Lecture 09: Causal Inference in Statistical Genetics

PUBH 8878, Statistical Genetics

Chiraag Gohel

The George Washington University

2025-10-29

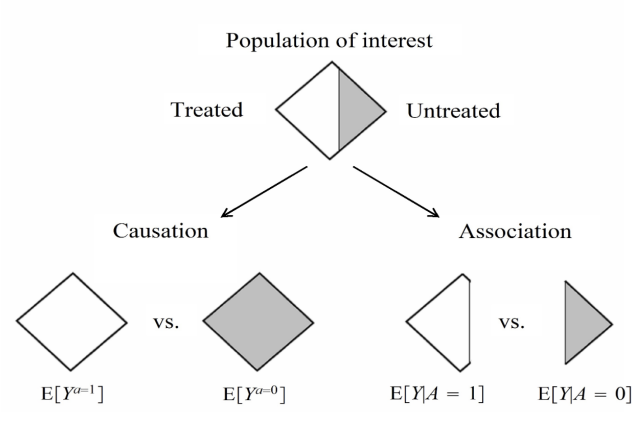
Motivation

- One major goal of epidemiology is to identify modifiable causes of health outcomes and disease (Celentano *et al.*, 2019)
- To enact interventions/treatment on some trait, we first want evidence that the trait **causes** the outcome of interest

Difference between causation and association

Consider

- Treatment A, where
$$A = \begin{cases} 1 \text{ if treated} \\ 0 \text{ if untreated} \end{cases}$$
- Outcome Y, where
$$Y = \begin{cases} 1 \text{ if death} \\ 0 \text{ if survival} \end{cases}$$
- $Y^{a=i}$ is the outcome that would have been observed under the treatment $a = i$



From Hernan and Robins (2025)

One solution: Randomization

Exchangeability

$$Y^a \perp A \text{ for all } a \implies \Pr[Y^a = 1 | A = 1] = \Pr[Y^a = 1 | A = 0]$$

- Or, independence between the counterfactual outcome and the observed treatment
- When group membership is randomized, in an ideal RCT, the groups are exchangeable
- Furthermore, this implies that $\mathrm{E}[Y^a | A = a'] = \mathrm{E}[Y^a]$

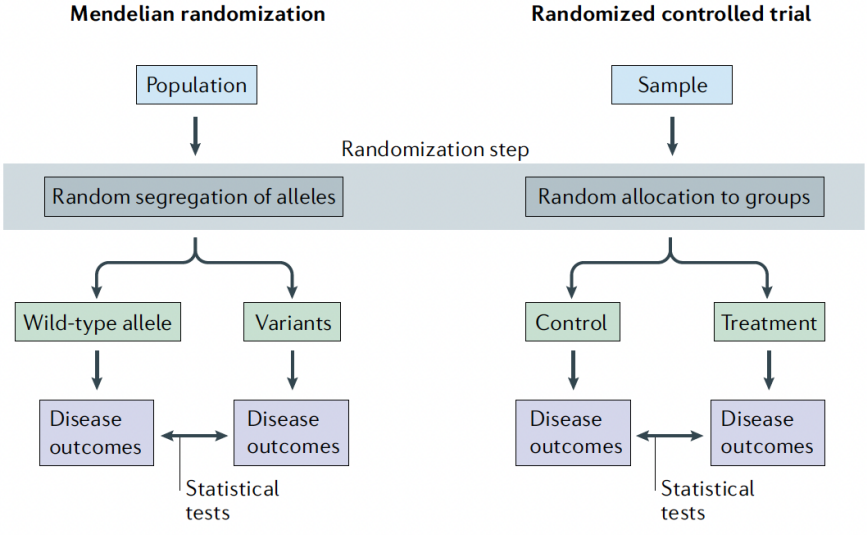
Randomization is often not possible

- Let's say a researcher wants to estimate the causal effect of smoking (A) on lung health (Y)
- We can't conduct an RCT experiment on the general population here
- Let U be common causes of A and Y (risk preferences, SES, environment)
- Note that
  $\Pr[Y^a = 1 \mid A = a] = \sum_u \Pr[Y^a = 1 \mid U = u, A = a] \Pr[U = u \mid A = a]$
- If $\Pr(U|A = 1) \neq \Pr(U|A = 0)$, then we do not have exchangeability

A genetic solution?

- However, we can use Mendelian randomization: the distribution of alleles/genes (G) is set at conception and is approximately random.
- Think of G as an as-if randomized assignment that nudges smoking (X)
  - We can compare lung health (Y) across G groups much like RCT arms, scaled by the G-induced difference in X
- After adjusting for ancestry/population structure (A), G should be independent of typical confounders (U)

MR compared to RCT



From Figure 1 in Sanderson *et al.* (2022)

How MR approximates exchangeability

- Use genotype G as an as-if randomized assignment determined at conception.
- Key approximation: for each $a \in \{0, 1\}$, $Y^a \perp G$ (often taken conditional on ancestry/population structure), so $\mathsf{E}[Y^a \mid G] = \mathsf{E}[Y^a]$.
- Timing: the "assignment" (G) happens at conception; effects reflect long-run exposure differences rather than acute treatment.
- Noncompliance: G only shifts X; not everyone changes behavior. MR targets the effect among those whose X is moved by G (a complier-type estimand).
- This mirrors RCT exchangeability, replacing randomized A with (approximately) randomized G.

A more rigorous treatment of concepts in RCT can be found in Evans and Ting (2021)

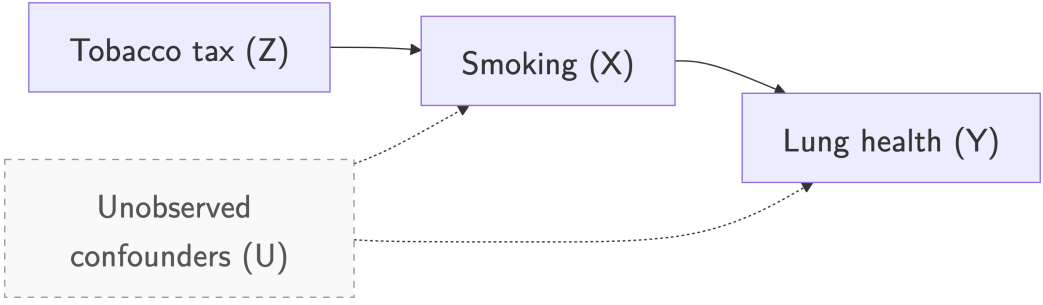
Instrumental Variables

Definition

An **instrument** is a variable that predicts the exposure, but conditional on the exposure shows no independent association with the outcome (Lousdal, 2018)

- Again, consider estimating the causal effect of smoking (A) on lung health (Y)
- Let U be common causes of A and Y
- Consider an instrument Z, recorded tobacco tax levels

Tobacco tax → Smoking → Lung health

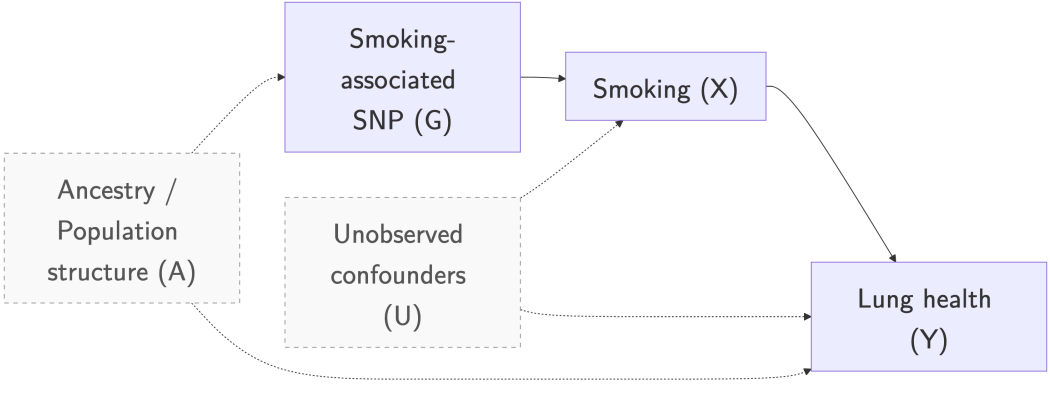


IV DAG

Examples of confounders

- $X \to Y$ confounders (U): socioeconomic status, occupational exposures, ambient air pollution, household/peer smoking, mental health and risk preferences, access to healthcare and preventive care, baseline respiratory conditions.
- $Z \to Y$ threats: smoke-free laws and anti-smoking campaigns, healthcare policy intensity, or regional socioeconomic trends that correlate with both tobacco taxes and lung health.

SNP → Smoking → Lung health



Mendelian randomization: SNP $\rightarrow$ Smoking $\rightarrow$ Lung health

Examples of confounders

- $X \rightarrow Y$ confounders (U): socioeconomic status, occupational exposures, ambient air pollution, household/peer smoking, mental health and risk preferences, access to healthcare and preventive care, baseline respiratory conditions.
- Population structure (A): ancestry differences, recruitment center/region, or subtle structure that links allele frequencies and lung health via environmental or clinical differences.
- $G \rightarrow Y$ threats: horizontal pleiotropy (genetic effects on lung health not via smoking), dynastic effects/assortative mating.

IV Conditions

In order for the instrument to provide a valid test of the null hypotehsis that the exposure has no effect on the outcome, certain conditions must hold (Didelez *et al.*, 2010):

1. Relevance

The IV must be associated with the exposure

2. Exchangeability

There are no causes of the IV that also influence the outcome through mechanisms other than the exposure of interest

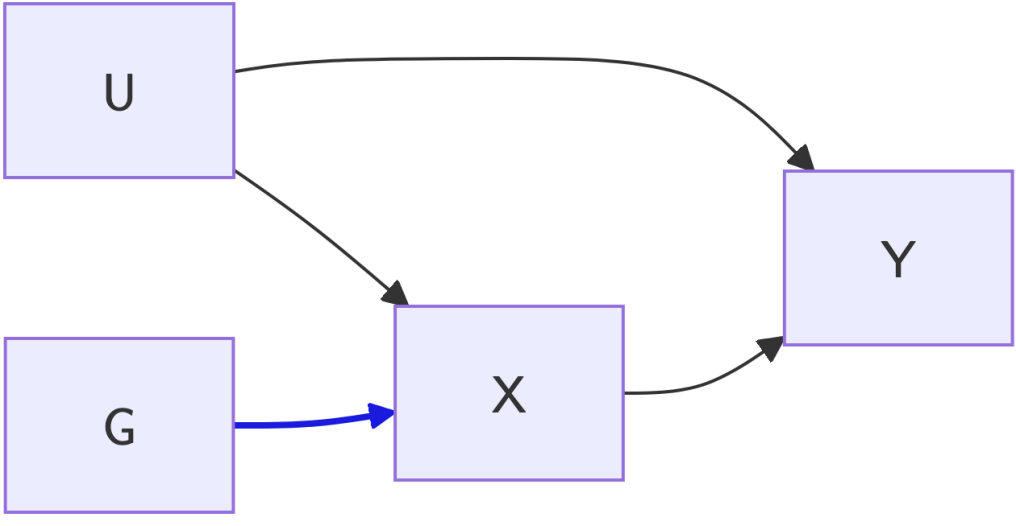
3. The Exclusion Restriction

The IV does not affect the outcome other than through the exposure and does not affect any other trait that has a downstream effect on the outcome of interest.

IV Conditions

1. Relevance

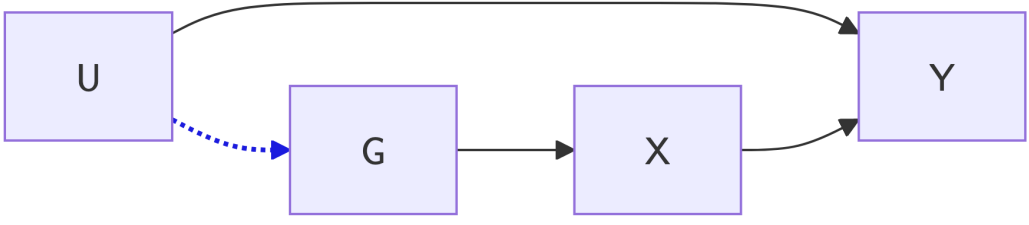
The IV must be associated with the exposure



IV Conditions

2. Exchangeability

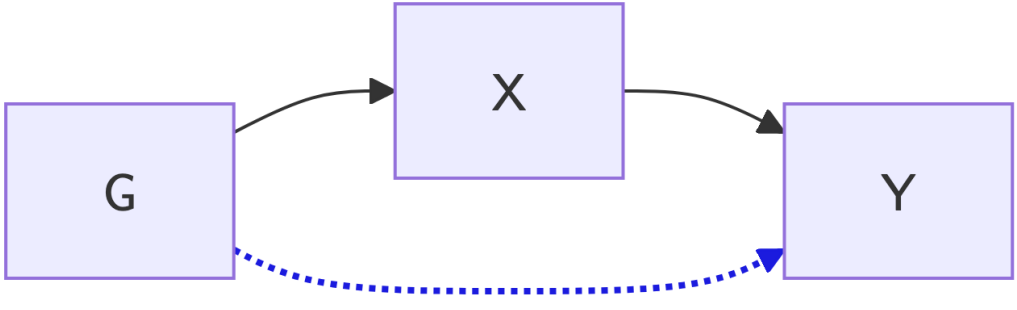
There are no causes of the IV that also influence the outcome through mechanisms other than the exposure of interest



IV Conditions

3. The Exclusion Restriction

The IV does not affect the outcome other than through the exposure and does not affect any other trait that has a downstream effect on the outcome of interest.



IV Conditions

1. Relevance
2. Exchangeability
3. The Exclusion Restriction

Only the first condition can be formally tested. The other two conditions can be disproved and otherwise assessed through a range of sensitivity analyses, but cannot be demonstrated to be true

---

For a deeper look into IV methods in Biostatistics, see Hernan and Robins (2025), Baiocchi *et al.* (2014), and Rubin and Imbens (2015)

Assessing IV Conditions

1. Relevance
   - The strength of the instrument can be assessed through the an F-statistic from the regression of the exposure on the instrument
   - A common rule of thumb is that an F-statistic > 10 indicates a sufficiently strong instrument (Staiger and Stock, 1997)

Assessing IV Conditions

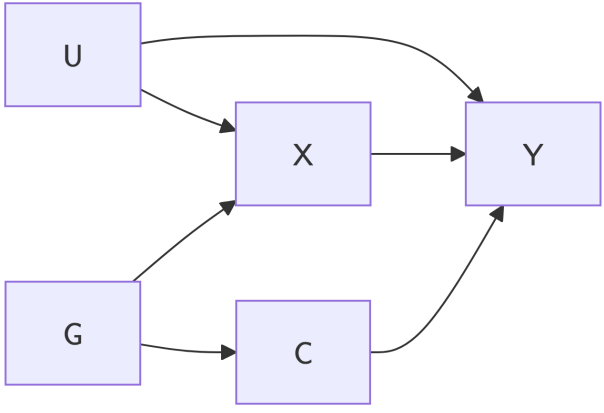
2. Exchangeability
   - Confounding of the genetic variants with the outcome can occur due to population structure, assortative mating, or dynastic effects
   - Common approaches to mitigate these issues include adjusting for principal components of ancestry, restricting to unrelated individuals, and within-family study designs (Davies *et al.*, 2019)

Assessing IV Conditions

3. Exclusion Restriction
   - Violations can occur due to pleiotropy, where a genetic variant influences multiple traits
   - Violations can occur due to linkage disequilibrium

Assessing IV Conditions

Horizontal Pleiotropy

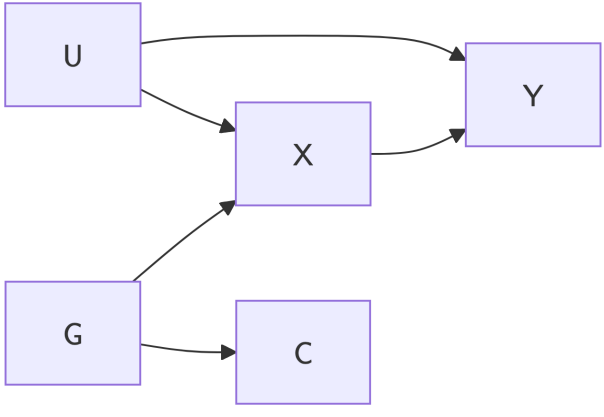


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- C : unmeasured phenotype

Assessing IV Conditions

Horizontal Pleiotropy

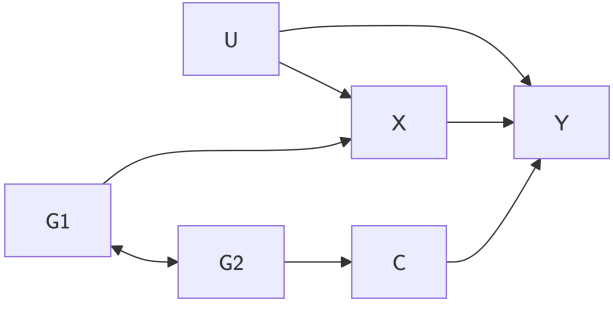


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- C : unmeasured phenotype

Assessing IV Conditions

Linkage disequilibrium

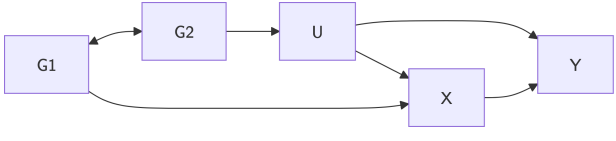


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- C : unmeasured phenotype

Assessing IV Conditions

Linkage disequilibrium

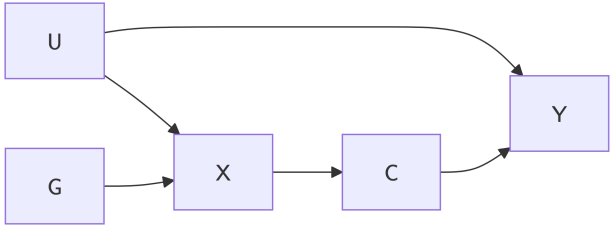


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder

Assessing IV Conditions

Vertical Pleiotropy

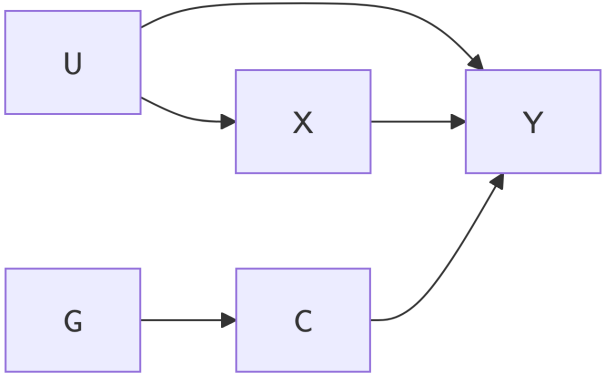


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- C : unmeasured phenotype

Assessing IV Conditions

Misspecification of the primary phenotype

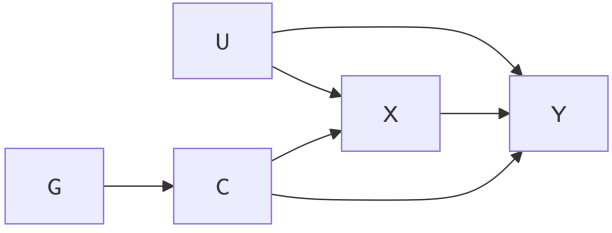


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- C : unmeasured phenotype

Assessing IV Conditions

Correlated pleiotropy

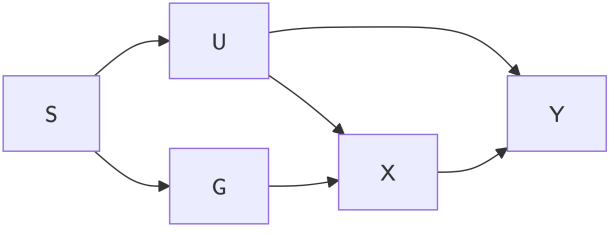


Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- C : unmeasured phenotype

Assessing IV Conditions

Population stratification



Key

- G : Genetic variant
- X : Exposure of interest
- Y : Outcome of interest
- U : unmeasured confounder
- S : population structure/ancestry

Assessing IV Conditions

For a deeper dive into understanding directed acyclic graphs (DAGs) and causal inference, see Pearl (2022)

Point-estimate identifying conditions

- **IV1–IV3** are sufficient to test the **exact null** (no causal effect), **not** to identify a numeric effect size.
- For a **point estimate**, add one of the following assumptions

Homogeneity

Either

1. The effect of the exposure on the outcome is the same for all individuals (estimate is the causal effect of the exposure on the outcome)
2. The effect of the exposure on the outcome is independent of the value of the instrument (estimate is the population average causal effect)

Monotonicity

The direction of the effect of the genetic variant on the exposure is the same for everyone

Two Stage Least Squares Estimation

Stage One

- Let X be the exposure of interest, **G** be a $n \times p$ matrix of genetic variants. We can then model

$$X = \pi_0 + \mathbf{G} + v_x$$

Stage Two

- The outcome is then regressed upon the predicted value of the exposure, $\hat{X}$

$$Y = \alpha + \beta\hat{X} + u$$

Where $\hat{\beta}$ is a consistent estimator of the causal effect of X on Y if the IV assumptions hold (Wooldridge, 2010)

TSLS Example

```
set.seed(8878)
n <- 2000              # sample size
U <- rnorm(n)           # unobserved confounder
Z <- rbinom(n, 1, 0.5) # instrument
e_x <- rnorm(n)         # error term for X
e_y <- rnorm(n)         # error term for Y
```

TSLS Example

```r
beta_true <- 1.5 # causal effect of X on Y
theta     <- 0.8 # effect of Z on X (relevance)
alpha     <- 1.0 # effect of U on Y (confounding)
lambda    <- 0.9 # effect of U on X (confounding)
```

```r
X <- theta * Z + lambda * U + e_x          # exposure
Y <- beta_true * X + alpha * U + e_y       # outcome
df <- tibble(Y = Y, X = X, Z = Z, U = U)
```

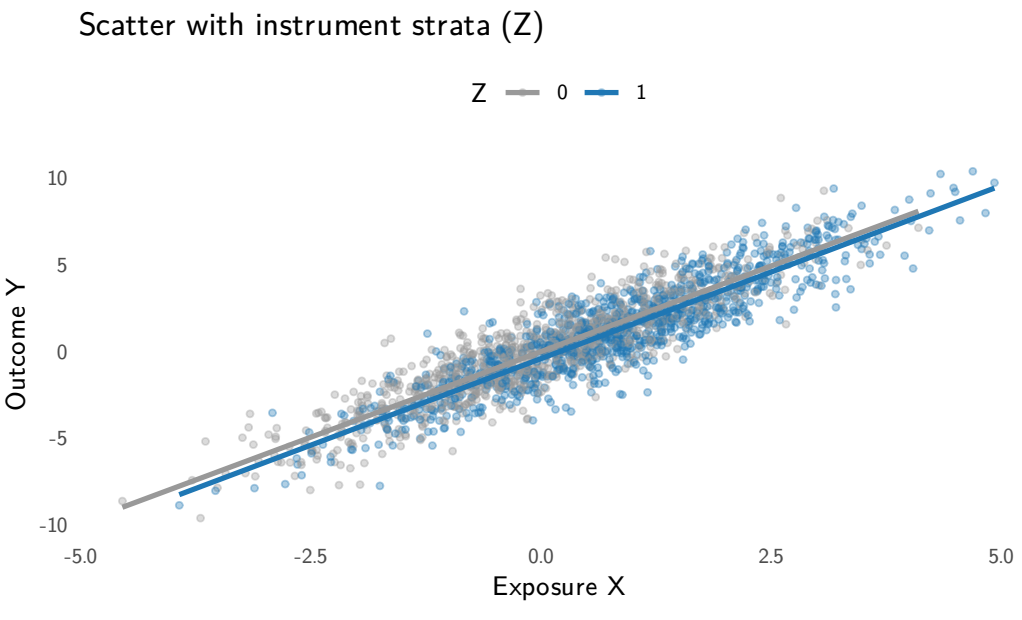
TSLS Example

```r
m_ols   <- lm(Y ~ X, data = df)   # OLS estimate

m_stage1 <- lm(X ~ Z, data = df)  # 1st stage: predict X from Z
Xhat <- fitted(m_stage1)          # predicted X from 1st stage

m_tsls <- lm(Y ~ Xhat, data = df) # 2nd stage: regress Y on Xhat

# First-stage strength (F-statistic on Z)
F1 <- unname(summary(m_stage1)$fstatistic[1])
```

TSLS Example

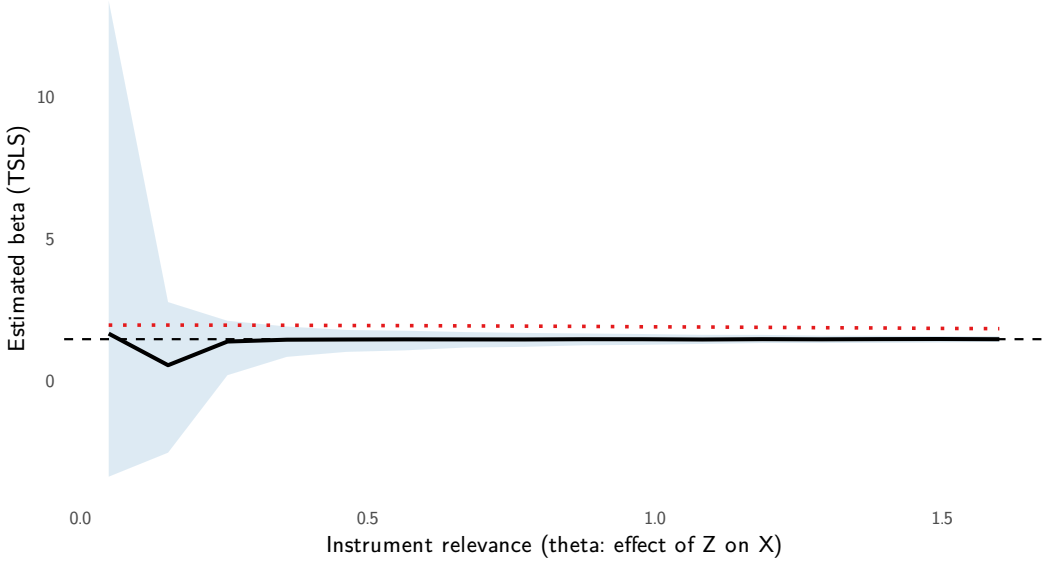
True beta	OLS beta-hat	2SLS beta-hat	F
1.5	1.943296	1.493856	182.5937

TSLS Example



TSLS across instrument strength

Ribbon: 2.5–97.5% across simulations; dashed = true beta; dotted = OLS mean



References I

Baiocchi, M., Cheng, J. and Small, D. S. (2014) Instrumental variable methods for causal inference. *Statistics in Medicine*, **33**, 2297–2340. DOI: 10.1002/sim.6128.

Celentano, D. D., Szklo, M. and Gordis, L. (2019) *Gordis Epidemiology*. 6th edition. Philadelphia, PA: Elsevier.

Davies, N. M., Howe, L. J., Brumpton, B., et al. (2019) Within family Mendelian randomization studies. *Hum Mol Genet*, **28**, R170–R179. DOI: 10.1093/hmg/ddz204.

Didelez, V., Meng, S. and Sheehan, N. A. (2010) Assumptions of IV Methods for Observational Epidemiology. *Statistical Science*, **25**, 22–40. Institute of Mathematical Statistics. DOI: 10.1214/09-STS316.

References II

Evans, S. R. and Ting, N. (2021) *Fundamental Concepts for Clinical Trialists*. Boca Raton: Chapman & Hall/CRC.

Hernan, M. A. and Robins, J. M. (2025) *Causal Inference: What If*. Boca Raton: CRC Press.

Lousdal, M. L. (2018) An introduction to instrumental variable assumptions, validation and estimation. *Emerg Themes Epidemiol*, **15**, 1. DOI: 10.1186/s12982-018-0069-7.

Pearl, J. (2022) *Causality: Models, Reasoning, and Inference*. Second edition, reprinted with corrections. Cambridge New York, NY Port Melbourne New Delhi Singapore: Cambridge University Press.

References III

Rubin, D. B. and Imbens, G. W. (eds) (2015) Instrumental Variables Analysis of Randomized Experiments with Two-Sided Noncompliance. In *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, pp. 542–559. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139025751.025.

Sanderson, E., Glymour, M. M., Holmes, M. V., et al. (2022) Mendelian randomization. *Nat Rev Methods Primers*, **2**, 1–21. DOI: 10.1038/s43586-021-00092-5.

Staiger, D. and Stock, J. H. (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65**, 557–586. [Wiley, Econometric Society]. DOI: 10.2307/2171753.

Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA, USA: MIT Press.