SISCER Module 15
Lecture 5: Instrumental variables

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 2022

Hierarchy of evidence

When the goal is to infer causation...

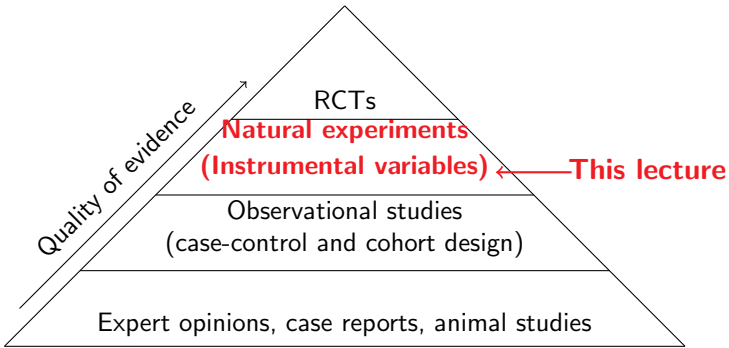


Figure: (A rough) Hierarchy of evidence in medical studies.

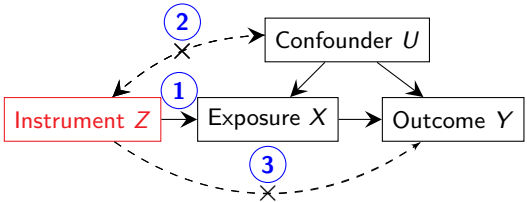
Fundamental challenge of observational studies

"Correlation does not imply causation".

Observational studies = Enumerating confounders

- Idea: Conditioning on possible sources of spurious correlation.
- Example: Possible confounders between smoking and lung cancer:
  - Age.
  - Sex.
  - Urban/Rural.
  - Working environment.
  - Socioeconomic class.
  - ...
- **Fundamental challenge: We can never be sure this list is complete.**
- The promise of instrumental variables: unbiased estimation of causal effect without enumerating confounders.

What is an instrument variable (IV)?



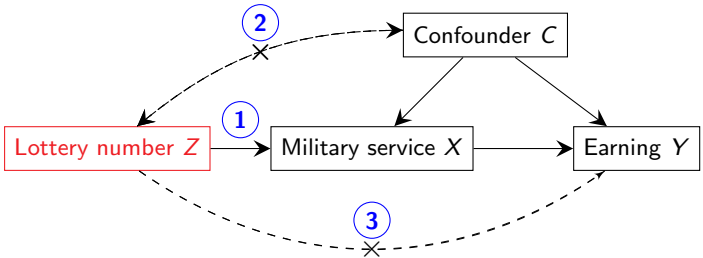
Core IV assumptions

1. **Relevance**: Z is associated with the exposure (X).
2. **Effective random assignment**: Z is independent of unmeasured confounder (U).
3. **Exclusion restriction**: Z cannot have any direct effect on the outcome (Y).

Wald's estimator based on Intention-to-treat (ITT) analysis

$$\text{Causal effect of } X \text{ on } Y \approx \frac{\text{ITT Effect of } Z \text{ on } Y}{\text{ITT Effect of } Z \text{ on } X}$$

IV in Economics: Effect of military service on earnings (Angrist, 1990)



- In 1970, the U.S. government conducted draft lottery to determine priority of conscription for the Vietnam war.
- Exercise: Justify the core IV assumptions.
- The draft lottery can be regarded as a "natural experiment" of military service.

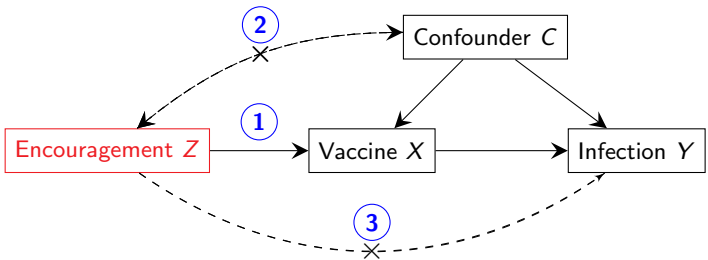
Results of the Vietnam-war lottery study

Table 4.1.3: Wald estimates of the effects of military service on the earnings of white men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

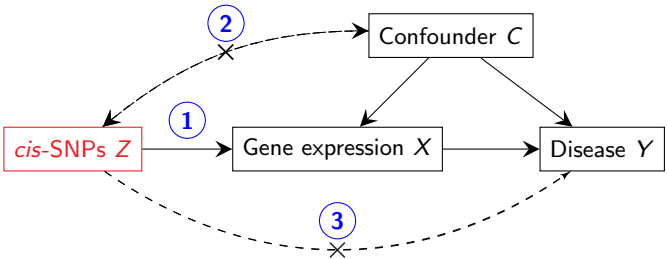
Notes: Adapted from Angrist (1990), Tables 2 and 3. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

IV in Public Health: Effectiveness of vaccine (Hirano et al., 2000)



- This is also called randomized encouragement design.
- The same idea can be applied to RCTs with non-compliance.

IV in Human Genetics: Gene testing (Gamazon et al., 2015)



- Compared to *trans*-SNPs, *cis*-SNPs are more likely to satisfy exclusion restriction (criterion 3).
- This is a special case of "Mendelian randomization" where genetic variation is used as IV and typically X is an epidemiological risk factor (more downstream).

Linear IV model

- The Wald ratio estimator becomes inadequate when $\boldsymbol{Z}$ and $\boldsymbol{X}$ are multivariate.
- The most commonly used IV estimators are based on the following linear model:

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\alpha} + U_i,$$
$$\boldsymbol{X}_i = \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \boldsymbol{V}_i.$$

IV assumptions in the linear model

1. Relevance: $\boldsymbol{\gamma} \neq 0$;
2. Exogeneity: $\boldsymbol{Z}_i \perp\!\!\!\perp (U_i, \boldsymbol{V}_i)$;
3. Exclusion restriction: $\boldsymbol{\alpha} = \boldsymbol{0}$.

- The exposure variable $\boldsymbol{X}_i$ is called *confounded* or *endogenous* if it is correlated with U_i (or equivalently, if $\boldsymbol{V}_i$ is correlated with U_i).

Identification of causal effect

Under the linear IV model, the causal effect $\boldsymbol{\beta}$ satisfies $\mathbb{E}[\boldsymbol{Z}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})] = \mathbf{0}$.

- Notice how this is different from the usual normal equation $\mathbb{E}[\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})] = \mathbf{0}$.
- To identify $\boldsymbol{\beta}$, we need $\dim(\boldsymbol{Z}_i) \geq \dim(\boldsymbol{X}_i)$.
- Just-identified case: When $\dim(\boldsymbol{Z}_i) = \dim(\boldsymbol{X}_i)$, we can estimate $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^{n} \boldsymbol{Z}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}) = \mathbf{0}.$$

  The solution in matrix-form is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{Y}.$$

- Over-identified case: When $\dim(\boldsymbol{Z}_i) > \dim(\boldsymbol{X}_i)$, we have some freedom to choose which (linear combinations of) equations to solve.

Two-stage least squares (TSLS)

- In the over-identified case, for any function $\boldsymbol{f} : \mathbb{R}^{\dim(\boldsymbol{Z}_i)} \to \mathbb{R}^{\dim(\boldsymbol{X}_i)}$, we have

$$\mathbb{E}[\boldsymbol{f}(\boldsymbol{Z}_i) \cdot (Y_i - \boldsymbol{X}_i^T \beta)] = \boldsymbol{0}.$$

- The most efficient choice of $\boldsymbol{f}$ is $\boldsymbol{f}(\boldsymbol{Z}_i) = \mathbb{E}[\boldsymbol{X}_i \mid \boldsymbol{Z}_i] = \boldsymbol{Z}_i^T \gamma$.
- The nuisance parameter γ is not known but can be estimated from the data. The most common estimator is least squares:

$$\hat{\gamma} = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}.$$

- This is called two-stage least squares, because (let $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\gamma}$)

$$\hat{\beta} = \texttt{lm}(\boldsymbol{Y} \sim \hat{\boldsymbol{X}}) = \texttt{lm}(\boldsymbol{Y} \sim \texttt{predict}(\texttt{lm}(\boldsymbol{X} \sim \boldsymbol{Z})))$$

- However, standard error of $\hat{\beta}$ cannot be obtained directly from `lm` because $\hat{\gamma}$ is estimated from the data.

Limited information maximum likelihood (LIML)

- Recall the linear IV model:

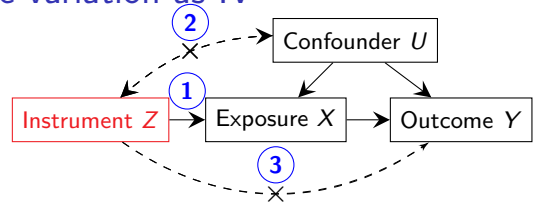
$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + U_i,$$
$$\boldsymbol{X}_i = \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \boldsymbol{V}_i.$$

- The LIML estimator assumes the noise variables $(U_i, \boldsymbol{V}_i)$ are jointly normal with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$.
- LIML maximizes the log-likelihood of this problem:

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) = -\frac{1}{2} \sum_{i=1}^{n} \log |\boldsymbol{\Sigma}^{-1}| + \begin{pmatrix} Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} \\ \boldsymbol{X}_i - \boldsymbol{Z}_i^T \boldsymbol{\gamma} \end{pmatrix}^T \boldsymbol{\Sigma}^{-1} \begin{pmatrix} Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} \\ \boldsymbol{X}_i - \boldsymbol{Z}_i^T \boldsymbol{\gamma} \end{pmatrix}.$$

- TSLS and LIML are asymptotically equivalent (when $n \to \infty$ and $\dim(\boldsymbol{X}_i)$ and $\dim(\boldsymbol{Z}_i)$ are fixed).
- LIML is more robust to weak instruments (small $\boldsymbol{\gamma}$).

MR = Using genetic variation as IV



Examine the core IV assumptions

Criterion (1)	✓	Modern GWAS have identified many causal variants
Criterion (2)	✓	Almost Comes for free due to Mendel's Second Law Possible concern: population stratification
Criterion (3)	?	Problematic because of wide-spread **pleiotropy** (multiple functions of genes).

Summary-data Mendelian randomisation

- Suppose $Z_1, \ldots, Z_p$ are independent SNPs.
- $\hat{\Gamma}_j \overset{ind.}{\sim} \mathrm{N}(\Gamma_j, \sigma_{1j}^2)$: SNP effect on outcome Y, obtained from $\texttt{lm}(Y \sim Z_j)$.
- $\hat{\gamma}_j \overset{ind.}{\sim} \mathrm{N}(\gamma_j, \sigma_{2j}^2)$: SNP effect on treatment A, obtained from $\texttt{lm}(A \sim Z_j)$.
- Model for pleiotropy:

$$\Gamma_j = \beta\gamma_j + \alpha_j,$$

  where β is causal effect of A on Y and $\alpha_j \sim \mathrm{N}(0, \tau^2)$ is direct effect of Z on Y.
- Inverse-variance weighted (IVW) estimator:

$$\hat{\beta} = \frac{\sum_{j=1}^{p}(1/\sigma_{1j}^2)(\hat{\Gamma}_j/\hat{\gamma}_j)}{\sum_{j=1}^{p}(1/\sigma_{1j}^2)}.$$

- Other methods: debiased IVW (Ye et al., 2021), weighted median, robust adjusted profile score (Zhao et al., 2020), and many others.

Mendelian randomisation: Example

- GWAS summary data from UK BioBank. Sample size around 500,000. 160 SNPs.
- Exposure is Body Mass Index (BMI); Outcome is systolic blood pressure (SBP).

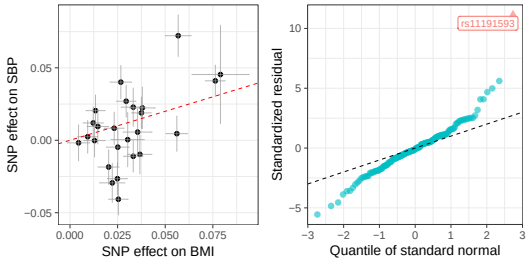


Figure: Left: $\hat{\Gamma}_j$ vs $\hat{\gamma}_j$; Right: Q-Q plot for $\hat{\alpha}_j = \hat{\Gamma}_j - \hat{\beta}\hat{\gamma}_j$ after standardisation.

- Estimated $\hat{\beta} = 0.402$ (standard error $= 0.106$). BMI and SBP were standardised.
- Reference: Zhao et al. (2020, 2019).

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *American Economic Review*, 80(3):313–336.

Gamazon, E. R., , Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., and Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88.

Ye, T., Shao, J., and Kang, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. *The Annals of statistics*, 49(4):2079–2100.

Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *International Journal of Epidemiology*, 48(5):1478–1492.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics*, 48(3):1742–1769.