Instrumental Variables

Md Moshi Ul Alam

April 13, 2025

With selection bias OLS estimates are biased and inconsistent

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

- Selection bias in observational data $\implies$
- Failure of MLR 4: $E(\varepsilon_i | D_i) \neq 0$: D_i **is endogenous.**

With selection bias OLS estimates are biased and inconsistent

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

- Selection bias in observational data $\implies$
- Failure of MLR 4: $E(\varepsilon_i|D_i) \neq 0$: D_i **is endogenous.**

$$E(\varepsilon_i|D_i) \neq 0 \implies cov(D_i, \varepsilon_i) \neq 0$$

With selection bias OLS estimates are biased and inconsistent

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

- Selection bias in observational data $\implies$
- Failure of MLR 4: $E(\varepsilon_i | D_i) \neq 0$: D_i **is endogenous.**

$$E(\varepsilon_i | D_i) \neq 0 \implies cov(D_i, \varepsilon_i) \neq 0$$

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \quad \text{where} \quad cov(D_i, \varepsilon_i) \neq 0$$

$\widehat{\beta_1}^{OLS} = \frac{\widehat{cov}(Y_i, D_i)}{\widehat{var}(D_i)}$ will be **biased** & **inconsistent** when D_i is **endogenous**

Introduce concept of consistency & Draw DAG

Motivating example

- Example: Suppose college attendance (D_i) depends on a bunch of factors,

  1. High school GPA
  2. Parents' income
  3. Intrinsic motivation

     - Comparing the earnings (Y_i) of students who do/do not attend college for first two reasons likely confounds causal effect of college by selection on unobservables

Motivating example

- Example: Suppose college attendance (D_i) depends on a bunch of factors,
  1. High school GPA
  2. Parents' income
  3. Intrinsic motivation
     - Comparing the earnings (Y_i) of students who do/do not attend college for first two reasons likely confounds causal effect of college by selection on unobservables
  4. An admissions lottery

Motivating example

- Example: Suppose college attendance (D_i) depends on a bunch of factors,
  1. High school GPA
  2. Parents' income
  3. Intrinsic motivation
     - Comparing the earnings (Y_i) of students who do/do not attend college for first two reasons likely confounds causal effect of college by selection on unobservables
  4. An admissions lottery
- The 4th factor may be *as good as randomly assigned*
- IV is a way of focusing on just the "clean exogenous variation"

Basic Framework for IV

- Try to find an IV Z_i which satisfies

  - $cov(Z_i, D_i) \neq 0$ (IV relevance) : testable

  - $cov(Z_i, \varepsilon_i) = 0$ (IV exogeneity) : non-testable

- Under these and some other standard assumptions we can use Z_i to consistently estimate β_1 even if D_i is endogenous.

Intuiton and Relevant questions for IV

- For IV, the relevant questions are,
    - Why does treatment D_i vary across units with the IV Z_i
    - Is the variation in treatment via the variation in the IV exogenous?"
- The IV (Z_i) isloates the exogenous variation in the endogenous variable (D_i) due to the exogenous variation in Z_i by virtue of its covariance with D_i
- let us quickly walk through an example : Angrist and Evans (1998)

Example 1:

- Angrist and Evans (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size"
  - Sample of mothers with at least two children
  - $Y_i = 1$ if mother i is employed, 0 otherwise
  - $D_i = 1$ if mother i has at least three children, 0 otherwise

- What is the causal question?
- What are some confounding factors?
- Is it possible to control for these factors with observables?

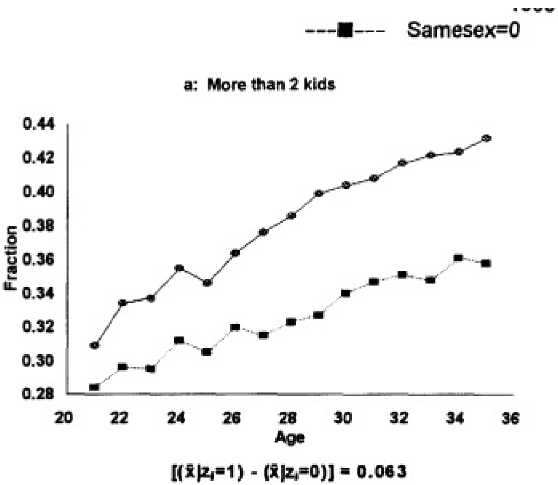
Example 1:

- Angrist and Evans (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size"
    - $Y_i = 1$ if mother i is employed, 0 otherwise
    - $D_i = 1$ if mother i has at least three children, 0 otherwise

- Why does D_i vary across women ?
- Are any of these reasons exogenous? (brainstorming possible IVs)

Example 1:

- Angrist and Evans (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size"
  - $Y_i = 1$ if mother i is employed, 0 otherwise
  - $D_i = 1$ if mother i has at least three children, 0 otherwise
  - $Z_i = 1$ if mother's first two children are same sex, 0 otherwise
- What are the signs of $cov(Z_i, D_i)$, and $cov(Z_i, \varepsilon_i)$?



we will come back to more details later.

IV in Bivariate model with one endogenous variable (D) and one exogenous (Z) variable

Framework

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \quad \text{where} \quad cov(D_i, \varepsilon_i) \neq 0$$

- Suppose instead that we have a **valid** instrument Z_i
- i.e., Z_i satisfies $\underbrace{cov(Z_i, \varepsilon_i) = 0}_{exogeneity}$ and $\underbrace{cov(Z_i, D_i) \neq 0}_{relevance}$.
- Note that in the general case with more covariates, all that is required is that the covariance of Z_i and ε_i conditional on covariates X_i equals zero

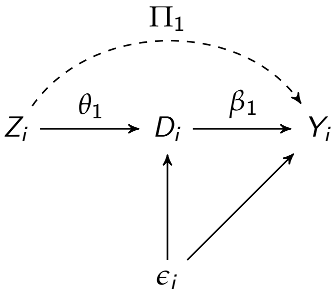
$$cov(Z_i, \varepsilon_i | X_i) = 0 \quad and \quad cov(Z_i, D_i | X_i) \neq 0$$

IV lingo–breaking it down to 3 effects

The structural equation
$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$

The first stage
$D_i = \theta_0 + \theta_1 Z_i + u_i$

The reduced form
$Y_i = \Pi_0 + \Pi_1 Z_i + e_i$



Another derivation from exogeneity and relevance

$cov(Z_i, \epsilon_i) = 0$

Quick implementation in R

- Install and load the `AER` library
- Use the `ivreg` function.
- While reporting results always compare with OLS

IV in Bivariate model with one endogenous covariate ($D \in \{0, 1\}$) and one exogenous ($Z \in \{0, 1\}$) variable

The Wald estimator

- A useful special case occurs when both the endogenous variable and the instrument are binary.
  - $Y_i =$ outcome
  - $D_i \in \{0, 1\}$ treatment (not randomly assigned)
  - $Z_i \in \{0, 1\}$ binary instrumental variable
- In this case, the IV formula reduces to a simple form called the Wald estimator.

The Wald estimator

The reduced form

- Consider the reduced form outcome model $Y_i = \Pi_0 + \Pi_1 Z_i + e_i$
- In the case of $Z_i \in \{0, 1\}$, the coefficient on the instrument is simply the difference in mean outcomes between units with $Z_i = 0, 1$.
- Estimate of the reduced form coefficient equals

$$\widehat{\Pi}_1 = \widehat{E}(Y_i | Z_i = 1) - \widehat{E}(Y_i | Z_i = 0)$$

where the expectations are estimated by sample means.

The Wald estimator

The first stage

- Now consider the first stage regression of the endogenous variable $D_i \in \{0, 1\}$ on the instrument $Z_i \in \{0, 1\}$ is a linear probability model in this case.

- The first stage coefficient estimate equals

$$\begin{aligned}
\widehat{\theta}_1 &= \widehat{E}(D_i|Z_i = 1) - \widehat{E}(D_i|Z_i = 0) \\
&=
\end{aligned}$$

The Wald estimator

- Combining the two estimates using the formula above gives the Wald estimator

$$\widehat{\beta}_{IV} = \frac{\widehat{\Pi}_1}{\widehat{\theta}_1} = \frac{\widehat{E}(Y_i|Z_i = 1) - \widehat{E}(Y_i|Z_i = 0)}{\widehat{Pr}(D_i = 1|Z_i = 1) - \widehat{Pr}(D_i = 1|Z_i = 0)}$$

- Observe how it captures exactly the essence of $\frac{cov(Z_i, Y_i)}{cov(Z_i, D_i)}$.
- Relate this to Angrist and Evans (1998) example

Illustrative example of the Wald estimator

An Illustrative example: the effect of job training on earnings

- Interested in the labor market success of workers following a lay off.

- Specifically, we are interested in whether a government job training program boosts the wages these workers earn once they become reemployed.

- Notation:
  - $Y_i =$ weekly earnings in i's first new job after a layoff
  - $D_i \in \{0, 1\}$ is an indicator for whether i chose to participate in the job training program

- What is wrong with using an OLS estimator of the impact of job training?

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$
$$\beta_1^{OLS} = E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)$$

Example: Job Training

Issue with OLS – endogeneity

- If participation is voluntary, we might worry that the people who choose to participate are **highly motivated.**
- $\implies$ experiences of those who choose not to participate are **not** representative of the "counterfactual" experiences of those who **do** participate.
- The people who participate in the job training program might have had high earnings (Y_i) even if they had not enrolled in job training.

Example: Job Training

Potential IV

- Let's think about using an example IV strategy instead.
- One common IV strategy exploits variables that are related to individuals "cost" of obtaining treatment.
- One notion of cost is the distance from where individuals live to the job training center.
- The idea is that a person lives far away from the treatment center is less likely to enroll in job training than that same person would be if they lived closer.
- There are important issues which we will ignore for now.

Example: Job Training

- Consider as an example a training center in one town that serves two towns.
  - Those in the same town as the center have a short distance to go
  - Those in the other town have a long distance to go

Example: Job Training

- The outcome in the absence of training, denoted by $Y_0 = 100$.
- Suppose the causal impact of training on those who take it is $\Delta = 10$ (i.e., $Y_1 = 110$)
- The tuition for the training course is 5.

Example: Job Training

Let's create an example scenario where everyone who is eligible in one town enrolls, while only half of those eligible in the other town enrolls

- 200 eligible persons in each town.
- For those in the near town, the cost is zero for everyone.
- The cost of travel for those in the far town varies across persons.
    - In the far town, for those with a car the cost is essentially zero;
    - Assume that half of the eligible persons have a car
    - for those without one the cost is 12. Will they enroll?
- Assume that everyone knows their cost/benefits and participates only when the benefits exceed the costs.

Example: Job Training

- Let $Z = 1$ for residence in the near town and ($Z = 0$ in the far town)
- Let $D = 1$ denote participation in training ($D = 0$ for nonpartipation)
- Now, let's write the information in the model's setup using our standard notation,
  - $Pr(D = 1|Z = 1) = 1$
  - $Pr(D = 1|Z = 0) = 0.5$
- Difference in participation probabilities between the near town means our instrument is correlated with the endogenous variable.
- At the same time, by construction the instrument is not correlated with the untreated outcome (which is a constant in this example).

Example: Job Training

- What will average outcomes be in the two towns?

$$\begin{aligned} E(Y|Z=1) &= Y_0 + \Delta \times Pr(D=1|Z=1) \\ &= 100 + 10x1.0 \\ &= 110 \end{aligned}$$

$$\begin{aligned} E(Y|Z=0) &= Y_0 + \Delta \times Pr(D=1|Z=0) \\ &= 100 + 10x0.5 \\ &= 105 \end{aligned}$$

Example: Job Training

- The IV estimator in this simple case is given by:

$$\widehat{\beta}_{IV} = \frac{\widehat{E}(Y|Z=1) - \widehat{E}(Y|Z=0)}{\widehat{Pr}(D=1|Z=1) - \widehat{Pr}(D=1|Z=0)}$$

- Numerator: mean difference in outcomes between the two towns
- The denominator scales up the numerator to account for the fact that the change in the IV changes the status for some but not all individuals.
- We will formalize this in the 2 lectures with the concept of *local average treatment effect (LATE) and compliers*.

Example: Job Training

- Inserting the numbers from the example into the formula gives:

$$\widehat{\beta}_{IV} = \frac{110 - 105}{1.0 - 0.5} = \frac{5}{0.5} = 10.$$

  which is the correct answer.

- The IV estimator isolates variation in the endogenous variable (D_i) due to specific exogenous factors (Z_i)
- thus is not biased by the selection on unobservables if one can:
  1. test the relevance of the IV $cov(D_i, Z_i) \neq 0$ and
  2. defend the exogeneity of the IV $cov(Z_i, \varepsilon_i) = 0$.

Quick IV Recap so far

- To tease out causal effects: IV satisfying *relevance* and *exogeneity* conditions
- Linked with the structural equation, the first stage and the reduced form

$$\frac{\text{Reduced form coeff}}{\text{First stage coeff}}$$

- IV estimator in bivariate models

$$\frac{cov(Y_i, Z_i)}{cov(D_i, Z_i)}$$

  - Discrete D_i and Z_i this becomes the Wald estimator

$$\frac{\widehat{E}(Y_i|Z_i = 1) - \widehat{E}(Y_i|Z_i = 0)}{\widehat{Pr}(D_i = 1|Z_i = 1) - \widehat{Pr}(D_i = 1|Z_i = 0)}$$

Asymptotic Bias and Weak Instruments

Asymptotic bias in IV

- Consider one IV (Z_i) and one endogenous variable (X_i)
- We know $\beta_{IV} = \dfrac{cov(z_i, y_i)}{cov(z_i, x_i)}$. So the IV estimate:

$$\widehat{\beta}_{IV} = \frac{\widehat{cov}(z_i, y_i)}{\widehat{cov}(z_i, D_i)} =$$

- From this, it follows that as $N \to \infty$,

$$\widehat{\beta}_{IV} \to \tag{1}$$

Asymptotic bias in IV

- As $N \to \infty$,

$$\widehat{\beta}_{IV} \to \beta_1 + \frac{cov(z_i, \epsilon_i)}{cov(z_i, D_i)}$$

- The asymptotic bias depends on the relative strength of the population relationships:
  - $cov(z_i, \epsilon_i)$: covariance between the instrument and the error term
  - $cov(z_i, D_i)$: covariance between the instrument and the endogenous variable
- Although it is assumed that $cov(z_i, \epsilon_i) = 0$, it is likely not zero but small
- If the **instruments are weak**, i.e. $cov(z_i, D_i) \simeq 0$, then even a small correlation can lead to a large bias.

Asymptotic bias: IV compared to OLS

- It is interesting to compare the asymptotic biases of OLS and IV.

$$\widehat{\beta}_{IV} \rightarrow \beta_1 + \frac{cov(z_i, \epsilon_i)}{cov(z_i, D_i)}$$

$$\hat{\beta}_{OLS} \rightarrow \beta_1 + \frac{cov(D_i, \epsilon_i)}{cov(D_i, D_i)}$$

- The ratio of the asymptotic bias of IV to that of OLS then equals

$$\frac{cov(z_i, \epsilon_i) cov(z_i, D_i)}{cov(D_i, \epsilon_i) cov(D_i, D_i)} = \frac{\rho_{z\epsilon}}{\rho_{D\epsilon}\rho_{zD}}$$

Bias comparison: IV vs OLS

- For IV to be preferred to OLS, the absolute value of the ratio of the biases should be less than one.

$$|\rho_{z\epsilon}| < |\rho_{D\epsilon}\rho_{zD}|$$

- Thus, IV is likely to have a lower bias than OLS when the instrument is not very correlated with the residual, when:
  - the endogenous variable is highly correlated with the residual
  - and the instrument is strongly correlated with the endogenous variable
- Otherwise we will have the problem of “**weak instruments**”
- Next assignment will have a problem with weak IVs

Illustrative Example: Weak Instrument

- Effect of extra study time on exam performance.

Illustrative Example: Weak Instrument

- Effect of extra study time on exam performance.
- **Endogeneity:** Motivated students study more and perform better.

Illustrative Example: Weak Instrument

- Effect of extra study time on exam performance.
- **Endogeneity:** Motivated students study more and perform better.
- **Instrument:** Distance from dorm to the campus library.

Illustrative Example: Weak Instrument

- Effect of extra study time on exam performance.
- **Endogeneity:** Motivated students study more and perform better.
- **Instrument:** Distance from dorm to the campus library.
- **Assumption:** Distance affects study time but not exam performance.

Let us simulate a dataset to see this problem in action IV with

- the true causal effect of 1.5
- endogenous D_i (study time)
- a weak IV Z_i (distance to library)

Variation in the IV

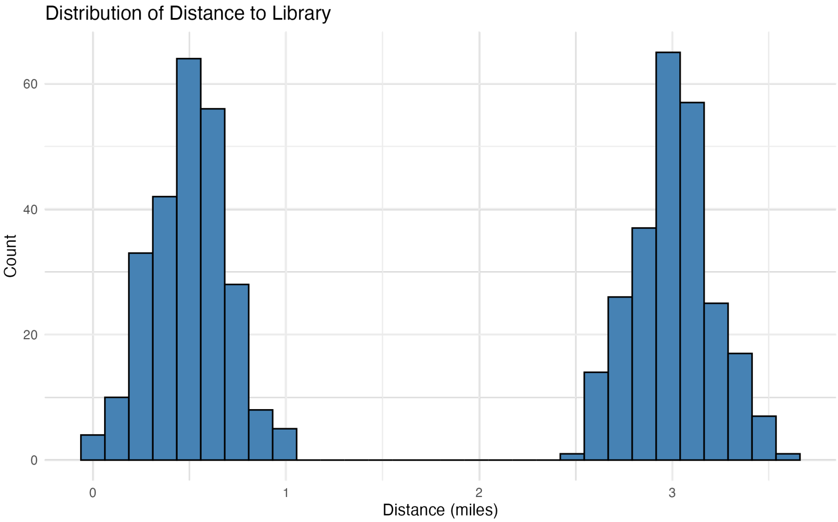


Figure: IV: Distance from dorm to library

Illustratively examining the strength of the IV

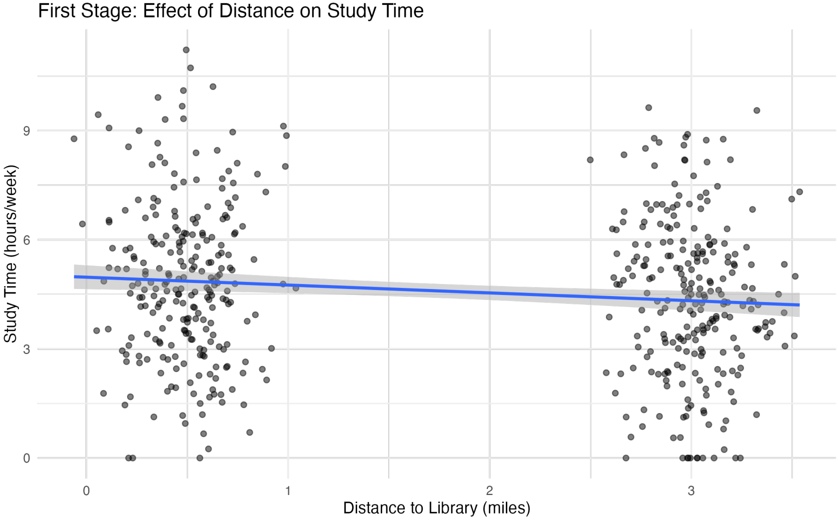


Figure: Study Time (D_i) vs. Distance (Z_i)

Weak IV Issue

- If distance (Z_i) only marginally influences study time (D_i)

  (or if most students live either nearby or far)
- Then IV estimates become imprecise and unreliable.

Weak IV Issue

- If distance (Z_i) only marginally influences study time (D_i)

  (or if most students live either nearby or far)
- Then IV estimates become imprecise and unreliable.
- Rule of thumb: F-statistic of 1st stage < 10. *(Why F and not t?)*
- In R: `summary(iv_model, diagnostics = TRUE)`

OLS vs IV results

Table: Comparison of OLS and IV Estimates

	Dependent variable:	
	Exam Score	
	OLS	*instrumental variable*
	(1)	(2)
Study Time	3.150***	0.330
	(0.118)	(1.386)
Constant	52.888***	65.840***
	(0.602)	(6.377)
Observations	500	500
R^2	0.588	0.117
Adjusted R^2	0.587	0.115
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
Call:
ivreg(formula = exam_score ~ study_time | distance, data = student_data)

Residuals:
     Min        1Q    Median        3Q       Max
-28.2201   -6.0128   0.3707    5.3254   27.6730

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.8399     6.3774  10.324   <2e-16 ***
study_time    0.3301     1.3860   0.238    0.812

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1 498     7.880  0.0052 **
Wu-Hausman         1 497     9.163  0.0026 **
Sargan             0  NA        NA      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.505 on 498 degrees of freedom
Multiple R-Squared: 0.1168,     Adjusted R-squared: 0.115
Wald test: 0.05672 on 1 and 498 DF,  p-value: 0.8119
```

Figure: Weak Instrument: IV Results

Multivariate models: IV and 2SLS

Two Stage Least Squares (2SLS): First Stage

- Using the linear projection of the endogenous variables on the exogenous variables as the instrument leads to the two-stage least squares estimator.

- Structural equation:

$$y = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_K x_{K,i} + \beta_D \underbrace{D_i}_{\textit{endogenous variable}} + \varepsilon_i$$

- First stage: regress the endogenous variable on **the exogenous variables**

$$D_i = \delta_0 + \delta_1 x_{1,i} + \ldots + \delta_K x_{K,i} + \theta Z_i + u_i$$

- Generate predicted values of the endogenous variable:

$$\widehat{D}_i = \widehat{\delta}_0 + \widehat{\delta}_1 x_{1,i} + \ldots + \widehat{\delta}_K x_{K,i} + \widehat{\theta} Z_i$$

Two Stage Least Squares (2SLS): Second stage

- Estimate the model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_K x_{K,i} + \beta_D \widehat{D}_i + e_i$$

where the error term is now uncorrelated with all of the independent variables, thus giving consistent (but not unbiased) estimates.

How does the system look?

Original regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \beta_D D_i + \epsilon_i$$

- To compactly “stack” observations, define matrix notation:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} \; ; \; X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} & D_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{iK} & D_i \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NK} & D_N \end{bmatrix} \; ; \; \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \\ \beta_D \end{bmatrix} \; ; \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_N \end{bmatrix}$$

How does the system look?

All exogenous variables: Included and excluded

- Consider the case where X_1 to X_K are exogenous, and X_K is endogenous

- Denote a matrix of all exogenous variables:

$$Z = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} & z_{11} & \cdots & z_{1M} \\ \vdots & \ddots & & \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{iK} & z_{i1} & \cdots & z_{iM} \\ \vdots & & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} & z_{N1} & \cdots & z_{NM} \end{bmatrix}$$

Assumption 2SLS1: Instrument Exogeneity

- The instruments are uncorrelated with the error term.
- More formally: $E[Z_i \epsilon_i] = 0$, where:
  - Z_i includes both the exogenous variables from the original regression and the IV.
  - ϵ_i is the error term from the structural equation.
- Intuition: The variables we use to instrument (i.e., Z_i) should not contain any information about the unobserved factors affecting the outcome.
- Why this matters: If the IV are correlated with the error term, they will inherit the same bias as the endogenous variables they replace.

Assumption 2SLS2: Instrument Relevance and No Redundancy

- two parts:
  - (a) The IV are not perfectly correlated with each other. (no redundancy)
  - (b) The IV are good predictors of the endogenous regressors. (relevance)
- Part (a): No instrument can be written as an exact combination of the others. This ensures we are using truly independent sources of variation.
- Part (b): Each endogenous variable must be “explained” to some extent by the IV. This is what allows us to isolate variation in the endogenous variables that is not contaminated by the error term.
- Intuition: If the IV are weak (i.e., poor predictors of the endogenous variables), then 2SLS will also perform poorly.

Deriving the IV parameter (just-identified case)

- Start with the model: $Y_i = \beta X_i + \epsilon_i$ where X_i is endogenous: $Cov(X_i, \epsilon_i) \neq 0$
- Use IV Z_i such that:
  - $Cov(Z_i, \epsilon_i) = 0$ (exogeneity)
  - $Cov(Z_i, X_i) \neq 0$ (relevance)
- Use the exogeneity condition: $E[Z_i \cdot \epsilon_i] = 0$
- Substitute $\epsilon_i = Y_i - \beta X_i$:

$$E[Z_i(Y_i - \beta X_i)] = 0$$
$$\Rightarrow E[Z_i Y_i] = \beta E[Z_i X_i]$$
$$\Rightarrow \beta = \frac{E[Z_i Y_i]}{E[Z_i X_i]}$$

- This is the IV estimand: it gives us β using only variation in X_i that comes from Z_i.

Consistency of the IV estimator

- In practice, we estimate expectations using sample averages:

$$\hat{\beta}_{IV} = \frac{\frac{1}{N}\sum Z_i Y_i}{\frac{1}{N}\sum Z_i X_i}$$

- As $N \to \infty$, by the Law of Large Numbers:

$$\hat{\beta}_{IV} \to \frac{E[Z_i Y_i]}{E[Z_i X_i]} = \beta$$

- So $\hat{\beta}_{IV}$ is consistent for β.
- This derivation works when there is only **one instrument** for each **endogenous variable**: the just-identified case.
    - Z has the same number of columns as X

Variance

Variance of the 2SLS Estimator

- one endogenous variable D_i, one IV Z_i, and k exogenous covariates $X_{1i}, \ldots, X_{ki}$
- Under homoskedasticity (constant error variance σ^2)
- The variance of the 2SLS estimator for the coefficient on D_i is:

$$\text{Var}(\hat{\beta}_{2SLS}) = \frac{\hat{\sigma}^2/(n-K)}{R^2_{D,Z} \cdot (1 - R^2_{D,X}) \cdot \sum_{i=1}^{n}(D_i - \bar{D})^2}$$

- $R^2_{D,Z}$ is the partial R^2 from the first stage
- $R^2_{D,X}$ is the R^2 from regressing D on exogenous covariates X

<title>

1. **First-stage strength:** $R^2_{D,Z}$ in denominator
   - Weaker instrument → Lower $R^2_{D,Z}$ → Higher variance
2. **Control variables:** $1 - R^2_{D,X}$ in denominator
   - Better prediction of D with X → Lower variance
3. **Sample variation in D:** $\sum(D_i - \bar{D})^2$ in denominator
   - More variation in D → Lower variance

Comparing 2SLS and OLS Variance

OLS Variance:

$$\text{Var}(\hat{\beta}_{OLS}) = \frac{\hat{\sigma}^2/(n-K)}{(1-R^2_{D,X}) \cdot \sum_{i=1}^{n}(D_i - \bar{D})^2}$$

2SLS Variance:

$$\text{Var}(\hat{\beta}_{2SLS}) = \frac{\hat{\sigma}^2/(n-K)}{R^2_{D,Z} \cdot (1-R^2_{D,X}) \cdot \sum_{i=1}^{n}(D_i - \bar{D})^2}$$

Since $R^2_{D,Z} < 1$, we have:

$$\text{Var}(\hat{\beta}_{2SLS}) > \text{Var}(\hat{\beta}_{OLS})$$

Efficiency comparison

1. 2SLS estimators are always less efficient than OLS estimators
2. The efficiency loss depends on instrument strength:
   - Stronger instruments (higher $R^2_{D,Z}$) $\rightarrow$ closer to OLS efficiency
   - Weak instruments $\rightarrow$ dramatically higher variance
3. Variance inflation factor: $1/R^2_{D,Z}$
   - $R^2_{D,Z} = 0.5 \rightarrow 2\times$ larger variance than OLS
   - $R^2_{D,Z} = 0.1 \rightarrow 10\times$ larger variance than OLS
4. Trade-off: Consistency (removing endogeneity bias) vs. Efficiency

Multiple Instruments

Multiple instruments and two-stage least squares

- Suppose instead that we have multiple instruments $z_1, \ldots, z_M$ all of which have non-zero covariance with the endogenous var

- The exogeneity assumption needs to hold for each and every instrument:
  - 1) $cov(Z_{1i}, \varepsilon_i) = 0$
  - 2) $cov(Z_{2i}, \varepsilon_i) = 0$

    $\vdots$
  - M) $cov(Z_{Mi}, \varepsilon_i) = 0$

Definition: Overidentified

- When there are more instruments (Z) than endogenous regressors (X), the model is said to be **“over-identified”**
    - Example: case with one endogenous variable, one instrument is needed
- The single instrument case is referred to as a “just-identified” model.
- If there are M instruments and one endogenous variable, then there are $M-1$ **“over-identifying restrictions”**
    - These restrictions can be tested (later)

Heterogeneous Treatment Effects

Table 4
OLS and IV estimates of the return to education with instruments based on features of the school system[a]

Author	Sample and instrument		Schooling coefficients	
			OLS	IV
1. Angrist and Krueger (1991)	1970 and 1980 Census Data, Men. Instruments are quarter of birth interacted with year of birth. Controls include quadratic in age and indicators for race, marital status, urban residence	1920–1929 cohort in 1970	0.070 (0.000)	0.101 (0.033)
		1930–1939 cohort in 1980	0.063 (0.000)	0.060 (0.030)
		1940–1949 cohort in 1980	0.052 (0.000)	0.078 (0.030)
2. Staiger and Stock (1997)	1980 Census, Men. Instruments are quarter of birth interacted with state and year of birth. Controls are same as in Angrist and Krueger, plus indicators for state of birth. LIML estimates	1930–1939 cohort in 1980	0.063 (0.000)	0.098 (0.015)
		1940–1949 cohort in 1980	0.052 (0.000)	0.088 (0.018)
3. Kane and Rouse (1993)	NLS Class of 1972, Women. Instruments are tuition at 2 and 4-year state colleges and distance to nearest college. Controls include race, part-time status, experience. Note: Schooling measured in units of college credit equivalents	Models without test scores or parental education	0.080 (0.005)	0.091 (0.033)
		Models with test scores and parental education	0.063 (0.005)	0.094 (0.042)
4. Card (1995b)	NLS Young Men (1966 Cohort) Instrument is an indicator for a nearby 4-year college in 1966, or the interaction of this with	Models that use college proximity as instrument (1976 earnings)	0.073 (0.004)	0.132 (0.049)

Table 4 (*continued*)

Author	Sample and instrument		Schooling coefficients OLS	IV
	parental education. Controls include race, experience (treated as endogenous), region, and parental education	Models that use college proximity × family background as instrument	–	0.097 (0.048)
5. Conneely and Uusitalo (1997)	Finnish men who served in the army in 1982, and were working full time in civilian jobs in 1994. Administrative earnings and education data. Instrument is living in university town in 1980. Controls include quadratic in experience and parental education and earnings.	Models that exclude parental education and earnings	0.085 (0.001)	0.110 (0.024)
		Models that include parental education and earnings	0.083 (0.001)	0.098 (0.035)
6. Maluccio (1997)	Bicol Multipurpose Survey (rural Philippines): male and female wage earners age 20–44 in 1994, whose families were interviewed in 1978. Instruments are distance to nearest high school and indicator for local private high school. Controls include quadratic in age.	Models that do not control for selection of employment status or location	0.073 (0.011)	0.145 (0.041)
		Models with selection correction for location and employment	0.063 (0.006)	0.113 (0.033)
7. Harmon and Walker (1995)	British Family Expenditure Survey 1978–1986 (men). Instruments are indicators for changes in the minimum school leaving age in 1947 and 1973. Controls include quadratic in age, year, survey and region and region		0.061 (0.001)	0.153 (0.015)

[a] Notes: see text for sources and information on individual studies.

So far..

- So far implicit baseline model has been a model of common treatment effects

$$Y_i = \beta X_i + \delta D_i + \varepsilon_i$$

  - Everyone has the same value for the coefficient δ on the treatment variable - ATT
  - Can interact treatment with other observed characteristics to estimate *subgroup ATT*

Heterogeneous treatment effects

- In many settings, there is no reason to expect that a given treatment everyone in the same way

$$Y_i = \beta X_i + \delta_i D_i + \varepsilon_i$$

- Lets simplify: Remove all covariates and see what an IV (Z_i) identifies in this world

So whats going on here?

- Consider a typical IV strategy in a treatment effects context: random variation in the cost of accessing treatment. (e.g. distance from the training center or college, presence of a bus strike)

- This generates exogenous variation in the probability of treatment.

- If $cov(Z_i, \epsilon_i) = 0$, then this constitutes a valid instrument in the common effect world.

- However, it is easy to show that such instruments are not likely to be valid in the heterogeneous treatment effects world, if agents know both their costs Z_i and their likely impacts from the program δ_i.

Some intuition before formalization

- Suppose all of you had the same benefits for a given treatment
- Any 2 of you with the same costs will have the same likelihood of participation
- But suppose benefits differ between you
- Then even with same costs your likelihood of participation will differ
- Taking up treatment can not be forced, even with random assignment, some may choose not to take up treatment
- This is a problem because as econometricians we do not observe individual specific benefits which individuals themselves maybe aware of thus driving another type of selection

Selection based on unit-specific treatment effects (δ_i)

- Suppose that participation is determined by

$$D_i^* = \gamma_0 + \gamma_1 \delta_i - \gamma_2 Z_i + v_i$$

- Where $D_i = 1$ iff $D_i^* > 0$ and $D_i = 0$ otherwise, and where Z_i is a variable that measures cost of participation

Selection based on unit-specific treatment effects (δ_i)

- Suppose that participation is determined by

$$D_i^* = \gamma_0 + \gamma_1 \delta_i - \gamma_2 Z_i + v_i$$

- Where $D_i = 1$ iff $D_i^* > 0$ and $D_i = 0$ otherwise, and where Z_i is a variable that measures cost of participation

- Likelihood of different participation status
  - $D_i = 1$: Need high δ_i relative to costs Z_i
  - $D_i = 0$: Need low δ_i relative to costs Z_i

-> Conditional on D_i, the value of the instrument Z_i is correlated with the unit-specific component of the impact δ_i.

Selections in Heterogeneous treatment effects world

- In essence $E[\delta_i|Z_i, D_i] \neq 0$
- Put differently, when we are interested in ATT, in a heterogenous treatment effects world, there are two kinds of selection
  - One on the unobserved component of the outcome equation (ϵ_i)
  - And one on the unobserved component of the impact (δ_i)
- However, all is not lost. As long as $cov(Z_i, \varepsilon_i) = 0$ (IV $\perp$ original outcome eqn. error term), IV still estimates a causal parameter *under some additional assumptions*

What does the IV identify?
Angrist and Imbens (1994)

Heterogeneous treatment effects, Wald Estimator

- Simple case of $Z \in \{0, 1\}$, a binary treatment $D \in \{0, 1\}$ and no covariates
- There are four possible distinct groups in terms of how the instrument (Z) relates to treatment (D)

Groups	$Z = 0$	$Z = 1$
Never takers (NT)	$D = 0$	$D = 0$
Defiers (DF)	$D = 1$	$D = 0$
Compliers (C)	$D = 0$	$D = 1$
Always takers (AT)	$D = 1$	$D = 1$

- The compliers are the key group here
- They are the units that respond to the instrument by taking the treatment when they are randomly assigned, otherwise would not

Heterogeneous treatment effects, Wald Estimator

- Now consider the formula for the Wald estimator once again.

$$\delta_{IV} = \frac{E(Y|Z=1) - E(Y|Z=0)}{Pr(D=1|Z=1) - Pr(D=1|Z=0)}$$

- We can express both the numerator and denominator using the four groups and their potential outcomes $\{Y_0, Y_1\}$

Groups	$Z=0$	$Z=1$
Never takers (NT)	$D=0$	$D=0$
Defiers (DF)	$D=1$	$D=0$
Compliers (C)	$D=0$	$D=1$
Always takers (AT)	$D=1$	$D=1$

Numerator of the Wald estimator-I

$E(Y|Z=1) - E(Y|Z=0)$

Groups	$Z=0$	$Z=1$
Never takers (NT)	$D=0$	$D=0$
Defiers (DF)	$D=1$	$D=0$
Compliers (C)	$D=0$	$D=1$
Always takers (AT)	$D=1$	$D=1$

So, with these 4 groups lets write out

$$E(Y|Z=0) =$$

Numerator of the Wald estimator-II

$E(Y|Z=1) - E(Y|Z=0)$

Groups	$Z=0$	$Z=1$
Never takers (NT)	$D=0$	$D=0$
Defiers (DF)	$D=1$	$D=0$
Compliers (C)	$D=0$	$D=1$
Always takers (AT)	$D=1$	$D=1$

$$\begin{aligned} E(Y|Z=1) \quad = \quad & E(Y_0|NT)\ Pr(NT) \\ & + E(Y_0|DF)\ Pr(DF) \\ & + E(Y_1|C)\ Pr(C) \\ & + E(Y_1|AT)\ Pr(AT) \end{aligned}$$

Heterogeneous treatment effects, Wald Estimator numerator

$E(Y|Z=1) - E(Y|Z=0)$

- Using the four groups, we can rewrite the terms in the numerator as:

$$
\begin{aligned}
E(Y|Z=0) \quad = \quad & E(Y_0|NT)\ Pr(NT) \\
& + E(Y_1|DF)\ Pr(DF) \\
& + E(Y_0|C)\ Pr(C) \\
& + E(Y_1|AT)\ Pr(AT)
\end{aligned}
$$

$$
\begin{aligned}
E(Y|Z=1) \quad = \quad & E(Y_0|NT)\ Pr(NT) \\
& + E(Y_0|DF)\ Pr(DF) \\
& + E(Y_1|C)\ Pr(C) \\
& + E(Y_1|AT)\ Pr(AT)
\end{aligned}
$$

Heterogeneous treatment effects, Wald Estimator denominator

$Pr(D = 1|Z = 1) - Pr(D = 1|Z = 0)$

$$\begin{aligned} Pr(D = 1|Z = 0) \quad = \quad & Pr(D = 1|Z = 0, NT)\ Pr(NT) \\ & + Pr(D = 1|Z = 0, DF)\ Pr(DF) \\ & + Pr(D = 1|Z = 0, C)\ Pr(C) \\ & + Pr(D = 1|Z = 0, AT)\ Pr(AT) \end{aligned}$$

and

Heterogeneous treatment effects, Wald Estimator denominator

$Pr(D = 1|Z = 1) - Pr(D = 1|Z = 0)$

$$\begin{aligned} Pr(D = 1|Z = 0) \quad = \quad & Pr(D = 1|Z = 0, NT)\ Pr(NT) \\ & + Pr(D = 1|Z = 0, DF)\ Pr(DF) \\ & + Pr(D = 1|Z = 0, C)\ Pr(C) \\ & + Pr(D = 1|Z = 0, AT)\ Pr(AT) \end{aligned}$$

and

$$\begin{aligned} Pr(D = 1|Z = 1) \quad = \quad & Pr(D = 1|Z = 1, NT)\ Pr(NT) \\ & + Pr(D = 1|Z = 1, DF)\ Pr(DF) \\ & + Pr(D = 1|Z = 1, C)\ Pr(C) \\ & + Pr(D = 1|Z = 1, AT)\ Pr(AT) \end{aligned}$$

Heterogeneous treatment effects, Wald Estimator

- Some of these terms cancel out in the subtraction, leaving

$$\delta_{IV} = \frac{\left[E(Y_1|C)Pr(C) + E(Y_0|DF)Pr(DF)\right] - \left[E(Y_0|C)Pr(C) + E(Y_1|DF)Pr(DF)\right]}{Pr(C) - Pr(DF)}$$

- Intuition is that terms for "always takers" and "never takers" will cancel out, since their behavior/outcomes are not affected by the IV
- The IV estimator in this case is a weighted average of the treatment effect on the compliers and the negative of the treatment effect on the defiers

LATE: need Two additional Identifying Assumptions

- Define the **Local Average Treatment Effect (LATE)** as the average treatment effect for compliers

$$\text{LATE} = E(Y_1 - Y_0 | C)$$

- The first assumption needed is of monotonicity:
    - The IV can only increase or only decrease the prob. of participation for all units
    - This assumption fits in very well for cost-based instruments, which theory suggests should have such a monotonic effect $\Pr(D_i = 1 \mid Z_i = 1) > \Pr(D_i = 1 \mid Z_i = 0)$
    - It means that $Pr(D) = 0$.
- The second assumption is that there are some compliers, $Pr(C) > 0$

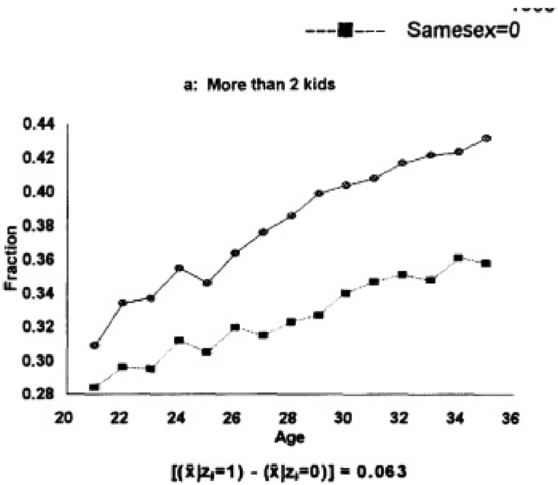
LATE

- Imposing these assumptions on the formula gives

$$\begin{aligned} \delta_{IV} &= \frac{[E(Y_1|C) - E(Y_0|C)]Pr(C)}{Pr(C)} \\ &= E(Y_1|C) - E(Y_0|C) \\ &= LATE \end{aligned}$$

- Thus, with monotonicity and some compliers, the IV estimator gives the Local Average Treatment Effect (LATE)
  - LATE $=$ the mean impact of treatment on the compliers.

Back to the Angrist and Evans (1998) example



LATE: Is it an interesting parameter?

- The LATE is a well-defined economic parameter.
- In some cases, as when the available policy option consists of moving the instrument, it may be the parameter of greatest interest.
- Thus, if we want to give a $500 tax credit for university attendance, the policy parameter of interest is the impact of university on individuals who attend with the tax credit but do not attend without it.
- The LATE provides no information about the impact of treatment on the always-takers, which could be large or small.
- If the mean impact of treatment on the treated is the primary parameter of interest, this is a very important omission.

Heterogeneous treatment effects: discussion

- When are there heterogeneous treatment effects?
- Institutional knowledge is helpful here.
- If the treatment is itself heterogeneous, as in the case of most training programs, this is suggestive that the treatment effects are likely to be as well.
- Looking for variation in the ATE across subgroups is also informative.
- If they vary a lot, then they are likely to vary with unobserved characteristics as well.
- Beyond scope of our discussion, there are some formal tests of null of zero unobserved variance in treatment effects

Summary in Heterogeneous treatment effects

- In many settings, treatment affects different units differently
- For IV, the key issue is that instruments can be correlated with the person-specific component of a treatment effect
- Can occur even if instruments are uncorrelated with the outcome equation error term
- In heterogenous treatment effects world, IV identifies LATE and not ATE
- Complier characteristics could be characterized to understand further who are being moved by the IV
- But beyond the scope of this class and covered in graduate courses