Econometrics **August 18, 2026**

Instrumental Variable Technique

Instructor: Kalyan Kolukuluri *Content: Lecture Notes*

Note: *Notes may be distributed outside this class only with the permission of the Instructor. The author does not make claims on the authenticity of the information.*

Contents

1 Observational Studies

Observational Studies are a type of research where the investigator observes the effect of a variable or intervention without actively controlling or manipulating the conditions. Unlike randomized controlled trials (RCTs), where participants are randomly assigned to treatment or control groups, observational studies rely on naturally occurring variations in treatment or exposure. These studies are commonly used in fields like epidemiology, social sciences, and economics to investigate causal relationships when controlled experiments are not feasible.

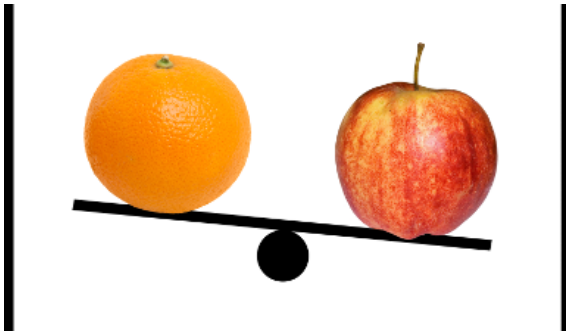
1. Cohort Studies: Follow a group of individuals over time, observing who receives the treatment or exposure and what outcomes occur.
2. Case-Control Studies: Compare individuals with a specific outcome (cases) to those without (controls), looking retrospectively at their exposures.
3. Cross-Sectional Studies: Analyze data at a single point in time to determine the relationship between exposure and outcome.

1.1 Non-Random (Treatment) Assignment

In observational studies, treatment (**T**) is **not randomly assigned**. Instead, it is typically assigned based on factors that may be related to the treatment itself or the outcome being studied. This lack of randomization introduces potential **biases** that must be carefully considered in the analysis. Some common reasons for non-random treatment assignment include:

- **Self-selection**: Individuals may choose to undergo a particular treatment based on their preferences, needs, or risk profiles. For example, people in poorer health may be more likely to seek a specific medical treatment.
- **Physician/Institution Decision**: In medical studies, doctors or institutions may decide which treatment is appropriate for a patient based on clinical judgment, past medical history, or severity of illness.
- **Economic or Social Factors**: Treatment or exposure might depend on economic, geographic, or social factors. For example, wealthier individuals might have better access to healthcare or healthier diets.

Figure 1: Apples (Enrolled) to Oranges (Non-enrolled) comparison



1.2 Issues with self-selection

Incorrect Counterfactual: Enrolled & Non-enrolled

Problem - Enrolled & Non-enrolled comparison

Reason for not enrolling may be correlated with outcome (Y) So, estimated impact is confounded with other things. This is **Selection Bias**

- Enrolled: Treatment group
- Non-enrolled: Control group (includes ineligible + eligible but not choosing to enrol)
- Non-comparability is due to *Selection Bias*.

1.3 (Self)-Selection into treatment (An illustration)

Problem - Enrolled & Non-enrolled comparison

Let us consider a job skills training program (non-random intervention), that allows particpants to voluntarily enrol. How do we measure the impact of the job skills training program?

Table 1: Self-Selection

Variable Definition	Notation	Ketan	Murali
Treatment (training) Status Chosen	T_i	1	0
Potential Outcome without skill training	Y_{0i}	$\boxed{3}$	5
Potential Outcome with skill training	Y_{1i}	4	$\boxed{5}$
Actual job outcome	Y_i	4	5
Treatment Effect (same-person comparison)	$Y_{1i} - Y_{0i}$	1	0
Treatment Effect (cross-person comparison)	$Y_{1i}^T - Y_{0i}^C$	$= 4 - 5 = -1$	

- **Ketan:** is a resident of rural Gujarat who just passed his tenth standard. Ketan knows he can clearly benefit from job training and enrolls ($T = 1$).
- **Murali:** is a resident of urban Gujarat who already worked as an assembly worker and is recently unemployed. Murali knows he does not benefit from job training and does not enroll ($T = 0$).
- **Ketan & Murali:** Both are part of your impact study.
- Net effect of comparing Ketan and Murali (Using actual job outcomes of each) = $4 - 5 = -1$
- *Negative effect due to job training !!! - this happened due to comparison b/w apples and oranges i.e. due to **selection bias**.*

1.4 Enrolment types

In some observational studies, an eligibility criterion (Z) is set-up for assigning treatment (T). The eligibility could perhaps be a numeric threshold (e.g. poverty percentile or age or income etc.) Despite the eligibility criterion /rules (Z), it is left to the individual if she would be willing to take up the treatment or not. In other words *the subjects of an observational study have a choice to enrol in the treatment.*

Now, in this context, we can subclassify the population into four groups based on their predisposition to enrol in treatment.

Table 2: Sub-groups with different attitudes towards enrolment

Complier	Never Taker	Always Taker	Defier
□□□□□□□□	⊗	△	smiling-face-with-horns
□□□□□□□□	⊗	△	smiling-face-with-horns
$T_{Z=1}=1; T_{Z=0}=0$	$T_{Z=1}=0; T_{Z=0}=0$	$T_{Z=1}=1; T_{Z=0}=1$	$T_{Z=1}=0; T_{Z=0}=1$

1. **Enroll if assigned (Compliers -** □**)**: Assume 80 % of a population and therefore a sample are compliers. They enroll when assigned and do not enroll when not assigned. They are good boys and girls. $T_{Z=1}=1; T_{Z=0}=0$
2. **Never takers (**⊗**)**: Regardless of their assignment status they DO NOT enroll. Assume 10% of population. $T_{Z=1}=0; T_{Z=0}=0$
3. **Always takers (**△**)**: Regardless of their assignment status they ALWAYS enroll. Assume 10% of population. $T_{Z=1}=1; T_{Z=0}=1$
4. **Always takers & Never takers are noncompliers (**△; ⊗**)**
5. **Defiers (smiling-face-with-horns):** Defiers are contrarians. They enrol when not assigned and do not enrol when assigned. $T_{Z=1}=0; T_{Z=0}=1$

1.5 Enrollment - Ideal vs Actual

If random assignment were adhered to in letter, each group would have equal distribution of all different types of subjects (compliers, defiers, always takers, and never takers).

Real (actual) enrollment But, in reality, Defiers smiling-face-with-horns have switched groups. Never takers ⊗ in Treatment group switched to Control group and Always takers △ in Control group switched to Treatment group.

Table 3: Ideal Program Enrolment (strict adherence by subjects to random assignment)

Group assigned to Treatment	Group assigned to Control
Ideal Enrollment (Randomized Assignment)	
□□□□□□□□⊗△ smiling-face-with-horns	□□□□□□□□⊗△ smiling-face-with-horns
Actual (real) Enrollment (Voluntary)	
□□□□□□□□△△ smiling-face-with-horns	□□□□□□□□⊗⊗ smiling-face-with-horns

1.6 Non-Parametric Identification

Can we get identification without parametric assumptions?

Before you read further, see Section A for potential outcomes notation.

C	Compliers	$T_{Z=1} = 1;\ T_{Z=0} = 0$
D	Defiers	$T_{Z=1} = 0;\ T_{Z=0} = 1$
A	Always takers	$T_{Z=1} = 1;\ T_{Z=0} = 1$
N	Never takers	$T_{Z=1} = 0;\ T_{Z=0} = 0$

These groups are called **principal strata** because membership is determined by potential outcomes and thus remains fixed regardless of actual treatment assignment.

Monotonicity Assumption: For every individual $\forall i$

$$\forall i, T_{Z=1}(i) \geq T_{Z=0}(i) \equiv \text{No defiers} smiling - face - with - horns$$

This means that switching the instrument from 0 to 1 either increases treatment uptake or leaves it unchanged for every individual—it never decreases treatment for anyone.

1.7 Average Treatment Effect ATE

We can decompose the ATE across principal strata:

$$\text{ATE} = \sum_{s \in \{C,A,N,D\}} P(s) \cdot E[Y_i(1) - Y_i(0) \mid s]$$

$$
\begin{aligned}
&E[Y_{T=1} - Y_{T=0}] \\
&= E[Y_{Z=1} - Y_{Z=0} | T_{Z=1} = 1, T_{Z=0} = 0] P(T_{Z=1} = 1, T_{Z=0} = 0) (\text{Compliers}) \\
&+ \\
&E[Y_{Z=1} - Y_{Z=0}] T_{Z=1} = 0; T_{Z=0} = 1] P(T_{Z=1} = 0, T_{Z=0} = 1) (\text{Defiers}) \\
&+ \\
&E[Y_{Z=1} - Y_{Z=0} | T_{Z=1} = 1; T_{Z=0} = 1] P(T_{Z=1} = 1, T_{Z=0} = 1) (\text{Always takers}) \\
&+ \\
&E | Y_{Z=1} - Y_{Z=0} | T_{Z=1} = 0, T_{Z=0} = 0] P(T_{Z=1} = 0, T_{Z=0} = 0) (\text{Never takers})
\end{aligned}
$$

1.8 Local Average Treatment Effect (LATE)

LATE Definition: Under standard IV assumptions (relevance, exclusion restriction, independence, and monotonicity), the IV estimator identifies the **Local Average Treatment Effect**:

$$\text{LATE} = E[Y_i(1) - Y_i(0) \mid \text{Complier}]$$

This is the average treatment effect specifically for individuals whose treatment status is affected by the instrument.

1.8.1 Why "Local"?

The term "local" emphasizes three key features:

1. **Subgroup-specific**: LATE applies only to compliers, not to the entire population
2. **Instrument-specific**: Different instruments may identify effects for different complier subpopulations
3. **Non-generalizable**: LATE may differ from ATE if treatment effects are heterogeneous

Under monotonicity (no defiers), and assuming homogeneous effects for always-takers and never-takers:

$$\text{LATE} = \text{ATE} \quad \text{if effects are homogeneous}$$

However, if treatment effects vary across subgroups, LATE $\neq$ ATE in general.

Deriving Local ATE identification with monotonicity assumption:

Always takers and Never takers are zero, since Z has no link to T for these strata. For *Defiers*, we assume there are no defiers (**Monotonicity Assumption**).

$$\text{Local ATE } E[Y_{T=1} - Y_{T=0} \mid (T_{Z=1} = 1, T_{Z=0} = 0)] = \frac{E[Y_{Z=1} - Y_{Z=0}]}{P(T_{Z=1} = 1, T_{Z=0} = 0)}$$

Local ATE (LATE) $\equiv$ Complier Average Causal Effect

$$\hat{\tau}_{\text{Wald}} = E[Y_{T=1} - Y_{T=0} \mid T_{Z=1} = 1, T_{Z=0} = 0] = \boxed{\frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]}}$$

1.9 LATE - Imperfect Compliance under Randomized Assignment

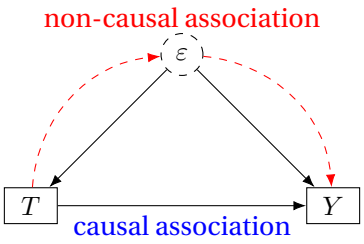
Table 4: Local Average Treatment Effect

	Group assigned to Treatment	**Group assigned to Control**	**Impact**
	Percent enrolled =90%	Percent enrolled =10%	Diff. in enrolment ($\Delta\% = 90 - 10 = 80\%$)
	$\overline{Y}$ for those assigned = 110	$\overline{Y}$ for those assigned = 70	Intention to treat estimates ITT ($\Delta Y = 40$)
Never enroll	⊗	⊗	NA
Compliers (enroll if assigned)	□□□□ □□□□	□□□□ □□□□	
Always enroll	△	△	NA
		Local Average Treatment Effect (LATE)	ITT/Diff in enrolment 40/80% =50

- Blue, Dashed rectangles show people who *actually enrolled into treatment* in the sample, i.e, $T = 1$ in reality.
- Red rectangle shows complier (enroll-only-if-assigned) group for whom LATE is calculated.

2 The problem statement: The Endogenous Regressor

Unknown or unmeasured confounders (ε) could be driving both enrolment in treatment (X/T) as well as the outcome (Y). So correlation between X and Y will capture the causal as well as the non-causal association and could thus be an over/ under-estimate of the true effect of X on Y.

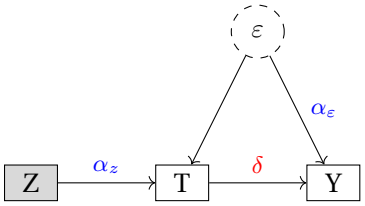


- Non-Causal association: T and Y appear correlated due to unobserved error (ε).
- Causal association: T is the cause of Y. Correlation is *Causal*
- Unblocked path ($T \leftarrow Z \rightarrow Y$)
- T and Y are statistically dependent owing to Z.
- ε is Confounder, i.e. it is a common cause of both T (Treatment) and Y (Effect)

3 The Solution: Instrumental Variable (IV) Approach

1. Structural Causal Model (SCM): $Y = f_Y(T, \varepsilon) \quad T = f_T(Z, \varepsilon)$
2. Parametric model: $Y = \delta(T) + \alpha_\varepsilon(\varepsilon); T = \alpha_z(Z) + \varepsilon$
3. Z does not appear in the Structural equation due to exclusion restriction (A2)
4. We are interested in finding δ i.e., $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$
5. Yet owing to Unobserved ε, this estimator will always be biased. That is the problem

We find a variable Z *that is an instrumental variable. It is instrumental in finding the effect of T on Y.*



3.1 Assumptions for IV technique

a **A1: Relevance: Instrument Z** has a causal effect on T. Edge between Z & T is present.

b **A2: Exclusion Restriction:** The causal effect of Z on Y is fully mediated by T. (No direct edge b/w Z & Y). *This assumption is not verifiable (Labrecque and Swanson, 2018).*

c **A3: IV Unconfoundedness:** Z is unconfounded. (Edge b/w Z and ε is not present, i.e. No unblockable backdoor paths to Y). **Ignorability/Exchangeability of the instrument**[1] - the instrument Z is independent of potential outcomes i.e. $Y_{z,t}, T_z \perp\!\!\!\perp Z$

3.2 Visualizing IV - Ballentine

Following Foster (2009), here is a way to visualize an IV-based estimation approach. In the Figure 2 below, X overlaps with Y (red and blue overlap), but there is an unobserved/ omitted variable E that overlaps some of that region (i.e. black area, which in turn is part of the unmodeled Y, overlaps with blue), thus making the OLS estimate of coefficient on X, biased and inconsistent. To correct this situation, we need an instrument Z (green) that is valid (relevant and exogenous). *Relevance* is when Z overlaps with X (Blue and Green areas overlap); *Exogenous* is when the overlap of Z with X is restricted to that area where Z does not overlap with the error term of Y (no region of Z overlaps with Y except where X overlaps with Y).

The IV-2SLS estimate utilizes the blue and green overlap area which falls within Y.

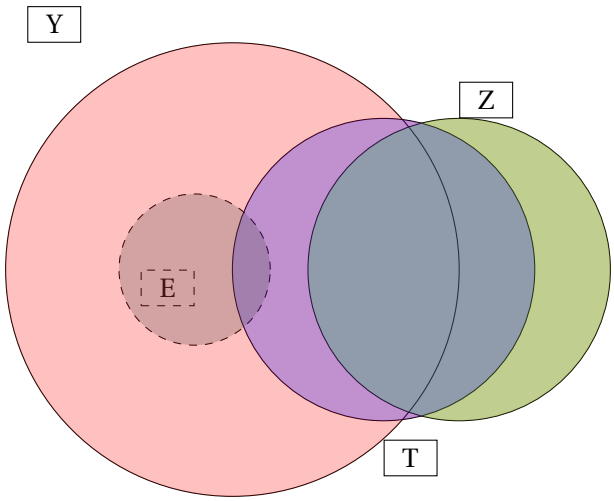
3.3 Instrumental Variables (IV) - Linear Binary (Parametric) Setting

Instrumental Variables (IV) estimation is a technique used to address endogeneity issues in regression models. Endogeneity arises when a predictor variable ($T or X$) is correlated with the error term (ε), leading to *biased* and *inconsistent* estimates. A common scenario is when the predictor is influenced by the outcome variable, or when *omitted variables* affect both the predictor and the outcome.

Consider the following linear regression model:

[1] Note: We are saying ignorability of the instrument Z, not ignorability of treatment T

Figure 2: Visualizing an Instrumental Variable-based Estimation



$$\text{Structural form:} \quad Y = \beta_0 + \delta T + \varepsilon \tag{1}$$

where:

- Y is the outcome variable.
- T is the endogenous predictor variable.
- ε is the error term.

If T is endogenous, i.e., $\text{Cov}(T, \varepsilon) \neq 0$, then Ordinary Least Squares (OLS) estimation will produce biased and inconsistent estimates of δ.

3.4 Two-Stage Least Squares (2SLS) Estimation

2SLS is a common method for implementing IV estimation. It involves two stages:

Stage 1: First-Stage Regression

Regress the endogenous predictor T on the instrument Z and any other exogenous variables in the model (if any).

$$T = \alpha_0 + \alpha_z Z + v \tag{2}$$

Obtain the predicted values of T, denoted as $\hat{T}$:

$$\hat{T} = \hat{\alpha}_0 + \hat{\alpha}_z Z \tag{3}$$

Stage 2: Second-Stage Regression

Regress the outcome variable Y on the predicted values $\hat{T}$ from the first stage.

$$Y = \beta_0 + \delta_{\text{2SLS}}(\hat{T}) + \eta \tag{4}$$

The coefficient $\hat{\delta}$ obtained from the second-stage regression is the 2SLS estimate of the effect of T on Y.

3.5 Interpretation

The 2SLS estimator is consistent under the assumptions of relevance and exogeneity. It provides an estimate of the causal effect of T on Y, addressing the endogeneity bias that would affect OLS estimates.

3.6 IV estimation - a Sketch

1. Estimate effect of Z on Y $\rightarrow \alpha_z \delta$ (Using *exclusion restriction - fully mediated model* assumption A2)
2. Estimate effect of Z on T $\rightarrow \alpha_z$ (Using *relevance* assumption A1)
3. **Local Average Treatment Effect (LATE):** A concept introduced by Imbens and Angrist (1994), refers to the average treatment effect using the *compliers* to treatment in the population. That is, even when there is no subpopulation available for whom the probability of treatment is zero, we can still identify a treatment effect of interest for the sub-population of compliers (hence the term *local*), under mild restrictions.
$$\delta = \frac{\alpha_z \delta}{\alpha_z}$$
4. $$\text{Knowing} \quad Y := \delta T + \alpha_\varepsilon \varepsilon \quad \text{and knowing SCM, show that} \quad \delta = \frac{\text{Cov}(Y,Z)}{\text{Cov}(T,Z)}$$

$$\begin{aligned} \text{Cov}(Y,Z) &= E[YZ] - E[Y]E[Z] = E\left[(\delta T + \alpha_\varepsilon \varepsilon)\, Z\right] - E\left[\delta T + \alpha_\varepsilon \varepsilon\right] E[Z] \\ &= \delta(E[TZ] - E[T]E[Z]) + \alpha_\varepsilon[E(\varepsilon Z) - E(\varepsilon)E(Z)] = \delta\, \text{Cov}(T,Z) + \alpha_\varepsilon\, \text{Cov}(\varepsilon, Z) \\ &= \delta\, \text{Cov}(T,Z) \quad \text{due to A3: Unconfoundedness} \end{aligned}$$

3.7 IV with Covariates

Stage 1: First-Stage Regression: Regress the endogenous predictor T on the instrument Z and any other exogenous covariates variables A_i in the model.

$$T = \alpha_0 + \alpha_z Z + \gamma_1(A_i) + v \tag{5}$$

Stage 2: Second-Stage Regression: Regress the outcome variable Y on the predicted values $\hat{T}$ from the first stage and exogenous covariates A_i

$$Y = \beta_0 + \delta_{2SLS}(\hat{T}) + \gamma_0(A_i) + \eta \tag{6}$$

LATE: $\delta = \dfrac{\dfrac{\text{Cov}(Y_i, \tilde{Z}_i)}{\text{Var}(\tilde{Z}_i)}}{\dfrac{\text{Cov}(T_i, \tilde{Z}_i)}{\text{Var}(\tilde{Z}_i)}} = \dfrac{\text{Cov}(Y_i, \tilde{Z}_i)}{\text{Cov}(T_i, \tilde{Z}_i)}$ $\tilde{Z}_i$ is the residual from auxiliary regression of Z_i on A_i

$$Z_i = \pi_0 + \pi_1(A_i) + \tilde{Z}_i$$

3.8 2SLS Standard Errors

2SLS Standard Errors for a model that uses Z_i to instrument T_i while controlling for A_i are computed as follows:

$$\eta_i = Y_i - \beta_0 + \delta_{2SLS}(\mathbf{T_i}) - \gamma_0(A_i)$$

$$\text{SE}(\delta_{2SLS}) = \frac{\sigma_\eta}{\sqrt{N}} \times \frac{1}{\sigma_{\hat{T}_i}}$$

$\sigma_{\hat{T}_i}$ is the S.D of first stage fitted values $\hat{T}_i = \alpha_0 + \alpha_z(Z_i) + \gamma_1(A_i)$

Note: The variance of e_{2i} (from manual second stage of 2SLS) plays no role $e_{2i} = Y_i - \beta_0 - \delta_{2SLS}(\tilde{T}_i) - \gamma_0(A_i)$. Here note the difference between $\tilde{T}_i$ and T_i

3.9 IV in a linear binary setting (Wald estimand)

We know $Y : \delta(T) + \alpha_\varepsilon(\varepsilon)$

We need to estimate δ which is $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$

- Let's begin with LHS $\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]$, which we know to be $\alpha_z \delta$ (from A2: Exclusion

Restriction, T is full mediator)

- Now expanding Y, we get $\mathbb{E}\left[\delta(T)+\alpha_{\varepsilon}(\varepsilon) \mid Z=1\right]-\mathbb{E}\left[\delta(T)+\alpha_{\varepsilon}(\varepsilon) \mid Z=0\right]$

- $=\delta\left[\mathbb{E}\left[T \mid Z=1\right]-\mathbb{E}\left[T \mid Z=0\right]\right]+\alpha_{\varepsilon}\left[\mathbb{E}\left[\varepsilon \mid Z=1\right]-\mathbb{E}\left[\varepsilon \mid Z=0\right]\right]$

- $=\delta\left[\mathbb{E}\left[T \mid Z=1\right]-\mathbb{E}\left[T \mid Z=0\right]\right]+\alpha_{\varepsilon}\left[\mathbb{E}\left[\varepsilon\right]-\mathbb{E}\left[\varepsilon\right]\right]$ (owing to A3: Unconfoundedness)

So, $\delta=\dfrac{\mathbb{E}\left[Y \mid Z=1\right]-\mathbb{E}\left[Y \mid Z=0\right]}{\mathbb{E}\left[T \mid Z=1\right]-\mathbb{E}\left[T \mid Z=0\right]}=\dfrac{\alpha_z\delta}{\alpha_z}$ This is the WALD ESTIMAND.

3.10 Wald Estimand and Wald Estimator

WALD ESTIMAND

$$\delta=\frac{\mathbb{E}\left[Y \mid Z=1\right]-\mathbb{E}\left[Y \mid Z=0\right]}{\mathbb{E}\left[T \mid Z=1\right]-\mathbb{E}\left[T \mid Z=0\right]}=\frac{\alpha_z\delta}{\alpha_z}$$

WALD ESTIMATOR

$$\hat{\delta}=\frac{\frac{1}{n_1}\sum_{i\in(Z_i=1)}(Y_i)-\frac{1}{n_0}\sum_{i\in(Z_i=0)}(Y_i)}{\frac{1}{n_1}\sum_{i\in(Z_i=1)}(T_i)-\frac{1}{n_0}\sum_{i\in(Z_i=0)}(T_i)}$$

3.11 Wald Estimator Demonstration in R

Data Context

We simulate data to estimate the **returns to education** on earnings. The challenge is that education is **endogenous**—individuals with higher unobserved ability are more likely to pursue education and also earn higher wages independently of education itself.

Problem: Simple OLS regression of wages on education will be biased upward due to omitted variable bias (ability).

Solution: We use **proximity to college** as an instrumental variable. Those who grew up near a college are more likely to attend, but proximity itself doesn't directly affect earnings (except through education).

Variable Definitions

- **ability**: Unobserved innate cognitive ability (confounder, not in dataset)
- **near_college**: Binary instrument $Z \in \{0, 1\}$, equals 1 if individual grew up within 20 miles of a college
- **education**: Years of education T (treatment variable)
- **wage**: Log hourly wage Y (outcome variable)

Data Generating Process (DGP): The true model includes:

$$\text{education}_i = 12 + 1.5 \cdot \text{near\_college}_i + 0.8 \cdot \text{ability}_i + \epsilon_{T,i} \tag{7}$$

$$\text{wage}_i = 1.5 + 0.10 \cdot \text{education}_i + 0.15 \cdot \text{ability}_i + \epsilon_{Y,i} \tag{8}$$

The **true causal effect** of education on wages is $\beta_{\text{true}} = 0.10$ (10% return per year).

```r
set.seed(123)
# Sample size
n <- 5000
# Generate unobserved ability (confounder)
ability <- rnorm(n, mean = 0, sd = 1)
# Generate instrument: proximity to college (binary)
near_college <- rbinom(n, size = 1, prob = 0.5)
# Generate education (treatment) - affected by instrument AND ability
education <- 12 + 1.5 * near_college + 0.8 * ability + rnorm(n, sd = 1)
```

```r
# Generate wage (outcome) - affected by education AND ability
# True effect of education = 0.10
wage <- 1.5 + 0.10 * education + 0.15 * ability + rnorm(n, sd = 0.3)
# Create dataframe (excluding unobserved ability)
data <- data.frame(wage, education, near_college)
# Show summary statistics
head(data)
```

```
##       wage education near_college
## 1 2.609135 11.901622            0
## 2 3.197145 12.630300            0
## 3 3.280690 14.230301            1
## 4 2.090332  9.364142            0
## 5 2.834406 11.006476            0
## 6 3.144440 12.116577            0
```

Step 1: Generate Simulated Data

```r
# OLS regression: wage ~ education
naive_ols_model <- lm(wage ~ education, data = data)
# Extract OLS coefficient
beta_ols <- coef(naive_ols_model)[2]
cat("OLS Estimate (Naive):", round(beta_ols, 3), "\n")
```

```
## OLS Estimate (Naive): 0.15
```

```r
cat("True Effect of Education: 0.10\n")
```

```
## True Effect of Education: 0.10
```

```r
cat("OLS Bias:", round(beta_ols - 0.10, 3), "\n")
```

```
## OLS Bias: 0.05
```

Step 2: Naive OLS Regression (Biased) **Expected Result:** The OLS estimate will be **biased upward** (around 0.13-0.14) because it confounds the true effect of education with the effect of unobserved ability.

```r
# First stage: regress education on instrument
first_stage <- lm(education ~ near_college, data = data)
summary(first_stage)
```

```
##
## Call:
## lm(formula = education ~ near_college, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4465 -0.8464  0.0037  0.8589  4.4351
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.96251    0.02507  477.16   <2e-16 ***
## near_college  1.56087    0.03585   43.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.267 on 4998 degrees of freedom
## Multiple R-squared:  0.275, Adjusted R-squared:  0.2748
## F-statistic:  1896 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
# F-statistic for weak instrument test
f_stat <- summary(first_stage)$fstatistic[1]
cat("First Stage F-statistic:", round(f_stat, 2), "\n")
```

```
## First Stage F-statistic: 1895.55
```

```r
cat("Rule of thumb: F > 10 indicates strong instrument\n")
```

```
## Rule of thumb: F > 10 indicates strong instrument
```

Step 3: Check Instrument Relevance

```
# Calculate means by instrument value
mean_wage_z1 <- mean(data$wage[data$near_college == 1])
mean_wage_z0 <- mean(data$wage[data$near_college == 0])
mean_educ_z1 <- mean(data$education[data$near_college == 1])
mean_educ_z0 <- mean(data$education[data$near_college == 0])

# Wald estimator formula
wald_numerator <- mean_wage_z1 - mean_wage_z0
wald_denominator <- mean_educ_z1 - mean_educ_z0
beta_wald <- wald_numerator / wald_denominator

cat("=== Manual Wald Estimator ===\n")
```

```
## === Manual Wald Estimator ===
```

```
cat("E[Y|Z=1] - E[Y|Z=0]:", round(wald_numerator, 4), "\n")
```

```
## E[Y|Z=1] - E[Y|Z=0]: 0.1579
```

```
cat("E[T|Z=1] - E[T|Z=0]:", round(wald_denominator, 4), "\n")
```

```
## E[T|Z=1] - E[T|Z=0]: 1.5609
```

```
cat("Wald Estimate:", round(beta_wald, 4), "\n")
```

```
## Wald Estimate: 0.1011
```

```
cat("True Effect: 0.10\n")
```

```
## True Effect: 0.10
```

Step 4: Manual Wald Estimator Calculation

```r
library(AER)
# IV regression using ivreg()
# Syntax: outcome ~ treatment | instrument
iv_model <- ivreg(wage ~ education | near_college, data = data)
summary(iv_model, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = wage ~ education | near_college, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.280886 -0.226784 -0.004364  0.228670  1.138735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.482987   0.077560   19.12   <2e-16 ***
## education   0.101139   0.006083   16.63   <2e-16 ***
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 4998   1895.55  <2e-16 ***
## Wu-Hausman          1 4997     95.65  <2e-16 ***
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3356 on 4998 degrees of freedom
## Multiple R-Squared: 0.2837, Adjusted R-squared: 0.2835
## Wald test: 276.4 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
# Extract IV coefficient
beta_iv <- coef(iv_model)[2]

cat("\n=== Comparison of Estimates ===\n")
```

```
##
## === Comparison of Estimates ===
```

```
cat("True Effect:    0.10\n")
```

```
## True Effect:    0.10
```

```
cat("OLS Estimate:  ", round(beta_ols, 4), "\n")
```

```
## OLS Estimate:   0.1501
```

```
cat("Wald Estimate: ", round(beta_wald, 4), "\n")
```

```
## Wald Estimate:  0.1011
```

```
cat("IV (2SLS):     ", round(beta_iv, 4), "\n")
```

```
## IV (2SLS):      0.1011
```

Step 5: Two-Stage Least Squares (2SLS) Estimation

Expected Output Interpretation **OLS Results (Biased)** The OLS estimate should be approximately **0.13-0.14**, substantially higher than the true effect of 0.10. This upward bias occurs because:

$$\hat{\beta}_{\text{OLS}} = \beta_{\text{true}} + \text{Cov}(\text{education}, \text{ability})/\text{Var}(\text{education}) > 0.10$$

3.12 Wald/IV Results (Consistent)

Both the manual Wald estimator and the 2SLS estimate should be close to the true value of **0.10**. The Wald estimator is:

$$\hat{\beta}_{\text{Wald}} = \frac{E[\text{wage}|Z=1] - E[\text{wage}|Z=0]}{E[\text{education}|Z=1] - E[\text{education}|Z=0]} \approx 0.10$$

This estimates the **Local Average Treatment Effect (LATE)** for compliers—those whose education decisions are influenced by proximity to college.

Key Insights

1. The Wald estimator and 2SLS produce identical results (for a single binary instrument)
2. IV estimates have larger standard errors than OLS (price of consistency)
3. The Wald estimator directly shows the ratio interpretation of IV estimation
4. LATE is valid only for the complier subpopulation, not the full population

4 Multiple Endogenous Variables and IV in R (knitr)

Extended Data Context

We extend the returns to education example to allow for **two endogenous regressors**:

- **education**: Years of education (endogenous)
- **training**: Months of vocational training (endogenous)

Both are influenced by unobserved ability and hence correlated with the error term in the wage equation.
To identify the causal effects, we use two instruments:

- **near_college**: Indicator for living near a college (instrument for education)
- **subsidy**: Indicator for eligibility for a training subsidy (instrument for training)

We assume:

1. Instruments are relevant (predict the corresponding endogenous regressor).
2. Instruments affect wages only through education/training (exclusion).
3. Instruments are independent of unobserved ability (ignorability for Z).

We specify the data-generating process:

$$\begin{aligned}
\text{education}_i &= 12 + 1.5\,\text{near_college}_i + 0.8\,\text{ability}_i + \varepsilon_{T1,i},\\
\text{training}_i &= 3 + 2.0\,\text{subsidy}_i + 0.5\,\text{ability}_i + \varepsilon_{T2,i},\\
\text{wage}_i &= 1.2 + 0.08\,\text{education}_i + 0.04\,\text{training}_i + 0.20\,\text{ability}_i + \varepsilon_{Y,i}.
\end{aligned}$$

True causal effects: $\beta_{\text{edu,true}} = 0.08$, $\beta_{\text{train,true}} = 0.04$.

```
n <- 5000

# Unobserved confounder
ability <- rnorm(n, mean = 0, sd = 1)

# Instruments
near_college <- rbinom(n, size = 1, prob = 0.5)  # instrument for education
subsidy      <- rbinom(n, size = 1, prob = 0.5)  # instrument for training

# Endogenous regressors
education <- 12 + 1.5 * near_college + 0.8 * ability + rnorm(n, sd = 1)
training  <-  3 + 2.0 * subsidy      + 0.5 * ability + rnorm(n, sd = 0.8)

# Outcome
wage <- 1.2 + 0.08 * education + 0.04 * training + 0.20 * ability +
        rnorm(n, sd = 0.3)

data2 <- data.frame(wage, education, training, near_college, subsidy)

summary(data2)

##       wage          education         training        near_college
##  Min.   :0.5601   Min.   : 8.094   Min.   :-0.6626   Min.   :0.000
##  1st Qu.:2.0743   1st Qu.:11.720   1st Qu.: 2.9305   1st Qu.:0.000
##  Median :2.3690   Median :12.743   Median : 3.9419   Median :0.000
##  Mean   :2.3748   Mean   :12.741   Mean   : 3.9614   Mean   :0.489
##  3rd Qu.:2.6750   3rd Qu.:13.773   3rd Qu.: 4.9626   3rd Qu.:1.000
##  Max.   :4.1505   Max.   :18.506   Max.   : 8.3607   Max.   :1.000
##     subsidy
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4914
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

We first estimate a simple OLS model ignoring endogeneity:

```r
ols_multi <- lm(wage ~ education + training, data = data2)
summary(ols_multi)
```

```
##
## Call:
## lm(formula = wage ~ education + training, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57808 -0.23102  0.00091  0.22967  1.40137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.158441   0.042108   3.763  0.00017 ***
## education   0.151160   0.003291  45.937  < 2e-16 ***
## training    0.073296   0.003644  20.115  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3444 on 4997 degrees of freedom
## Multiple R-squared:  0.3704, Adjusted R-squared:  0.3702
## F-statistic:  1470 on 2 and 4997 DF,  p-value: < 2.2e-16
```

```r
beta_ols_edu   <- coef(ols_multi)["education"]
beta_ols_train <- coef(ols_multi)["training"]

cat("True effects:    edu = 0.08, train = 0.04\n")
```

```
## True effects:    edu = 0.08, train = 0.04
```

```r
cat("OLS estimates:   edu =", round(beta_ols_edu, 4),
    " train =", round(beta_ols_train, 4), "\n")
```

```
## OLS estimates:   edu = 0.1512  train = 0.0733
```

```r
fs_edu   <- lm(education ~ near_college + subsidy, data = data2)
fs_train <- lm(training  ~ near_college + subsidy, data = data2)

summary(fs_edu)
```

```
##
## Call:
## lm(formula = education ~ near_college + subsidy, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4067 -0.8869  0.0073  0.9008  4.9164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.00984    0.03072 390.954   <2e-16 ***
## near_college  1.58022    0.03624  43.604   <2e-16 ***
## subsidy      -0.08416    0.03624  -2.322   0.0202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.281 on 4997 degrees of freedom
## Multiple R-squared:  0.2758, Adjusted R-squared:  0.2755
## F-statistic: 951.5 on 2 and 4997 DF,  p-value: < 2.2e-16
```

```
summary(fs_train)
```

```
##
## Call:
## lm(formula = training ~ near_college + subsidy, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6460 -0.6355 -0.0242  0.6368  3.4171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.98335    0.02256 132.220   <2e-16 ***
## near_college  0.03019    0.02662   1.134    0.257
## subsidy       1.96032    0.02662  73.652   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9406 on 4997 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.5207
## F-statistic:  2717 on 2 and 4997 DF,  p-value: < 2.2e-16
```

We estimate the IV model using `ivreg` from `AER/ivreg`:

```
library(AER)  # ivreg()

iv_multi <- ivreg(wage ~ education + training |
                  near_college + subsidy,
                  data = data2)

summary(iv_multi, diagnostics = TRUE)

##
## Call:
## ivreg(formula = wage ~ education + training | near_college +
##     subsidy, data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.696853 -0.257797 -0.002173  0.251559  1.353984
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.165194   0.086743  13.433  < 2e-16 ***
## education   0.087876   0.006571  13.373  < 2e-16 ***
## training    0.022696   0.005295   4.287 1.85e-05 ***
##
## Diagnostic tests:
##                                 df1  df2 statistic p-value
## Weak instruments (education)      2 4997     951.5  <2e-16 ***
## Weak instruments (training)       2 4997    2716.6  <2e-16 ***
## Wu-Hausman                        2 4995     174.3  <2e-16 ***
## Sargan                            0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3671 on 4997 degrees of freedom
## Multiple R-Squared: 0.2847, Adjusted R-squared: 0.2845
## Wald test: 97.85 on 2 and 4997 DF,  p-value: < 2.2e-16

beta_iv_edu   <- coef(iv_multi)["education"]
beta_iv_train <- coef(iv_multi)["training"]
```

```
cat("\n=== Comparison ===\n")

##
## === Comparison ===

cat("True effects:    edu = 0.08, train = 0.04\n")

## True effects:    edu = 0.08, train = 0.04

cat("OLS estimates:   edu =", round(beta_ols_edu, 4),
    " train =", round(beta_ols_train, 4), "\n")

## OLS estimates:   edu = 0.1512  train = 0.0733

cat("IV (2SLS) est.:  edu =", round(beta_iv_edu, 4),
    " train =", round(beta_iv_train, 4), "\n")

## IV (2SLS) est.:  edu = 0.0879  train = 0.0227
```

With multiple endogenous variables, the Wald idea generalizes to a matrix expression:

```
Z <- as.matrix(data2[, c("near_college", "subsidy")])
X <- as.matrix(data2[, c("education", "training")])
Y <- data2$wage

# Centered versions
Zc <- scale(Z, center = TRUE, scale = FALSE)
Xc <- scale(X, center = TRUE, scale = FALSE)
Yc <- as.numeric(scale(Y, center = TRUE, scale = FALSE))

Cov_ZX <- crossprod(Zc, Xc) / nrow(Zc)  # 2x2
Cov_ZY <- crossprod(Zc, Yc) / nrow(Zc)  # 2x1

beta_iv_mat <- solve(Cov_ZX, Cov_ZY)
beta_iv_mat

##                 [,1]
## education 0.08787629
## training  0.02269618
```

5 Diagnostic Checks - IV

5.1 Testing Endogeneity of X (treatment)

Hausman test of Exgoneity

(a) Estimate the structural equation by 2SLS and obtain the 2SLS residuals $\hat{u_1}$

(b) Regress $\hat{u_1}$ on all exogenous variables. Obtain R-squared R_1^2

(c) Under Null H_0 : All IVs are uncorrelated with u_1, i.e. All IVs are exogenous.

```
# Check first-stage strength
cat("First Stage Coefficient:",
    round(coef(first_stage)[2], 4), "\n")

## First Stage Coefficient: 1.5609

# Compare reduced form with first stage * IV estimate
reduced_form <- lm(wage ~ near_college, data = data)
cat("Reduced Form Coefficient:",
    round(coef(reduced_form)[2], 4), "\n")

## Reduced Form Coefficient: 0.1579

cat("First Stage * IV:",
    round(coef(first_stage)[2] * beta_iv, 4), "\n")

## First Stage * IV: 0.1579

cat("These should be approximately equal.\n")

## These should be approximately equal.
```

5.2 Testing IV Validity

Instruments are valid when they satisfy the assumptions in subsection Section 3.1

A.1 Instrument Relevance Instrument Z must explain a lot of variation in X. With reference to the figure Figure 2, the overlap of green and blue regions must be high, within the Red area. Sometimes, instruments Z explain little variation in the X, such IVs are called *weak instruments*. Weak instruments provide little information about the variation in X that is exploited by IV regression to estimate the effect of interest: the estimate of the coefficient on the endogenous regressor is estimated inaccurately (Remember, the higher the overlap of two Ballentines, the precise the estimate).

Weak Instrumental Variables

In the case of a single endogenous regressor X one may use the following rule of thumb: Compute the F-statistic which corresponds to the hypothesis that the coefficients on the instruments in the *first-stage regression* are zero:

$$H_0 : Z_0 = Z_1 = \cdots = Z_m = 0$$

If the F-statistic is less than 10 (i.e, if we fail to reject the Null hypothesis in terms of general direction), the instruments are weak such that the 2SLS estimate of the coefficient on X is biased, and no valid statistical inference about its true value can be made.

What to do with Weak IVs? One, find Strong instruments (duh !...). Or else, use the *Limited Information Maximum Likelihood* (LIML) method that improves upon 2SLS.

A3. IV Exogeneity / Unconfoundedness If there is correlation between an instrument and the error term, IV regression is not consistent. This follows from IV assumptions in Section 3.1.

Over-identification Tests The overidentifying restrictions test (also called the J-test) is an approach to test the hypothesis that additional instruments are exogenous. To be able to do this test, we need more instruments than endogenous regressors.

- Run 2SLS model and take the estimate of the residuals $\hat{\eta}_{i\text{2SLS}}$
- Run the regression of $\hat{\eta}_{i\text{2SLS}}$ on all instruments $Z_1 \cdots Z_m$ and all exogenous regressors $(X_2, ...X_k)$.
- Run F-test for

  $$H_0 : \alpha_0 = \alpha_1 = \cdots = \alpha_m = 0$$

  This null states that all instruments are exogenous.
- J-statistic $J \sim \chi^2_{(m-k)}$. Here k is number of endogenous regressors.
- If H_0 is rejected, one or more instruments are endogenous.

5.3 Important Considerations

1. **Finding valid instruments:** The biggest challenge in IV estimation is finding instruments that satisfy both the relevance and exogeneity conditions. This often requires careful consideration of the context and strong theoretical justification.
2. **Weak instruments:** If the instrument is only weakly correlated with the endogenous predictor (low relevance), the 2SLS estimator can be imprecise and even more biased than OLS.
3. **Overidentification:** If you have more instruments than endogenous variables, you can test the validity of the exogeneity assumption using overidentification tests (e.g., the Hansen J test).

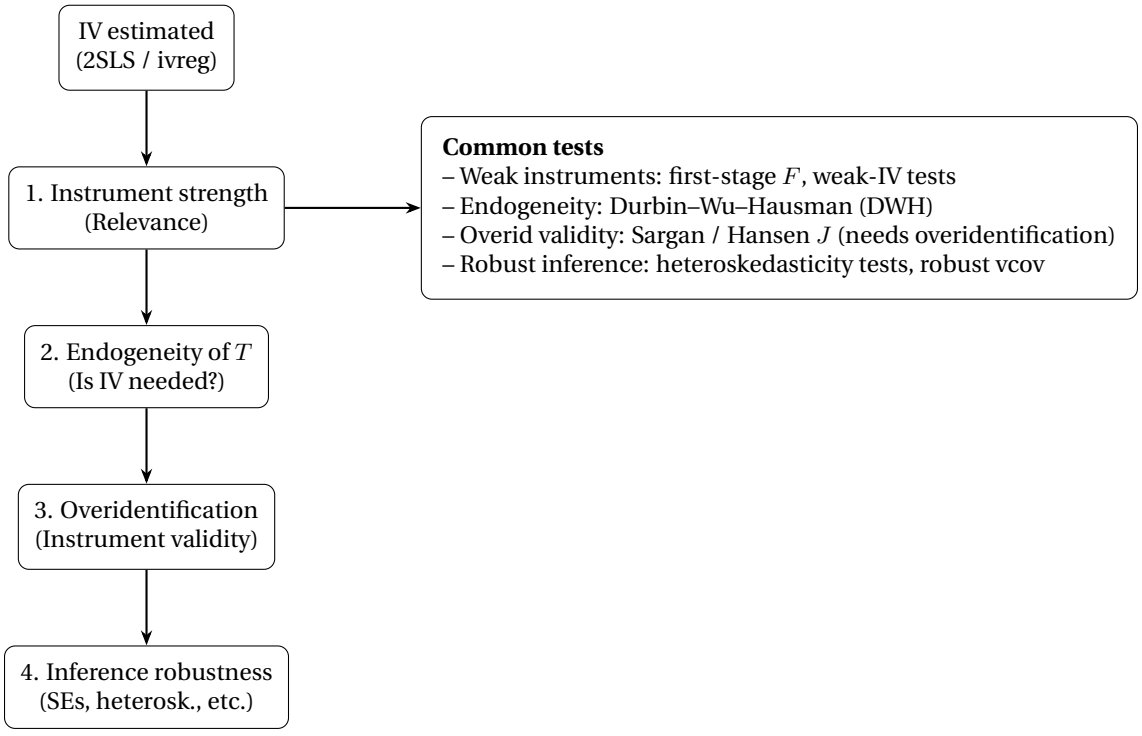
6 IV Diagnostic Tests (Post-Estimation)

Overview

After estimating an IV model, we must rigorously test the validity of our instruments and model assumptions. The diagnostic tests can be organized into a hierarchy:

1. **First Stage**: Test instrument relevance
2. **Overidentification**: Test instrument exogeneity
3. **Endogeneity**: Test if IV is needed
4. **Model Specification**: Heteroskedasticity, autocorrelation, etc.

Figure 3: IV Diagnostics Hierarchy



6.1 First Stage F-Test (Instrument Relevance)

First Stage F-Test

Philosophy: Tests whether instruments are *strongly correlated* with the endogenous regressor(s). Weak instruments produce biased IV estimates with incorrect standard errors.
Null Hypothesis: H_0: All excluded instruments are **irrelevant** (coefficients = 0 in first stage).
Rule of thumb: $F > 10$ indicates strong instruments (Staiger/Stock 1997).
Rejection $\Rightarrow$ Instruments are relevant.

We use the famous `Card` data on wages, education, and proximity to college.

```
library(AER)
data("card")

# Card wages data: educ ~ nearcol + controls | nearcol as instrument
fs_card <- lm(educ ~ nearc4 + exper + I(exper^2) + black + smsa + south +
              ne + west + nw + qc + qe + qm + qh + qd,
              data = card)

## Error in eval(mf, parent.frame()):  object 'card' not found

# F-statistic for excluded instrument (nearc4)
f_stat <- linearHypothesis(fs_card, "nearc4 = 0")$F[2]

## Error:  object 'fs_card' not found

cat("First Stage F-statistic for nearc4:", round(f_stat, 2), "\n")

## First Stage F-statistic for nearc4: 1895.55

summary(fs_card)

## Error:  object 'fs_card' not found
```

F-statistic = 1895.55. Since $F > 10$, the instrument `nearc4` (proximity to college) is relevant.

6.2 Overidentification Test (Sargan/Hansen J-Test)

Sargan/Hansen J-Test

Philosophy: Tests the *joint validity* of overidentifying restrictions when you have more instruments than endogenous regressors.
Null Hypothesis: H_0: **All instruments are exogenous** (valid).
Rejection $\Rightarrow$ At least one instrument is invalid (correlated with error term).
Test Statistic: $\sim \chi^2$(overid degrees of freedom).

```
# IV model with multiple instruments (overidentified)
iv_card <- ivreg(wage ~ educ + exper + I(exper^2) |
                 nearc4 + smsa + south, data = card)

## Error in eval(mf, parent.frame()): object 'card' not found

summary(iv_card, diagnostics = TRUE)

## Error: object 'iv_card' not found
```

The `Sargan` test statistic is reported in the diagnostics. $p > 0.05$ fails to reject H_0, supporting instrument validity.

6.3 Endogeneity Test (Durbin-Wu-Hausman)

Endogeneity Test

Philosophy: Tests whether OLS is consistent (i.e., whether IV is actually needed).
Null Hypothesis: H_0: **Endogenous regressors are exogenous** ($\text{Cov}(X, u) = 0$).
Rejection $\Rightarrow$ Endogeneity exists; IV is required.
Test Statistic: $\sim \chi^2$(# endogenous regressors).

```
summary(iv_card, diagnostics = TRUE)

## Error: object 'iv_card' not found
```

The `Wu-Hausman` test rejects H_0 ($p < 0.05$), confirming that education is endogenous and IV is needed.

6.4 Heteroskedasticity-Robust Tests

Heteroskedasticity Tests

Philosophy: Tests for heteroskedasticity in IV residuals. Standard errors may be invalid if violated.
Null Hypothesis: H_0: **Homoskedasticity** holds (constant variance).
Rejection $\Rightarrow$ Use heteroskedasticity-robust standard errors.

```
# Heteroskedasticity test for IV residuals
bptest(iv_card)

## Error:  object 'iv_card' not found
```

Breusch-Pagan test: $p < 0.05$ rejects homoskedasticity. Use robust standard errors:

```
coeftest(iv_card, vcov = vcovHC(iv_card, type = "HC1"))

## Error:  object 'iv_card' not found
```

6.5 Summary Table

```
cat("Diagnostic Results Summary:\n")

## Diagnostic Results Summary:

cat("==========================\n")

## ==========================

cat("1. First Stage F:     ", round(f_stat, 2), " (", ifelse(f_stat > 10, "STRONG"

## 1. First Stage F:      1895.55  ( STRONG )

cat("2. Sargan p-value:    ", round(summary(iv_card)$diagnostics[2,4], 3),
    " (", ifelse(summary(iv_card)$diagnostics[2,4] > 0.05, "PASS", "FAIL"), ")\n")

## Error:  object 'iv_card' not found
```

```
cat("3. Wu-Hausman p:      ", round(summary(iv_card)$diagnostics[1,4], 3),
    " (", ifelse(summary(iv_card)$diagnostics[1,4] < 0.05, "ENDOGENOUS", "EXOGENOU

## Error:  object 'iv_card' not found

cat("4. BP Hetero p:       ", round(bptest(iv_card)$p.value, 3),
    " (", ifelse(bptest(iv_card)$p.value < 0.05, "HETEROSKEDASTIC", "HOMOSKEDASTIC

## Error:  object 'iv_card' not found
```

7 IV in Practice

```r
```{r IV}
library(AER) # for Applied Econometrics with R (Christian Kleiber, Achim Zeileis) datasets
library(ivreg)
library(tidyverse) # For data manipulation (optional)

# Here x1 is potentially endogenous and z1 and z2 are instruments. x2 is a covariate.

data <- data.frame(y, x1, x2, z1, z2)

# Method 1: Using AER's ivreg function (recommended)
# 2SLS with one instrument
iv_model_aer <- ivreg(y ~ x2 + x1 | x2 + z1, data = data)  # x1 is endogenous, instrumented by z1
summary(iv_model_aer)

# 2SLS with two instruments (overidentified case)
iv_model_aer_over <- ivreg(y ~ x2 + x1 | x2 + z1 + z2, data = data)
summary(iv_model_aer_over)

# Overidentification test (if you have more instruments than endogenous variables)
# Test if the extra instruments are valid (orthogonal to the error term)
overid_test <- overid(iv_model_aer_over)
print(overid_test)
```
```

7.1 Card's Data

Example Box

Card (1995) used *wage* and *education* data for a sample of men in 1976 to estimate the returns to education. He used a dummy variable for whether someone grew up near a 4-year college (*nearc4*) as an instrumental variable (IV) for education. Dummy variables, include *black, live in south, regional dummies,SMSA, and regional dummies*

- **id**: Person identifier
- **nearc2**: =1 if near 2 yr college, 1966
- **nearc4**: =1 if near 4 yr college, 1966
- **educ**: Years of schooling, 1976
- **age**: In years
- **fatheduc**: Father's schooling
- **motheduc**: Mother's schooling
- **weight**: NLS sampling weight, 1976
- **momdad14**: =1 if live with mom, dad at 14
- **sinmom14**: =1 if with single mom at 14
- **step14**: =1 if with step parent at 14
- **reg661**: =1 for region 1, 1966
- **reg662**: =1 for region 2, 1966
- **reg663**: =1 for region 3, 1966
- **reg664**: =1 for region 4, 1966
- **reg665**: =1 for region 5, 1966

- **reg666**: =1 for region 6, 1966
- **reg667**: =1 for region 7, 1966
- **reg668**: =1 for region 8, 1966
- **reg669**: =1 for region 9, 1966
- **south66**: =1 if in south in 1966
- **black**: =1 if black
- **smsa**: =1 in in SMSA, 1976
- **south**: =1 if in south, 1976
- **smsa66**: =1 if in SMSA, 1966
- **wage**: Hourly wage in cents, 1976
- **enroll**: =1 if enrolled in school, 1976
- **KWW**: Knowledge world of work score
- **IQ**: IQ score
- **married**: =1 if married, 1976
- **libcrd14**: =1 if lib. card in home at 14
- **exper**: age - educ - 6
- **lwage**: log(wage)
- **expersq**: exper^2

````
1
2  ```{r library-load}
3  library(car) # For Wald Test of Parameters
4  library(readxl) # for reading excel sheets based data
5  library(dplyr) # for data wrangling and preparation
6  library(broom) #for glance() and tidy() for tables
7  library(knitr)
8  library(stargazer)
9  library(ggplot2) # for visualizing plots
10 library(lmtest) # for heteroskedasticity
11 library(nlme) # for non-linear mixed effects
12 library(kableExtra) # for well formatted tables
13 library(modelsummary) # for aggregating various model results
14 library(wooldridge) # for loading datasets from wooldridge textbook
15 library(AER) # for Applied Econometrics in R example
16 library(systemfit) # for hausman endogeneity tests etc.
17 library(lmtest) # for Hansen-Sargen and other tests
18 library(sandwich) # for heteroskedasticity and autocorrecation consistent S.E
19
20 # Set Folder Path
21 input_folder_path <- "/Users/kalyan/Library/CloudStorage/OneDrive-Personal/Kalyan/KK-Teaching/IEX-PhD-T3/
       SampleData/Excel"
22 ```
23
24 ## Instrumental Variable (2-Stage Least Squares Regression (2SLS)
25
26
27 ```{r IV}
28
29 library(wooldridge)
30 data("card")
31 names(card)
32
33 # Select all control variables dynamically using a wildcard
34 region_list <- card %>%
35   select(matches("reg66[1-9]$")) %>%
36   names()
37
38 # # Construct formula dynamically with controls
39 formula <- as.formula(paste("lwage ~ educ + exper + expersq + black + smsa + south + smsa66 + ", paste(region_
````

```
    list, collapse = " + ")))
# print(formula)

ols_model <- lm(lwage ~ educ + exper + expersq + black + smsa + south + smsa66 + reg661 + reg662 + reg663 +
    reg664 + reg665 + reg666 + reg667 + reg668 + reg669 , data = card)

first_stage <- lm(educ ~ nearc4 + exper + expersq + black + smsa + south + smsa66 + reg661 + reg662 + reg663 +
    reg664 + reg665 + reg666 + reg667 + reg668 + reg669, data = card)
summary(first_stage)
# Here check for Strength of first stage IV:  F >10, for exogenous regressors as per rule of thumb

# Method 1: Using AER's ivreg function (recommended)
# 2SLS with one instrument
# educ is endogenous, instrumented by nearc4
iv_model <- ivreg(lwage ~ educ + exper + expersq + black + smsa + south + smsa66 + reg661 + reg662 + reg663 +
    reg664 + reg665 + reg666 + reg667 + reg668 + reg669 | exper + expersq + black + smsa + south + smsa66 +
    reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + reg669 + nearc4, data = card)

models <- list("OLS Model" = ols_model, "IV Model" = iv_model)  # Named list for better output

# Create the model summary table with standard errors beneath estimates
table <- modelsummary(models,
                      estimate = "{estimate}{stars} ({std.error})",
                      statistic = NULL  # Prevents duplicate S.E
                      )
# Display the table
table
```

Is IV warranted in the first Place?

Conduct the Hausman Endogeneity test or Durbin-Wu-Hausman Test

Check whether T is endogenous in the model.

```{r Hausman-endogen test}
# Perform the Hausman test
# hausman.systemfit(ols_model, iv_model)

# Perform Endogenity test : Manual Durbin-Wu-Hausman Approach
# Get residuals from first stage (reduced form)
residuals_educ <- residuals(first_stage)

# # Step 3: Second-Stage Regression (Test for endogeneity)
second_stage <- lm(lwage ~ educ + exper + expersq + black + smsa + south + smsa66 + reg661 + reg662 + reg663 +
    reg664 + reg665 + reg666 + reg667 + reg668 + reg669 + residuals_educ, data = card)

summary(second_stage)

```

Overidentification test

```{r IV-Overidentification}
library(ivreg)

# 2SLS with two instruments
# educ is endogenous, instrumented by nearc4 and IQ
iv_model <- ivreg(lwage ~ educ + exper + expersq + black + smsa + south + smsa66 + reg661 + reg662 + reg663 +
    reg664 + reg665 + reg666 + reg667 + reg668 + reg669 | exper + expersq + black + smsa + south + smsa66 +
    reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + reg669 + nearc4 + IQ, data = card)
```

```
93
94
95 # Overidentification test (if you have more instruments than endogenous variables)
96 # Test if the extra instruments are valid (orthogonal to the error term)
97 hansen_test <- waldtest(iv_model, vcov = vcovHC(iv_model, type = "HC0"))
98 print(hansen_test)
99
100 ```
```

8 Alternate IV estimator: IV-GMM

The Generalized Method of Moments (GMM) is an econometric estimation technique that generalizes the Method of Moments (MM) framework. It is widely used when standard assumptions for Ordinary Least Squares (OLS) regression, such as strict exogeneity, are violated due to endogeneity issues **?**. Instrumental Variables (IV) estimation is a related approach used to address endogeneity, where instruments are variables correlated with endogenous regressors but uncorrelated with the error term.

GMM -Methodology GMM is based on the idea that a set of moment conditions derived from economic theory can be used to estimate parameters. Consider a model:

$$Y_i = X_i\beta + \varepsilon_i \tag{9}$$

where X_i is endogenous. If Z_i is a valid instrument such that:

$$E[Z_i'\varepsilon_i] = 0 \tag{10}$$

then the moment condition is:

$$g(\beta) = \frac{1}{N}\sum_{i=1}^{N} Z_i'(Y_i - X_i\beta) = 0. \tag{11}$$

GMM minimizes a quadratic function:

$$Q(\beta) = g(\beta)'Wg(\beta), \tag{12}$$

where W is a weighting matrix.

Advantages and Applications

GMM is robust to heteroskedasticity and does not require knowledge of the full likelihood function. It is used in dynamic panel models and asset pricing **??**.

Discussion on GMM is advanced study. It will be held separately in the context of *dynamic panel data* models.

8.1 Randomized Promotion

Randomized Promotion

Randomized Promotion is an Instrumental variable method that allows us to estimate impact unbiasedly. It randomly assigns *promotion or encouragement* to participate in a program. It is a useful strategy to evaluate programs open to everyone who is eligible.

- Randomized promotion is a good strategy when randomized assignment is not possible, and any individual who wishes to enroll in the program is free to do so.

8.2 Checklist: Randomized Assignment or Randomized Promotion

1. Baseline characteristics must be balanced between units that received the promotion campaign and those that did not.
2. **Validity of IV/promotion:** Verify that the promotion campaign substantially improves the take-up of the program. Compare program take-up rates in promoted vs non-promoted subsamples.
3. **Exclusion criteria of IV:** Promotion must not affect outcomes. This cannot be tested, so we rely on theory, contextual information, and common sense.

A The Potential Outcomes Framework

The Potential Outcomes Framework The potential outcomes framework, also known as the Rubin Causal Model, provides a formal structure for defining and estimating causal effects.

Basic Setup For each individual i in the population, we define:

- $T_i \in \{0, 1\}$: the treatment received (1 = treated, 0 = control)
- $Y_i(1)$: the potential outcome if individual i receives treatment
- $Y_i(0)$: the potential outcome if individual i does not receive treatment

The **individual treatment effect** for unit i is:

$$\tau_i = Y_i(1) - Y_i(0)$$

A.1 The Fundamental Problem of Causal Inference

We observe only one potential outcome for each individual:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

This means we can never observe τ_i directly. Instead, we estimate population-level average effects.

A.2 Average Treatment Effects

The **Average Treatment Effect (ATE)** is defined as:

$$\text{ATE} = E[Y_i(1) - Y_i(0)]$$

This represents the average causal effect of treatment across the entire population.

B Instrumental Variables and Potential Outcomes

B.1 Extended Potential Outcomes Notation

With an instrument $Z_i \in \{0, 1\}$, we extend the framework to include potential treatment assignments:

- $T_i(z)$: the treatment individual i would receive if $Z_i = z$
- $Y_i(z,t)$: the outcome individual i would have with $Z_i = z$ and $T_i = t$

B.2 Principal Strata

Based on potential treatment responses, we partition the population into four types:

$$\begin{aligned}
\text{Compliers (C):} \quad & T_i(1) = 1,\ T_i(0) = 0 \\
\text{Always-takers (A):} \quad & T_i(1) = 1,\ T_i(0) = 1 \\
\text{Never-takers (N):} \quad & T_i(1) = 0,\ T_i(0) = 0 \\
\text{Defiers (D):} \quad & T_i(1) = 0,\ T_i(0) = 1
\end{aligned}$$

These groups are called **principal strata** because membership is determined by potential outcomes and thus remains fixed regardless of actual treatment assignment.

C Local Average Treatment Effect (LATE)

C.1 Definition

Under standard IV assumptions (relevance, exclusion restriction, independence, and monotonicity), the IV estimator identifies the **Local Average Treatment Effect**:

$$\text{LATE} = E[Y_i(1) - Y_i(0) \mid \text{Complier}]$$

This is the average treatment effect specifically for individuals whose treatment status is affected by the instrument.

C.2 Why "Local"?

The term "local" emphasizes three key features:

1. **Subgroup-specific**: LATE applies only to compliers, not to the entire population
2. **Instrument-specific**: Different instruments may identify effects for different complier subpopulations
3. **Non-generalizable**: LATE may differ from ATE if treatment effects are heterogeneous

C.3 Relationship to ATE

We can decompose the ATE across principal strata:

$$\text{ATE} = \sum_{s \in \{C,A,N,D\}} P(s) \cdot E[Y_i(1) - Y_i(0) \mid s]$$

Under monotonicity (no defiers), and assuming homogeneous effects for always-takers and never-takers:

$$\text{LATE} = \text{ATE} \quad \text{if effects are homogeneous}$$

However, if treatment effects vary across subgroups, LATE $\neq$ ATE in general.

C.4 Identification via Wald Estimator

The LATE can be consistently estimated using the Wald estimator:

$$\hat{\tau}_{\text{LATE}} = \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[T_i \mid Z_i = 1] - E[T_i \mid Z_i = 0]}$$

This follows because:

- The numerator captures how the instrument affects outcomes
- The denominator captures the proportion of compliers (under IV assumptions)
- Only compliers contribute to both numerator and denominator

C.5 Interpretation in the Potential Outcomes Framework

LATE provides a causal interpretation within the potential outcomes framework by:

1. Identifying a well-defined subpopulation (compliers)
2. Estimating the average treatment effect for that subpopulation
3. Maintaining internal validity even when external validity to the full population is limited

The key insight is that IV methods answer the question: "What is the causal effect of treatment for those individuals who would be induced to change their treatment status by the instrument?"

References

Foster, G. (2009). A diagrammatic exposition of regression and instrumental variables for the beginning student. *The Journal of Economic Education*, 40:278–296.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475.

Labrecque, J. and Swanson, S. A. (2018). Understanding the assumptions underlying instrumental variable analyses: a brief review of falsification strategies and related tools. *Current Epidemiology Reports*, 2018:214–220.

The End