator is only available when the number of endogenous variables equals the number of instruments (exactly identified).
- When the model is just-identified, the 2SLS estimator is algebraically identical to the Wald estimator.
- In general, 2SLS is more efficient, and it is the most efficient IV estimator under homoskedasticity.

4 IV with Heterogeneous Treatment Effects: LATE

So far, we have maintained the assumption of a constant treatment effect: every individual has the same causal effect ρ. In many applications, this is unrealistic. Some people benefit more from schooling than others; some patients respond more to a drug than others. What does the IV estimator identify when treatment effects are heterogeneous?

This question was answered in a landmark paper by Imbens and Angrist (1994). The answer is the **Local Average Treatment Effect (LATE)**.

4.1 Setting: The Vietnam Draft Lottery

To motivate the framework, consider the classic study by Angrist (1990) on the effect of military service on earnings. During the Vietnam War, young men in the United States were drafted into the army based on a lottery: each birthday was assigned a random draft lottery number, and men with numbers below a cutoff were likely to be drafted. The key variables are:

- Y_i: wage earnings,
- D_i: whether individual i served in the military (treatment),
- z_i: whether the draft lottery number is below the cutoff (instrument).

The draft number is a plausible instrument: it was randomly assigned (exogenous) and it affected the probability of military service (relevant), but it should not affect wages through any channel other than military service (exclusion restriction).

4.2 Potential Outcomes and Four Assumptions

We now define potential outcomes more carefully. Let $Y_i(d, z)$ denote the potential outcome (wage) for individual i given treatment status d and instrument value z. Let D_{1i} and D_{0i} denote the potential treatment status when the instrument takes value 1 and 0, respectively.

The LATE theorem requires four assumptions:

Four Assumptions for LATE

1. **Independence:**

$$\{Y_i(D_{1i}, 1),\ Y_i(D_{0i}, 0),\ D_{1i},\ D_{0i}\} \perp\!\!\!\perp z_i.$$

   The instrument is as good as randomly assigned—it is independent of both potential outcomes and potential treatment status (agent types).

2. **Exclusion Restriction:**

$$Y_i(d, 0) = Y_i(d, 1) \equiv Y_{di} \quad \text{for } d = 0, 1.$$

   The instrument can affect the outcome *only* through treatment. In the draft example, the lottery number affects future wages only by changing military service, not through any other channel (e.g., education decisions).

3. **Existence of First Stage:**

$$E[D_{1i} - D_{0i}] \neq 0.$$

   The instrument must shift the probability of treatment on average.

4. **Monotonicity:**

$$\forall\, i, \quad D_{1i} - D_{0i} \geq 0 \quad \text{(or vice versa)}.$$

   The instrument moves treatment in the same direction for everyone (or has no effect). There are no **defiers**—individuals who do the opposite of what the instrument encourages.

4.3 Compliance Types

The monotonicity assumption partitions the population into three types based on how individuals respond to the instrument:

- **Compliers** ($D_{1i} > D_{0i}$): People who change their treatment status because of the instrument. In the draft example, these are men who served in the military only because their lottery number was below the cutoff.
- **Always-takers** ($D_{1i} = D_{0i} = 1$): People who always take treatment regardless of the instrument. These are men who would serve voluntarily even without being drafted.
- **Never-takers** ($D_{1i} = D_{0i} = 0$): People who never take treatment regardless of the instrument. These are men who would not serve even if drafted (e.g., obtained exemptions).

Monotonicity rules out a fourth type, **defiers** ($D_{1i} < D_{0i}$), individuals who would serve when their number is above the cutoff but refuse when it is below. This assumption is often plausible in practice.

4.4 The LATE Theorem

LATE Theorem (Theorem 4.4.1 in Angrist and Pischke, 2009)

Under Assumptions 1–4, the IV (Wald) estimand identifies the average treatment effect for the complier subpopulation:

$$\frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[D_i \mid z_i = 1] - E[D_i \mid z_i = 0]} = E[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]. \tag{12}$$

Proof. Let A denote always-takers, C denote compliers, and N denote never-takers. Decompose the intention-to-treat (ITT) effect:

$$\begin{aligned}
&E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0] \\
&= P(A_i \mid z_i = 1)\, E[Y_{1i} \mid A_i, z_i = 1] + P(C_i \mid z_i = 1)\, E[Y_{1i} \mid C_i, z_i = 1] \\
&\quad + P(N_i \mid z_i = 1)\, E[Y_{0i} \mid N_i, z_i = 1] \\
&\quad - \big[P(A_i \mid z_i = 0)\, E[Y_{1i} \mid A_i, z_i = 0] + P(C_i \mid z_i = 0)\, E[Y_{0i} \mid C_i, z_i = 0] \\
&\quad + P(N_i \mid z_i = 0)\, E[Y_{0i} \mid N_i, z_i = 0]\big].
\end{aligned}$$

By the independence assumption, z_i is independent of compliance type and potential outcomes. Therefore:

- $P(A_i \mid z_i = 1) = P(A_i \mid z_i = 0) = P(A_i)$, and similarly for C and N.
- $E[Y_{di} \mid \text{type}, z_i = 1] = E[Y_{di} \mid \text{type}, z_i = 0]$ for each type.

The always-taker terms cancel (both equal $P(A_i)\, E[Y_{1i} \mid A_i]$). The never-taker terms cancel (both equal $P(N_i)\, E[Y_{0i} \mid N_i]$). Only the complier terms remain:

$$E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0] = P(C_i)\big(E[Y_{1i} \mid C_i] - E[Y_{0i} \mid C_i]\big).$$

Similarly, the denominator equals:

$$E[D_i \mid z_i = 1] - E[D_i \mid z_i = 0] = P(C_i).$$

Dividing:

$$\frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[D_i \mid z_i = 1] - E[D_i \mid z_i = 0]} = E[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}].$$

□

4.5 Interpreting LATE

LATE has several important properties:

- **LATE is internally valid:** it identifies a well-defined causal effect for a well-defined subpopulation (compliers).
- **LATE is policy-relevant:** compliers are precisely those whose behavior can be changed by the policy instrument. If a policymaker is considering expanding the draft, the LATE tells them the average effect on the marginal group that would be affected.
- **Monotonicity is crucial:** without it, the effects of compliers and defiers would be mixed together, and the IV estimand would not have a clear causal interpretation.

However, LATE also has important limitations:

- **External validity:** LATE is specific to the complier group induced by a particular instrument. If the policy changes, the complier group changes, and the

LATE may differ.

- **Multi-valued instruments and treatments:** When the instrument or treatment is not binary, the number of compliance types grows exponentially, making the traditional LATE framework difficult to apply. As Pinto (2015) discusses, we need new approaches, such as combining IV with choice models, to handle these cases. This will be the subject of Chapter 6.

5 Multiple IV and the GMM Framework

So far, we have considered settings with a single endogenous variable and a single instrument. In practice, researchers often have multiple endogenous variables and multiple instruments. The **Generalized Method of Moments (GMM)** provides a unified framework for understanding all IV-related estimators.

5.1 Moment Equation Models

Let $g_i(\beta)$ be a known $l \times 1$ function of a $k \times 1$ parameter vector β.

Definition 5.1 (Moment Equation Model)**.** A moment equation model is defined by the condition:

$$E[g_i(\beta)] = 0. \tag{13}$$

In this system, we have l known equations and k unknown parameters.

Example 5.2. The linear regression model is a moment equation model with $l = k$ and $g_i(\beta) = x_i(Y_i - x_i'\beta)$. The moment condition $E[x_i(Y_i - x_i'\beta)] = 0$ is just the population orthogonality condition for OLS.

Given that these l equations are not linearly dependent, the relationship between l and k determines the identification status:

- If $l = k$: the model is **just-identified**. There are exactly as many equations as unknowns.
- If $l > k$: the model is **over-identified**. There are more equations than unknowns.
- If $l < k$: the model is **under-identified**. There are fewer equations than unknowns, and β cannot be uniquely recovered.

5.2 Method of Moments and GMM

When $l = k$ (just-identified), we can solve the system of equations directly using the **Method of Moments Estimator (MME)**: set the sample analogue of the moment conditions to zero and solve for $\hat{\beta}$:

$$\bar{g}_n = \frac{1}{n} \sum_{i=1}^{n} g_i(\hat{\beta}) = 0. \tag{14}$$

OLS is a special case: $\bar{g}_n = \frac{1}{n} \sum_{i=1}^{n} x_i(Y_i - x_i'\hat{\beta}) = 0$ gives exactly the OLS normal equations.

When $l > k$ (over-identified), we have more equations than unknowns, so in small sample, we generally cannot set all moment conditions to zero simultaneously. Here is an important conceptual difference we need to clarify. Although we assume that in population, we have the moment conditions $E[g_i(\beta)] = 0$, its sample analogue does not necessarily hold. Because there is uncertainty when we draw samples from the population.

Instead, we minimize the *distance* between the sample moment vector and zero:

$$J(\beta) = n\, \bar{g}_n(\beta)'\, W\, \bar{g}_n(\beta), \tag{15}$$

where W is a positive-definite **weighting matrix**. The **GMM estimator** is:

$$\hat{\beta}_{\text{gmm}} = \arg\min_{\beta}\; J(\beta). \tag{16}$$

The function J measures the weighted Euclidean distance between $\bar{g}_n$ and zero. The MME (and thus OLS) is a special case of GMM when $l = k$.

5.3 Linear GMM

In the linear IV model, let X_i denote the (potentially endogenous) regressors and Z_i the instruments. The moment conditions are:

$$E[g_i(\beta)] = E[Z_i(Y_i - X_i'\beta)] = 0. \tag{17}$$

Stacking over the sample, the GMM estimator minimizes:

$$\hat{\beta}_{\text{gmm}} = \arg\min_{\beta} n \underbrace{(Z'Y - Z'X\beta)' W (Z'Y - Z'X\beta)}_{J(\beta)}.$$

Here the dimension for X is $n \times k$, for Z is $n \times l$, for Y is $n \times 1$.

Linear GMM Estimator (Theorem 13.1 in Hansen (2022))

For the over-identified linear IV model with l endogenous variables and k instruments:

$$\hat{\beta}_{\text{gmm}} = (X'ZWZ'X)^{-1}(X'ZWZ'Y). \tag{18}$$

This is a remarkably general formula. Many familiar estimators are special cases, depending on the choice of W and the relationship between X and Z.

5.4 Special Cases of the Linear GMM Estimator

Case 1: OLS. When $X = Z$ (no endogenous variables) and $W = \mathbf{I}$:

$$\begin{aligned}\hat{\beta}_{\text{gmm}} &= (X'X\mathbf{I}X'X)^{-1}(X'X\mathbf{I}X'Y) \\ &= (X'X)^{-1}(X'X)^{-1}(X'X)(X'Y) \\ &= (X'X)^{-1}X'Y = \hat{\beta}_{\text{ols}}.\end{aligned}$$

The second line follows because $X'X$ is a square matrix when $X = Z$ (so $l = k$). Similarly, with a different weighting matrix, GMM becomes WLS.

Case 2: Wald/IV Estimator. When $l = k$ (just-identified) and $W = \mathbf{I}$:

$$\begin{aligned}\hat{\beta}_{\text{gmm}} &= (X'Z\mathbf{I}Z'X)^{-1}(X'Z\mathbf{I}Z'Y) \\ &= (Z'X)^{-1}(X'Z)^{-1}(X'Z)(Z'Y) \\ &= (Z'X)^{-1}Z'Y.\end{aligned}$$

Since $l = k$, the matrix $X'Z$ is square and invertible. This is the Wald/IV estimator.

Case 3: 2SLS Estimator. Let $P_z = Z(Z'Z)^{-1}Z'$ be the projection matrix onto the column space of Z, so that the first-stage fitted values are $\hat{X} = P_zX$. Projection matrix P has an important property called *idempotent*, such that $P \cdot P = P$.

When $W = (Z'Z)^{-1}$:

$$\begin{aligned}\hat{\beta}_{\text{gmm}} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y) \\ &= (X'P_z X)^{-1}X'P_z Y \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y.\end{aligned}$$

This is exactly the 2SLS estimator: regress Y on $\hat{X}$, where $\hat{X}$ is the first-stage prediction.

Remark 5.3. The fact that OLS, Wald, and 2SLS are all special cases of GMM is not merely an algebraic curiosity. It means that the asymptotic theory for GMM (consistency, asymptotic normality, efficient weighting) applies to all of these estimators as special cases. Understanding GMM gives you a unified toolkit for inference in IV models.

5.5 Over-identification Test

When we have more instruments than endogenous variables ($l > k$), we can test whether the moment conditions hold, that is, whether the instruments are valid. The logic is intuitive: if all instruments are valid, the minimized value of J should be close to zero. The test result below builds on Hansen (2022).

Hansen's J-Test (Theorem 13.14 in Hansen (2022))

Under the null hypothesis that all moment conditions are satisfied, as $n \to \infty$:

$$J = J(\hat{\beta}_{\text{gmm}}) \xrightarrow{d} \chi^2_{l-k}. \tag{19}$$

For c satisfying $\alpha = 1 - G_{l-k}(c)$, we reject H_0 if $J > c$. The test has asymptotic size α.

Caution with the Over-identification Test

The J-test has several important caveats:

1. **You should *hope* for non-rejection.** A valid IV strategy implies that J should be small. Rejection means something is wrong with at least one of the moment conditions.

2. **Feasibility requires over-identification.** The test can only be conducted when $l > k$. In the just-identified case, $J = 0$ by construction, and there is nothing to test.
3. **Rejection has multiple interpretations.** The J-test is a specification test: rejection does not necessarily mean $E[g_i] \neq 0$. It could also reflect non-linearity or other forms of model misspecification.
4. **Non-rejection does not prove validity.** The J-test has power against specific alternatives, but it cannot detect all forms of instrument invalidity.
5. **Finite-sample size distortion.** The actual rejection rate in finite samples tends to be too large (the test over-rejects).

6 Oster Bound: Endogeneity without IV

Finding a valid instrument is often extremely difficult. In many empirical settings, no credible instrument is available. What can we do when we face an endogeneity problem but cannot find an IV?

Oster (2019) proposes a method, **Oster bound**, that uses the pattern of coefficient stability and changes in R^2 across specifications to assess the robustness of OLS estimates to omitted variable bias. It is important to be clear about what this method does and does not do.

6.1 Point Identification vs. Set Identification

Before presenting the method, it is useful to clarify two concepts:

Definition 6.1 (Point Identification)**.** A parameter β is **point-identified** if there is a unique mapping from the data distribution to the parameter value. No other parameter value could generate the same data.

Definition 6.2 (Set Identification)**.** A parameter β is **set-identified** if the data restrict β to lie within a set, but do not uniquely pin it down. No parameter value *outside* the identified set is consistent with the data.

IV provides point identification of ρ (under its assumptions). The Oster bound provides only *set identification*: it gives a range within which the true effect plausibly lies. It is not a method for point estimation. It provides suggestive evidence when nothing else is available.

6.2 The Intuition

The core intuition of the Oster bound is simple and elegant: **the relationship between the treatment variable and unobservable confounders can be partially recovered from the relationship between the treatment variable and observable controls.**

Consider the following thought experiment. You run two regressions:

1. A "short" regression of Y on X alone (no controls), obtaining coefficient $\mathring{\beta}$ and R-squared $\mathring{R}$.
2. An "intermediate" regression of Y on X and a set of observed controls ω, obtaining $\tilde{\beta}$ and $\tilde{R}$.

If there is large omitted variable bias, then when we add informative controls, the coefficient on X should change substantially and the R^2 should increase a lot. The Oster bound exploits this logic: if the coefficient is *stable* (changes only a little) when we add a *strong, informative* control (one that substantially increases R^2), this is evidence that omitted variable bias is small.

6.3 The Formal Setup

Suppose we are interested in the effect of X on Y. There are two sets of variables W_1 and W_2, both correlated with X and Y. The variable W_1 can be represented by some observed proxies ω, while W_2 is unobservable. The model is:

$$\begin{aligned} Y &= \beta X + \Psi\omega + W_2 + \epsilon, \\ W_1 &= \Psi\omega, \end{aligned} \tag{20}$$

where W_1 and W_2 are assumed to be orthogonal to each other.

Define the **proportional selection relationship** δ:

$$\delta \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}, \tag{21}$$

where $\sigma_{jX} = \mathrm{Cov}(W_j, X)$ and $\sigma_j^2 = \mathrm{Var}(W_j)$ for $j = 1, 2$. The parameter δ measures the relative importance of the observed controls W_1 versus the unobserved confounders W_2 in their relationship with the treatment X:

- When δ is large, the unobserved variable W_2 is relatively more important than the observed control W_1.
- When $\delta = 1$, the unobservable and observable are equally related to the treatment.

6.4 Three Regressions, Two Key Parameters

Denote the coefficient on X and the R-squared from three regressions:

1. **Short regression:** Y on $X \Rightarrow \mathring{\beta}$, $\mathring{R}$.
2. **Intermediate regression:** Y on X and $\omega \Rightarrow \tilde{\beta}$, $\tilde{R}$.
3. **Full regression:** Y on X, ω, and $W_2 \Rightarrow R_{\max}$.

We can compute $\mathring{\beta}$, $\mathring{R}$, $\tilde{\beta}$, and $\tilde{R}$ from the data. The two unknowns are δ (the proportional selection parameter) and $R_{\max}$ (the R-squared from the hypothetical regression that includes all confounders). Given $\mathring{R}$ and $\tilde{R}$, we can infer how much variation is explained by observed controls; $R_{\max}$ tells us the share explained by the additional unobserved control W_2.

6.5 Key Theoretical Results

Oster (2019) provides three propositions connecting δ, $R_{\max}$, and bias.

Proposition 6.3 (Proposition 1 in Oster (2019): Debiased Estimator under $\delta = 1$)**.** *When $\delta = 1$ (observables and unobservables are equally important), the bias-adjusted estimator is*

$$\beta^* = \tilde{\beta} - \underbrace{[\mathring{\beta} - \tilde{\beta}]\frac{R_{\max} - \tilde{R}}{\tilde{R} - \mathring{R}}}_{bias}, \tag{22}$$

and this debiased estimator is asymptotically consistent: $\beta^* \xrightarrow{p} \beta$.

The formula has an intuitive interpretation. The bias is positively related to the *coefficient change* $\mathring{\beta} - \tilde{\beta}$ (how much the coefficient moves when controls are added) and the ratio $\frac{R_{\max} - \tilde{R}}{\tilde{R} - \mathring{R}}$ (how much additional variation the unobservables explain, relative to what the observables explained).

Proposition 6.4 (Proposition 2 in Oster (2019))**.** *Given δ and $R_{\max}$, we can calculate the bias and find a debiased estimator. In some cases, there may be multiple solutions, requiring solution selection. In short: $\delta, R_{\max} \to$ bias$, \beta$.*

Proposition 6.5 (Proposition 3 in Oster (2019))**.** *Given $R_{\max}$ and any targeted value of the treatment effect β, we can find the value of δ that makes the bias zero. In short: $R_{\max}, \beta$, bias $= 0 \to \delta$.*

Of course, we never observe δ or $R_{\max}$ (since W_2 is unobservable). But these propositions give us a powerful tool for *bounding* β under assumptions about these parameters.

6.6 Implementation in Practice

There are two practical methods for using the Oster bound:

Method 1: How robust is our result? Assume a value for $R_{\max}$ and calculate the value of δ for which $\beta = 0$ based on Proposition 6.5. This tells us how important unobservables would need to be (relative to observables) to completely erase the estimated effect.

1. As a rule of thumb, set $R_{\max} = \min\{1,\ 1.3\tilde{R}\}$. The factor 1.3 is derived from letting 90% of RCT studies in top journals pass this test.
2. Set $\beta = 0$ and find the corresponding δ.
3. If $\delta > 1$, our results are robust: unobservables would need to be more important than observables to overturn the finding.

Method 2: What is a conservative bound? Assume conservative values for both $R_{\max}$ and δ, and calculate the debiased estimate β^* based on Proposition 6.4.

1. Set $R_{\max} = \min\{1,\ 1.3\tilde{R}\}$ and $\delta = 1$.
2. Calculate β^* using (22).
3. The conservative bound on the treatment effect is $[\tilde{\beta}, \beta^*]$.
4. If the interval does not contain zero, the result is robust.

Limitations of the Oster Bound

1. The Oster bound is a **last resort**—use it when no valid instrument or other identification strategy is available.
2. The choice of R_{max} and δ is inherently arbitrary. The bound provides a *sense* of robustness, not a definitive answer.
3. It can also be used as a robustness check alongside other methods.
4. An application in economics: Clark et al. (2021) use the Oster bound to assess the robustness of their estimates of the effect of online learning on student performance during the COVID-19 pandemic.

7 Conclusion

This chapter introduced the instrumental variables framework as the primary tool for addressing endogeneity in empirical economics. The key takeaways are:

1. **IV requires two assumptions:** the exclusion restriction (the instrument affects the outcome only through the endogenous variable) and the existence of a first stage (the instrument is correlated with the endogenous variable). Both must be carefully justified in any application.

2. **With heterogeneous treatment effects, single binary endogenous variable and instrument,** the IV/Wald estimator identifies the **Local Average Treatment Effect (LATE)**, the average effect for the complier subpopulation. LATE is internally valid and policy-relevant, but not necessarily externally valid.

3. **GMM is the general framework** for IV estimation. OLS, the Wald estimator, and 2SLS are all special cases of the linear GMM estimator, differing only in the weighting matrix and the relationship between the number of regressors and the number of instruments.

4. **Over-identification tests** (Hansen's J-test) allow us to assess whether the moment conditions hold, but they should be interpreted cautiously.

5. **When no valid IV is available,** the Oster bound provides a way to assess robustness to omitted variable bias. It is based on the idea that the relationship between treatment and unobservables can be inferred from the relationship between treatment and observables. However, it provides only set identification

and relies on somewhat arbitrary parameter choices.

In the next chapter, we will relax the remaining simplifying assumptions: we will move beyond binary variables and binary instruments, and consider how to combine IV with choice models to handle more complex settings.

8 Application of IV: Angrist (1990)

We have invoked the Vietnam draft lottery several times in this chapter as the canonical motivating example for IV with heterogeneous treatment effects. The empirical study that established this design as a textbook reference is Angrist (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," published in the *American Economic Review*. The research question is simple but consequential: did military service in the Vietnam era cause a long term reduction in civilian earnings? The naive answer, comparing earnings of veterans and nonveterans, is contaminated by selection into the military: men who enlisted or were drafted differ from those who did not in many ways that also affect earnings (education, health, family background, expected labor market prospects). Without a credible source of exogenous variation, the OLS comparison cannot be given a causal interpretation.

The methodological contribution of the paper is to use the Vietnam era draft lottery as an instrument for veteran status. From 1970 to 1972, the United States Selective Service assigned each draft age man a random sequence number based on his date of birth. Men whose number fell below an administratively determined ceiling were called for induction, while those above the ceiling were exempt. The lottery number is plausibly random with respect to potential earnings and affects civilian earnings only through its effect on the probability of military service, so it satisfies the relevance, exclusion, and independence conditions discussed in Section 2. Angrist constructs draft eligibility as the binary instrument and uses it together with Social Security administrative earnings records, computing Wald style estimates by birth cohort and confirming the results with 2SLS that absorbs cohort fixed effects. Under the heterogeneous treatment effects framework of Section 4, the estimand is a LATE: the average earnings loss for compliers, that is, men who served because they were draft eligible but would not have served otherwise.

The main empirical finding is that Vietnam era military service caused a sizable and persistent decline in the earnings of white veterans, on the order of fifteen percent of nonveteran earnings in the early 1980s, roughly a decade after discharge. For nonwhite men the estimated effect is much smaller and statistically indistinguishable from zero. Beyond the substantive finding, the paper has been hugely influential as a methodological template: it shows in a single setting how to combine a randomized policy lever with administrative data and the IV/LATE machinery to recover a credible causal effect. I strongly encourage students to read the original article to see how the formal framework of this chapter plays out in a real empirical study.

A Homework Problems

1. **IV Identification with Covariates.**

   Consider the model $Y_i = \alpha + \beta_1 X_i + \beta_2 D_i + \epsilon_i$, where D_i is endogenous ($\text{Cov}(D_i, \epsilon_i) \neq 0$) and X_i is exogenous. Suppose z_i is a valid instrument: $\text{Cov}(z_i, \epsilon_i) = 0$ and $\text{Cov}(z_i, D_i) \neq 0$.

   (a) Show that the IV estimator of β_2 can be written as a ratio of covariances involving the residuals from projecting Y_i and D_i on X_i.

   (b) Explain intuitively why we need to "partial out" X_i before applying IV.

2. **LATE with Compliance Data.**

   Suppose 1000 individuals are randomly assigned to treatment ($z_i = 1$, 500 people) or control ($z_i = 0$, 500 people). Among those assigned to treatment, 300 actually take treatment ($D_i = 1$) and 200 do not. Among those assigned to control, 50 take treatment and 450 do not. Average outcomes: $\bar{Y}(z = 1) = 10$, $\bar{Y}(z = 0) = 8$.

   (a) Calculate the Wald/LATE estimator.

   (b) What fraction of the population are compliers? Always-takers? Never-takers?

   (c) Under what assumption can we rule out defiers?

3. **Oster Bound Calculation.**

   A researcher estimates the effect of class size on test scores. The short regression (no controls) gives $\mathring{\beta} = -2.5$ with $\mathring{R} = 0.04$. Adding school-level controls gives

$\tilde{\beta} = -1.8$ with $\tilde{R} = 0.22$.

(a) Using $R_{\max} = 1.3 \times \tilde{R}$ and $\delta = 1$, compute the bias-adjusted estimate β^*.

(b) Does the interval $[\tilde{\beta}, \beta^*]$ contain zero? What does this tell us about robustness?

(c) Using the same $R_{\max}$, find the value of δ that would make $\beta^* = 0$. Is the result robust?

Practice Example: IV Estimation of Returns to Schooling

In this practice example, we simulate data to illustrate the endogeneity problem and the IV solution. The setting mirrors the classic returns-to-schooling question.

Data Generating Process. We simulate $N = 2{,}000$ individuals. Each individual has an unobserved ability $A_i \sim N(0, 1)$. A binary instrument z_i (e.g., proximity to a college) is randomly assigned with $P(z_i = 1) = 0.5$, independent of ability. Schooling is determined by:

$$s_i = 12 + 2z_i + 1.5A_i + u_i, \quad u_i \sim N(0, 1),$$

so the instrument shifts schooling by 2 years on average. Wages are:

$$Y_i = 20 + 3s_i + 2A_i + \epsilon_i, \quad \epsilon_i \sim N(0, 5),$$

so the true causal effect of one additional year of schooling is $\rho = 3$. Because ability A_i enters both equations, OLS on the short regression will be upward biased.

Analysis. The Stata do-file (located at `src/chapter 5/ch5_iv.do`) performs the following steps:

1. Simulate the data and save as `data/chapter 5/ch5_iv.dta`.
2. Run OLS of Y on s (biased estimate).
3. Run the first-stage regression of s on z.
4. Run the reduced-form regression of Y on z.
5. Compute the Wald estimator as the ratio of reduced-form to first-stage coefficients.
6. Run 2SLS using `ivregress 2sls`.
7. Compare OLS and IV estimates to the true effect $\rho = 3$.

Expected Results. The OLS estimate should be substantially above 3 (upward biased due to ability). The IV/2SLS estimate should be close to 3. The first-stage F-statistic should be large, confirming instrument relevance.

References

Angrist, Joshua D. 1990. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.” *American Economic Review*, 80(3): 313–336.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

Clark, Andrew E., Huifu Nong, Hongjia Zhu, and Rong Zhu. 2021. “Compensating for Academic Loss: Online Learning and Student Performance during the COVID-19 Pandemic.” *China Economic Review*, 68: 101629.

Hansen, Bruce. 2022. *Econometrics*. Princeton University Press.

Imbens, Guido W. and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica*, 62(2): 467–475.

Oster, Emily. 2019. “Unobservable Selection and Coefficient Stability: Theory and Evidence.” *Journal of Business & Economic Statistics*, 37(2): 187–204.

Pinto, Rodrigo. 2015. “Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity.” Unpublished Ph.D. Thesis, University of Chicago, Department of Economics.