Econometrics 2

Instrumental Variables Regression – Part I

Lasha Chochua

2026

Introduction

Introduction to Instrumental Variables

- Sometimes, the error term in a regression is **correlated with the regressor**.
- This can happen due to:
  - **Omitted variables**
  - **Errors in variables**
  - **Simultaneous causality**
- In these cases, **OLS estimates become biased and inconsistent**.
- Adding the omitted variable to the regression can fix the bias – but only if we have data on it.
- If we do not, or if **causality runs in both directions**, we need a different method.
- **Instrumental Variables (IV) regression** provides a solution.

What Is IV Regression?

- IV regression helps us **estimate causal effects** even when X is correlated with the error term u.
- Think of X as having two parts:
  - One part **correlated with** u – causes bias.
  - One part **uncorrelated with** u – safe for estimation.
- If we can isolate the second part, we can remove the bias.
- This is done using **instrumental variables**, or **instruments**.

What Do Instruments Do?

- Instruments are **extra variables** that:
  - Explain variation in X,
  - But are **uncorrelated with the error term** u.
- These instruments act as tools to **extract the clean part of** X.
- Using them, we can estimate regression coefficients **consistently**.

Lecture Overview

- The lecture covers:
  - Why IV regression works.
  - What makes a **valid instrument**.
  - How to **implement and interpret** IV regression.
- Main method: **Two Stage Least Squares (2SLS)**.
- Key challenges:
  - Finding valid instruments.
  - Estimating real-world models, like demand for cigarettes.
  - Understanding **where valid instruments come from**.

IV with One Regressor and One Instrument

IV with One Regressor and One Instrument

- Suppose we have one regressor X, which is correlated with the error term u.
- Then, OLS will give **biased and inconsistent estimates**.
- This bias may come from:
  - Omitted variables
  - Measurement error
  - Simultaneous causality
- If we can find a valid instrument Z, we can still estimate the effect of X on Y.

IV Model and Assumptions

- Let β_1 be the causal effect of X on Y. The model is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n \tag{1}$$

- u_i captures unobserved factors affecting Y_i.
- If X_i and u_i are correlated, OLS is inconsistent.
- IV regression uses a variable Z that is **correlated with** X but **uncorrelated with** u.

Endogeneity and Exogeneity

- A variable **correlated with the error term** u is **endogenous**.
- A variable **uncorrelated with** u is **exogenous**.
- Endogeneity often arises in models where:
  - Variables influence each other (simultaneous equations),
  - Or where omitted variables affect both X and Y.
- Example: Student–teacher ratio and test scores may affect each other.

Valid Instrument Conditions

Conditions for a Valid Instrument Z

1. **Relevance**: $\text{corr}(Z_i, X_i) \neq 0$
2. **Exogeneity**: $\text{corr}(Z_i, u_i) = 0$

- Relevance ensures Z is related to X.
- Exogeneity ensures Z is uncorrelated with the error term.
- If both hold, Z can isolate exogenous variation in X.
- This allows consistent estimation of the causal effect β_1.

Two Stage Least Squares

Two Stage Least Squares (TSLS)

- If Z is a valid instrument, we can estimate β_1 using **Two Stage Least Squares (TSLS)**.
- TSLS works in two steps:
  - **First stage**: Regress X_i on Z_i to isolate the exogenous part of X_i.
  - **Second stage**: Regress Y_i on the predicted values $\hat{X}_i$.

TSLS – First Stage

- We start by regressing X_i on Z_i:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \tag{2}$$

- π_0 is the intercept, π_1 is the slope, and v_i is the error term.
- This regression splits X_i into two parts:
  - $\pi_0 + \pi_1 Z_i$: the part of X_i predicted by Z_i
  - v_i: the remaining part, possibly correlated with u_i
- Since Z_i is exogenous, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i.

TSLS – First Stage (Cont.)

- In practice, we do not know π_0 and π_1.
- We estimate them using OLS:

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

- $\hat{X}_i$ is the **fitted value** of X_i from the first stage.
- It captures only the variation in X_i that is explained by Z_i.

TSLS – Second Stage

- Regress Y_i on the predicted value $\hat{X}_i$ using OLS:

$$Y_i = \beta_0^{TSLS} + \beta_1^{TSLS} \hat{X}_i + \text{error}$$

- The TSLS estimator β_1^{TSLS} gives a **consistent estimate** of the causal effect of X on Y.

Why Does IV Regression Work?

Why Does IV Regression Work?

- IV regression solves the problem of correlation between X_i and u_i.
- This is shown through **two examples**.
- The first example is based on a historical case by **Philip Wright**.

Example 1: Philip Wright's Problem

- In the 1920s, Philip Wright studied **tariffs on imported goods**.
- He wanted to estimate **supply and demand elasticities** for products like butter.
- He needed to estimate how **price affects demand** for butter.
- This leads to the demand equation:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i \tag{3}$$

- Q_i^{butter}: quantity of butter consumed
- P_i^{butter}: price of butter
- u_i: factors affecting demand, like income and tastes

Why OLS Fails Here

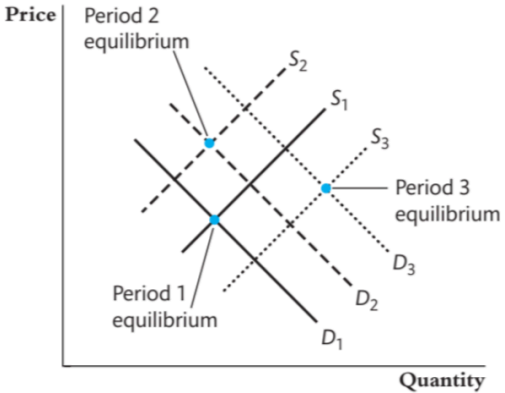
- Wright had data from 1912–1922 on:
  - Quantity of butter consumed
  - Price of butter
- But: price is **determined by both supply and demand**.
- This means $\ln(P_i^{butter})$ is **correlated with** u_i, causing endogeneity.
- OLS gives **biased estimates** of β_1.

Graphical Insight from Supply and Demand

- Imagine three years with different supply and demand curves:
  - Year 1: D_1, S_1 $\rightarrow$ first equilibrium
  - Year 2: D_2, S_2 $\rightarrow$ new equilibrium (demand $\uparrow$, supply $\downarrow$)
  - Year 3: D_3, S_3 $\rightarrow$ another shift
- Price and quantity keep changing due to **both** curves moving.
- A scatterplot of price and quantity won't reveal **true demand or supply** curves.
- OLS cannot recover the demand slope in such a setup.

Equilibrium Determination

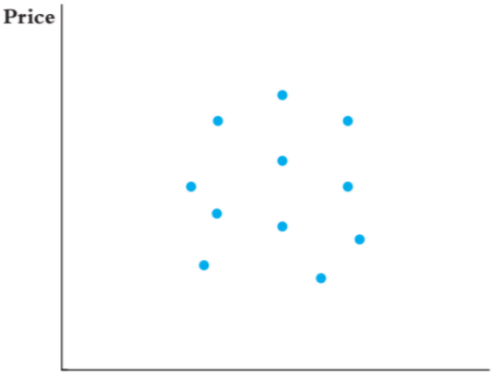
(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the third period is the intersection of D_3 and S_3.



(a) Demand and supply in three time periods

A Scatterplot of Prices and Quantities

(b) This scatterplot shows equilibrium price and quantity in 11 different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?



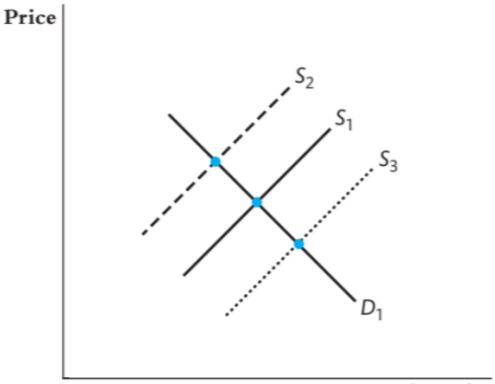
(b) Equilibrium price and quantity for 11 time periods

The Instrumental Variable Solution

- Wright's insight: use a variable that **shifts supply but not demand**.
- Example: **rainfall in dairy regions**
  - Less rainfall $\rightarrow$ less grazing $\rightarrow$ less butter $\rightarrow$ supply shifts left
  - Demand remains stable
- This satisfies both IV conditions:
  - **Relevance**: rainfall affects supply and thus price.
  - **Exogeneity**: rainfall does not directly affect demand $\Rightarrow$ uncorrelated with u_i.
- Using rainfall as an instrument gives a **valid estimate** of demand elasticity.

IV in Action

(c) When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1, the equilibrium prices and quantities trace out the demand curve.



(c) Equilibrium price and quantity when only the supply curve shifts

From the Figure to the Formula

- When rainfall shifts supply but **not** demand, each year traces a point **along the same demand curve**.
- Across years, all price variation is driven by Z_i (rainfall) – not by shifts in demand.
- Therefore:
  - $\text{cov}(Z_i, Y_i)$: how much log-quantity co-moves with rainfall
  - $\text{cov}(Z_i, X_i)$: how much log-price co-moves with rainfall
- Their ratio picks up movement **along the demand curve only**:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \approx \frac{\text{cov}(Z_i,\, \ln Q^{butter})}{\text{cov}(Z_i,\, \ln P^{butter})} = \beta_1^{demand}$$

Note

The figure is the formula made visible: the instrument uses supply shifts to trace the demand curve.

Example 2: Class Size and Test Scores

- Estimating the effect of **class size** on **student test scores** is difficult.
- Even after controlling for student and school characteristics, **omitted variable bias** may remain.
  - Example: unmeasured learning opportunities outside school
  - Or unmeasured teacher quality
- If we cannot observe or control for such factors, OLS estimates will be biased.

The IV Strategy

- IV regression can help solve this problem.
- Consider a **hypothetical example** involving California schools:
  - A **summer earthquake** damages some school buildings.
  - Schools near the epicenter are forced to **double up classes**, increasing class size.
- So, **distance to the epicenter** is correlated with class size:
  - Schools closer $\rightarrow$ bigger class sizes
  - Schools farther $\rightarrow$ smaller class sizes

Relevance and Exogeneity

- Distance to the epicenter is:
  - **Relevant**: correlated with class size
  - **Exogenous**: if it does **not** affect student performance **directly**, it is uncorrelated with u_i
- For example, distance to the epicenter does not affect:
  - Whether students are learning English
  - Family background
  - Teaching quality

A Valid Instrument

- Distance to the epicenter can serve as a **valid instrument**.
- It allows us to estimate the causal effect of class size on test scores.
- This is an example of how IV regression **fixes omitted variable bias**.

Sampling Distribution of the TSLS Estimator

Sampling Distribution of the TSLS Estimator

- In small samples, the exact distribution of the TSLS estimator is complex.
- But in large samples, it is **simple and useful**:
  - The TSLS estimator is **consistent**.
  - It is **normally distributed** in large samples.

Formula for the TSLS Estimator

- When there is one regressor X and one instrument Z, the TSLS formula is simple.
- Let:
  - s_{ZY} = sample covariance between Z and Y
  - s_{ZX} = sample covariance between Z and X
- Then, the TSLS estimator of β_1 is:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \tag{4}$$

- It is the **ratio of the covariance** of Z with Y to the **covariance** of Z with X.

TSLS Consistency – The Derivation

- Start from model (1) and take the covariance of both sides with Z_i:

$$\text{cov}(Z_i, Y_i) = \text{cov}(Z_i,\ \beta_0 + \beta_1 X_i + u_i)$$

- Expand using linearity of covariance ($\text{cov}(Z_i, \beta_0) = 0$):

$$\text{cov}(Z_i, Y_i) = \beta_1 \,\text{cov}(Z_i, X_i) + \text{cov}(Z_i, u_i) \tag{5}$$

- Apply the two instrument conditions:
  - **Exogeneity**: $\text{cov}(Z_i, u_i) = 0 \Rightarrow$ contaminating term vanishes.
  - **Relevance**: $\text{cov}(Z_i, X_i) \neq 0 \Rightarrow$ we can divide safely.
- Solving (5) for β_1:

$$\beta_1 = \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)} \tag{6}$$

Why Exogeneity Is the Key Assumption

- Equation (5) shows exactly why endogeneity breaks OLS but not IV:

OLS	TSLS
Identifies β_1 via $\text{cov}(X_i, Y_i)$	Identifies β_1 via $\text{cov}(Z_i, Y_i)$
If $\text{cov}(X_i, u_i) \neq 0$: extra term remains $\Rightarrow$ biased	If $\text{cov}(Z_i, u_i) = 0$: extra term vanishes $\Rightarrow$ consistent

- Exogeneity of Z is what **kills the bias** – it is the single most important assumption in IV.

TSLS Consistency

- From the sample covariances:

$$s_{ZY} \xrightarrow{p} \text{cov}(Z_i, Y_i), \quad s_{ZX} \xrightarrow{p} \text{cov}(Z_i, X_i)$$

- Then the TSLS estimator as defined in (4):

$$\widehat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)} = \beta_1 \tag{7}$$

- So, $\widehat{\beta}_1^{TSLS}$ is **consistent**.

Derivation of Asymptotic Variance of TSLS

- Consider the model in (1). The TSLS estimator is:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{\frac{1}{n}\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum(Z_i - \bar{Z})(X_i - \bar{X})}$$

- As $n \to \infty$, we approximate:

$$\hat{\beta}_1^{TSLS} \approx \frac{\frac{1}{n}\sum(Z_i - \mu_Z)Y_i}{\frac{1}{n}\sum(Z_i - \mu_Z)X_i}$$

- Substitute $Y_i = \beta_1 X_i + u_i$:

$$\hat{\beta}_1^{TSLS} \approx \beta_1 + \frac{\frac{1}{n}\sum(Z_i - \mu_Z)u_i}{\frac{1}{n}\sum(Z_i - \mu_Z)X_i}$$

Asymptotic Distribution (I)

- Multiply both sides by $\sqrt{n}$ to get a non-degenerate limit:

$$\sqrt{n}\left(\hat{\beta}_1^{TSLS} - \beta_1\right) = \frac{\frac{1}{\sqrt{n}}\sum(Z_i - \mu_Z)u_i}{\frac{1}{n}\sum(Z_i - \mu_Z)X_i}$$

- **Numerator:** by CLT, $\frac{1}{\sqrt{n}}\sum(Z_i - \mu_Z)u_i \xrightarrow{d} N(0,\ \text{var}[(Z_i - \mu_Z)u_i])$
- **Denominator:** by LLN, $\frac{1}{n}\sum(Z_i - \mu_Z)X_i \xrightarrow{p} \text{cov}(Z_i, X_i)$
- So, by Slutsky's theorem:

$$\sqrt{n}\left(\hat{\beta}_1^{TSLS} - \beta_1\right) \xrightarrow{d} N\left(0,\ \frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}\right)$$

Asymptotic Distribution (II)

- Therefore, the **asymptotic variance** of $\hat{\beta}_1^{TSLS}$ is:

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n} \cdot \frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2} \tag{8}$$

Inference with TSLS Estimator

- The asymptotic variance of $\hat{\beta}_1^{TSLS}$ from (8) is:

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n} \cdot \frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}$$

- In practice, we estimate the variance and covariance terms from data.
- The **standard error** of $\hat{\beta}_1^{TSLS}$ is the square root of this estimate.
- **Two important implications:**
  - TSLS is **always less precise than OLS** when X is exogenous – $[\text{cov}(Z_i, X_i)]^2 \leq [\text{var}(X_i)]^2$ since Z explains only part of X.
  - When $\text{cov}(Z_i, X_i) \approx 0$ (weak instrument), the denominator collapses $\Rightarrow$ variance explodes $\Rightarrow$ inference breaks down. (More on this in Part II.)

Conducting Hypothesis Tests

- Since $\hat{\beta}_1^{TSLS}$ is **approximately normal** in large samples:
  - We can use **t-tests** for inference.
- A **95% confidence interval** for β_1 is:

$$\hat{\beta}_1^{TSLS} \pm 1.96 \cdot SE(\hat{\beta}_1^{TSLS})$$

- These steps are automatically performed by standard TSLS routines in econometric software.

Application: Cigarette Demand Elasticity

Application: Cigarette Demand Elasticity

- Following Philip Wright's idea, IV methods are used to estimate **important elasticities**.
- Example: **elasticity of demand for cigarettes**.
- Similar methods are used in:
  - Health economics (e.g., effect of spending on mortality)
  - Public finance

Why Use TSLS for Cigarettes?

- OLS regression of log quantity on log price is biased due to **simultaneity**.
- Price and quantity are both influenced by demand and supply.
- TSLS solves this by using an **instrument**: cigarette **sales tax**.
- Instrument validity check:
  - **Relevance**: Higher sales tax raises price $\rightarrow$ satisfied.
  - **Exogeneity**: Tax rates set by public finance decisions, not demand $\rightarrow$ plausible, but requires judgment.

Step 1 – First Stage

- Regress log price on the instrument $SalesTax_i$:

$$\ln(\widehat{P_i^{cigarettes}}) = 4.62 + 0.031 \cdot SalesTax_i \quad (\mathrm{R}^2 = 47\%) \tag{9}$$
$$(0.03) \quad (0.005)$$

- The coefficient on $SalesTax_i$ is positive and highly significant.
- $R^2 = 47\%$ indicates **strong first-stage fit** – the instrument is relevant.

Step 2 – Second Stage

- Replace $\ln(P_i^{cigarettes})$ with fitted values $\ln(\widehat{P_i^{cigarettes}})$ and estimate by OLS:

$$\ln(\widehat{Q_i^{cigarettes}}) = 9.72 - 1.08 \ln(P_i^{cigarettes}) \quad (10)$$
$$(1.53) \quad (0.32)$$

- Interpretation: A **1% increase in price** leads to a **1.08% decrease** in quantity demanded.
- Demand is **price-elastic** – a 1% price rise reduces consumption by more than 1%.

A Problem – Income as Confounder

- The estimate may be misleading: **income** is a likely omitted variable.
  - Richer states may set higher taxes **and** consume less – for unrelated reasons.
  - If cov$(SalesTax_i,\, u_i) \neq 0$ because of income, exogeneity fails.
- Solution: **include** $\ln(Inc_i)$ **as a control variable** W_i.
  - Once we condition on income, $SalesTax_i$ should be uncorrelated with u_i.
  - This extends the model to the general IV framework.

Step 3 – Adding Income as a Control

- Model with income as an included exogenous variable:
  - Endogenous: $\ln(P_i^{cigarettes})$, ; Instrument: $SalesTax_i$, ; Control: $\ln(Inc_i)$

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14}\ln(P_i^{cigarettes}) + \underset{(0.31)}{0.21}\ln(Inc_i) \tag{11}$$

- Price elasticity rises slightly in magnitude: $-1.08 \rightarrow -1.14$.
- Income elasticity is positive $(+0.21)$ – higher income raises cigarette demand.

Step 4 – Adding a Second Instrument

- States also levy cigarette-specific taxes ($CigTax_i$), raising after-tax prices.
- If $CigTax_i$ is uncorrelated with u_i, it is a second valid instrument.
- Now we have:
  - 2 instruments: $SalesTax_i$, $CigTax_i$
  - 1 endogenous regressor: $\ln(P_i^{cigarettes})$
  - $\Rightarrow$ **Overidentified** ($m = 2 > k = 1$)

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28}\ln(P_i^{cigarettes}) + \underset{(0.25)}{0.28}\ln(Inc_i) \tag{12}$$

Step 5 – Comparing the Estimates

Specification	Price elasticity	SE	Instruments
1 instrument, no income	-1.08	0.32	$SalesTax$
1 instrument, with income	-1.14	0.37	$SalesTax$
2 instruments, with income	-1.28	0.25	$SalesTax, CigTax$

- Adding income changes the point estimate slightly – income was a confounder.
- Adding the second instrument **reduces SE from 0.37 to 0.25** – more relevant variation $\Rightarrow$ more precise.
- Next: Are both instruments truly exogenous? $\rightarrow$ Part II (overidentification test).

The General IV Regression Model

The General IV Regression Model

- The general IV model involves four types of variables:
  - Y: Dependent variable
  - X: Endogenous regressors (correlated with error)
  - W: Included exogenous variables (e.g., controls)
  - Z: Instrumental variables (instruments for X, excluded exogenous variables)
- We can have multiple X's, W's, and Z's.
- **Goal**: Use Z to isolate variation in X that is uncorrelated with the error term in Y.

The General IV Regression Model (Cont.)

The General IV Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \quad i = 1, \ldots, n \quad (13)$$

- Y_i: Dependent variable
- $X_{1i}, \ldots, X_{ki}$: k endogenous regressors (correlated with u_i)
- $W_{1i}, \ldots, W_{ri}$: r included exogenous regressors (uncorrelated with u_i)
- $Z_{1i}, \ldots, Z_{mi}$: m instrumental variables
- u_i: Error term (omitted variables or measurement error)

Identification Conditions

- For IV regression to work:
  - We need **at least as many instruments (Z's) as endogenous regressors (X's)**.
- Terminology:
  - **Exactly identified**: $m = k$ (number of instruments = number of endogenous regressors)
  - **Overidentified**: $m > k$
  - **Underidentified**: $m < k$ (cannot estimate the model)
- Identification is necessary for computing the TSLS estimator.

Included Exogenous and Control Variables

- W can be:
  - **Exogenous variables**: satisfy $E(u_i \mid W_i) = 0$
  - **Control variables**: included to remove correlation between Z and u_i
- Example:
  - If **sales tax** is correlated with **income**, and income affects cigarette demand,
    - then omitting income leads to **correlation between instrument and error**.
    - Including income (as W) removes this problem.

Role of Control Variables in IV

- If W is an effective control, then including it makes instrument Z exogenous:
  - TSLS estimator of β_1 becomes consistent.
- If W is correlated with u, then:
  - TSLS coefficient on W is biased.
  - Coefficient on X may also lose its causal interpretation.
- This mirrors the logic of **controls in OLS regression**.

Control Variables in IV Regression

Condition for Valid Control Variables in IV Regression

To be an effective control variable in IV regression, W must satisfy:

$$E(u_i \mid Z_i, W_i) = E(u_i \mid W_i)$$

This ensures that, once we control for W_i, the instrument Z_i is uncorrelated with the error u_i.

- This condition is called **conditional mean independence**.
- It means: after controlling for W_i, the instrument Z_i provides exogenous variation.
- If W_i is omitted and is correlated with both Z_i and u_i, the IV estimate is biased.

Two Special Cases

- **When W_i is exogenous** (e.g., income is included and uncorrelated with u_i):

$$E(u_i \mid W_i) = 0$$

- In this case, the IV exogeneity condition simplifies to:

$$E(u_i \mid Z_i, W_i) = 0$$

- **If W_i is a control variable** (possibly correlated with u_i), then:
  - We only require that Z_i is exogenous **conditional on** W_i.
  - This is:

$$E(u_i \mid Z_i, W_i) = E(u_i \mid W_i)$$

TSLS in the General IV Model

- When there is a single endogenous regressor X and additional included exogenous variables $W_1, \ldots, W_r$, the structural equation of interest is:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i \tag{14}$$

- X_i may be correlated with the error term u_i
- $W_{1i}, \ldots, W_{ri}$ are exogenous and uncorrelated with u_i

First Stage – Reduced Form for X_i

- The first-stage regression relates X_i to all available exogenous variables (instruments and controls):

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i \tag{15}$$

- This is sometimes called the **reduced form** for X_i
- It captures the variation in X_i that is exogenous

Example: Effect of Education on Earnings

- Goal: Estimate the causal effect of education (X_i) on wages (Y_i).
- Problem: Education may be **endogenous** (e.g., due to unobserved ability).

- People living closer to college tend to obtain more education.
- Z_i is a valid instrument **only if** it affects wages **only through** education.

- Z_i might be correlated with **family background**:
  - Wealthier families tend to live near colleges.
  - Family wealth (W_i) also affects wages directly.
- If we omit W_i:
  - The variation in education caused by Z_i may still reflect **endogenous influences**.
  - $\Rightarrow$ Instrument is **not exogenous**.

- Controlling for W_i **purges** Z_i of endogenous variation.
- The fitted values $\hat{X}_i$ now reflect only the **exogenous variation** in education.

Second Stage and TSLS Estimation

- In the first stage, estimate (15) by OLS and compute predicted values $\hat{X}_1, \ldots, \hat{X}_n$
- In the second stage, estimate (14) by OLS, replacing X_i with $\hat{X}_i$
- Final regression:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + \text{error}$$

- The resulting estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{1+r}$ form the **TSLS estimator**

Extension to Multiple Endogenous Regressors

- When there are multiple endogenous regressors $X_{1i}, \ldots, X_{ki}$, TSLS proceeds similarly to the single-regressor case, but with one key change:
- Each endogenous regressor requires its **own first-stage regression**.
- Each first-stage regression has the same form as (15):

$$X_{\ell i} = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i$$

- Where $\ell = 1, \ldots, k$, and $X_{\ell i}$ is each endogenous regressor.
- All instruments (Z's) and included exogenous variables (W's) are used as regressors.

Second Stage of TSLS

- After obtaining $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$ from the first stage,
- Replace all $X_{1i}, \ldots, X_{ki}$ in the structural equation with their predicted values.
- Then using **OLS** estimate (13):

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \cdots + \beta_k \hat{X}_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

- The resulting TSLS estimator is $\hat{\beta}_0, \ldots, \hat{\beta}_{k+r}$.
- In practice, both stages are handled automatically by TSLS commands in standard econometric software.

Instrument Relevance and Exogeneity in the General IV Model

Instrument Relevance and Exogeneity in the General IV Model

- The conditions for instrument relevance and exogeneity extend naturally to the general IV model, but with some important adjustments.

Instrument Relevance

- With one endogenous regressor X and multiple instruments (Z's), we require:
  - At least one Z must be correlated with X, **conditional on** W.
  - This ensures that the instrument provides useful variation in X after accounting for included exogenous variables.
- With multiple endogenous regressors:
  - We must avoid **perfect multicollinearity** in the second-stage regression.
  - Instruments must contain enough **independent variation** to isolate the exogenous part of each endogenous regressor.

Instrument Exogeneity

- The general requirement is:
  - Each instrument must be **uncorrelated** with the error term u_i.

$$\text{cov}(Z_{ji}, u_i) = 0 \quad \text{for all } j = 1, \ldots, m$$

- This ensures that instruments affect Y only through the endogenous regressors and not directly.
- These conditions are formalized in the next slide.

The Two Conditions for Valid Instruments

Two Conditions for Valid Instruments $Z_{1i}, \ldots, Z_{mi}$

1. **Instrument Relevance**
   - *In general*: Let $\hat{X}^*_{1i}$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z's) and included exogenous variables (W's).
   - Let "1" denote a constant regressor. Then the vector $(\hat{X}^*_{1i}, \ldots, \hat{X}^*_{ki}, W_{1i}, \ldots, W_{ri}, 1)$ must not be perfectly multicollinear.
   - *If there is only one* X: at least one instrument Z must have a non-zero coefficient in the regression of X on the Z's and W's.
2. **Instrument Exogeneity**
   - Instruments must be uncorrelated with the error term:

$$\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$$

IV Regression Assumptions and Sampling Distribution

IV Regression Assumptions and Sampling Distribution of TSLS

- Under the IV regression assumptions, the TSLS estimator is:
  - **Consistent**, and
  - Has a **sampling distribution** that is approximately **normal** in large samples.

The IV Regression Assumptions (1/2)

1. **Conditional Mean Zero for Included Exogenous Variables**
   - The conditional mean of u_i given the included exogenous variables (W's) is zero:

$$\mathbb{E}[u_i \mid W_i] = 0$$

2. **i.i.d. Sampling**
   - The observations are independently and identically distributed, as if from simple random sampling.
3. **No Large Outliers**
   - The data are not affected by large outliers that could distort inference.
4. **Instrument Validity**
   - The two conditions for valid instruments (relevance and exogeneity) must hold.
   - This includes:
     - Instruments are correlated with endogenous regressors (conditional on W),
     - Instruments are uncorrelated with the error term u_i.

The IV Regression Assumptions (2/2)

- The IV assumptions are a modification of the least squares assumptions for causal inference.
- The instrument relevance assumption also ensures that the regressors in the second-stage regression are not perfectly multicollinear.
- Assumptions ensure that the TSLS estimator has the same large-sample properties as the OLS estimator: consistency and asymptotic normality.

Sampling Distribution of the TSLS Estimator

- Under IV assumptions, the TSLS estimator is:
  - **Consistent**, and
  - **Asymptotically normal** in large samples.
- This holds even with multiple instruments and exogenous regressors.

Inference Using the TSLS Estimator

- Standard inference tools apply:
  - **Confidence intervals**: $\hat{\beta}_j^{TSLS} \pm 1.96 \cdot SE(\hat{\beta}_j^{TSLS})$
  - **Hypothesis testing**: Use t-statistics or F-tests for joint hypotheses.
- These rely on the large-sample normality of TSLS.

Calculation of TSLS Standard Errors

- **Important warning**: Standard errors from OLS in the second stage are incorrect unless adjustments are made.
  - TSLS standard errors must account for the two-stage nature of the estimator.
- In practice:
  - Software packages implement the correct formulas.
- **Heteroskedasticity**:
  - As usual, robust standard errors should be used if the error term u_i is not homoskedastic.

Required Reading

- **Stock and Watson (2020)**
- Chapter 12, Sections 12.1 and 12.2