STA 640 — Causal Inference

Chapter 6.1 Instrumental Variables

Fan Li

Department of Statistical Science
Duke University

Instrumental Variables

- Unmeasured confounding, i.e. unmeasured factors that affect both treatment assignment and outcome, is the major challenge in causal inference
- Instrumental variables (IV) is a main method in handling unmeasured confounding, essentially a natural experiment
- Originated from economics (bread and butter) and widely adopted in social sciences and recently genetics
- Main idea
  1. Find a variable (i.e. IV) that influences treatment assignment but is independent of unmeasured confounders and has no direct effect on the outcome except through its effect on treatment;
  2. Use this variable to extract variation in the treatment that is free of the unmeasured confounders;
  3. Use this confounder-free variation in the treatment to estimate the causal effect of the treatment

Instrumental Variables: DAG

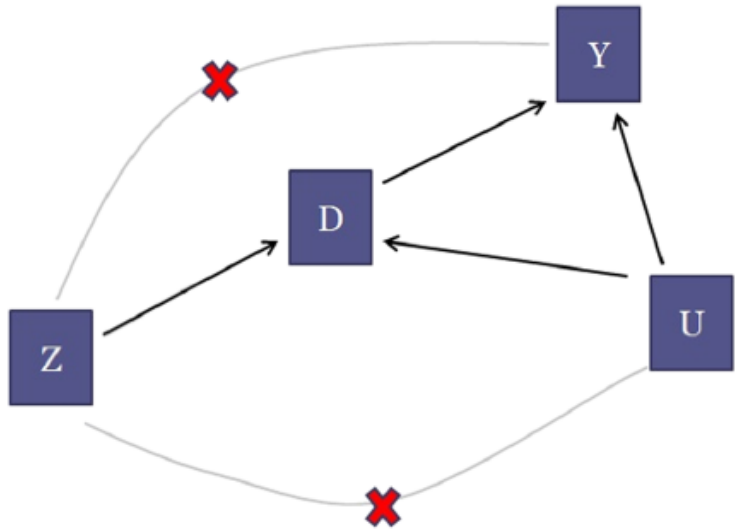

Figure 1. Directed acyclic graph for the relationship between an instrumental variable Z, a treatment D, unmeasured confounders U, and an outcome Y.

Notations:

- IV Z; treatment D; outcome Y; covariates X; unmeasured confounder U

Example of IV: distance to speciality care provider

- A classic example is McClellan et al. (1994, JAMA): study the effect of cardiac catheterization (treatment) for patients suffering a heart attack
- IV: the differential distance the patient lives from the nearest hospital that performs cardiac catheterization to the nearest hospital that does not perform
- Rationale: how close one lives to an advanced hospital is largely random (natural experiment), but it affects whether a patient got the treatment and thus outcome.
- More generally, for emergent conditions, proximity to a specialty care provider particularly enhances the chance of being treated by the specialty care provider.

Brief review of econometric approach to IV

Wooldridge, 2002; Imbens, 2014

- Traditional linear model of an outcome Y being related to a scalar treatment (i.e. endogenous variable) D given a set of covariates (i.e. exogenous variables) X:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2' X_i + \varepsilon_i \tag{1}$$

  where the estimand is the coefficient β_1
- Note: we no longer assume unconfoundedness
- Challenge: the error term ε_i is dependent of treatment variable D_i (confounded). Direct OLS estimator of β_1 is biased
- IV: a vector of dimension K of IVs Z, which satisfies (1) $\varepsilon_i \perp X_i$, (2) $\varepsilon_i \perp Z_i | X_i$. Together

$$\varepsilon_i \perp (Z_i, X_i)$$

- When the dimension of IV $K > 1$, over-identified; when $K = 1$, just-identified

Case of just-identified, no covariates

- The IV estimator of β_1 is the ratio of covariance:

$$\hat{\beta}_{1,\text{iv}} = \frac{\widehat{\text{Cov}}\left(Y_i, Z_i\right)}{\widehat{\text{Cov}}\left(D_i, Z_i\right)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\left(Z_i - \bar{Z}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(D_i - \bar{D}\right)\left(Z_i - \bar{Z}\right)}$$

  where $\bar{Y}, \bar{D}$ are the sample means

- With a binary IV Z, this is the Wald estimator:

$$\hat{\beta}_{1,\text{iv}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$$

  where $\bar{Y}_z, \bar{D}_z$ are the sample means within group $z(= 0, 1)$

Interpretation I: indirect least squares

- Two interpretations of the IV estimator $\hat{\beta}_{1,\text{iv}}$.
- Interpretation I: indirect least squares. Two reduced forms of regressions:

$$\begin{aligned} Y_i &= \pi_{10} + \pi_{11} \cdot Z_i + \varepsilon_{1i} \\ D_i &= \pi_{20} + \pi_{21} \cdot Z_i + \varepsilon_{2i} \end{aligned}$$

- The indirect least squares (ILS) estimator is the ratio of the least squares estimates of π_{11} and π_{21}: $\hat{\beta}_{1,\text{ils}} = \hat{\pi}_{11}/\hat{\pi}_{21}$
- In the case of randomized trial with binary treatment, $\hat{\beta}_{1,\text{ils}}$ is the ratio of the ITT estimates (Angrist, Imbens, Rubin, 1996)

Interpretation II: two stage least squares (2SLS)

- Stage 1: predict treatment value from IV via OLS:

$$\hat{D}_i = \hat{\pi}_{20} + \hat{\pi}_{21} \cdot Z_i$$

- Stage 2: plug in the predicted treatment in Stage 1 in the outcome model:

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \eta_i$$

- Estimate β_1 from Stage 2 via OLS to obtain the 2SLS estimator of β_1: $\hat{\beta}_{1,\text{2sls}}$
- Intuition: D is confounded, but we can use the IV to recover the "unconfounded portion" of D and plug into the outcome model
- Easy to verify

$$\hat{\beta}_{1,\text{iv}} = \hat{\beta}_{1,\text{ils}} = \hat{\beta}_{1,\text{2sls}}$$

Case of just-identified, with covariates

- The above discussion is straightforward to extend to the case with covariates
- Indirect LS: two reduced-form regressions

$$\begin{aligned} Y_i &= \pi_{10} + \pi_{11} \cdot Z_i + \pi_{12}' X_i + \varepsilon_{1i} \\ D_i &= \pi_{20} + \pi_{21} \cdot Z_i + \pi_{22}' X_i + \varepsilon_{2i} \end{aligned}$$

  ILS estimator: $\hat{\beta}_{1,\text{ils}} = \hat{\pi}_{11}/\hat{\pi}_{21}$
- 2SLS:

$$\begin{aligned} \hat{D}_i &= \hat{\pi}_{20} + \hat{\pi}_{21} \cdot Z_i + \hat{\pi}_{22}' X_i \\ Y_i &= \beta_0 + \beta_1 \hat{D}_i + \beta_2' X_i + \eta_i \end{aligned}$$

  2SLS estimator: OLS estimator of β_1 in Stage 2
- $\hat{\beta}_{1,\text{iv}} = \hat{\beta}_{1,\text{ils}} = \hat{\beta}_{1,\text{2sls}}$

Variance estimation

- The standard error for 2SLS estimate $\hat{\beta}_{1,2\text{sls}}$ is NOT the standard error of coefficient of $\hat{D}$ from Stage 2, because one also needs to account for the sampling uncertainty in using $\hat{E}(D|Z)$ as an estimate of $E(D|Z)$
- Assume homoskedasticity of the residuals in the IV model $\varepsilon_i \sim N(0, \sigma^2_\epsilon)$
- In large samples the distribution of the IV estimator $\hat{\beta}_{\text{iv}}$ is approximately normal, centered around the true value $\hat{\beta}$, with variance (Wooldridge, 2002):

$$\widehat{\mathbb{V}} = \hat{\sigma}^2_\varepsilon \cdot \left(\begin{pmatrix} 1 \\ \hat{D}_i \\ X_i \end{pmatrix} \begin{pmatrix} 1 \\ \hat{D}_i \\ X_i \end{pmatrix}' \right)^{-1}$$

2SLS with non-linear models: Forbidden regressions

- The relation $\hat{\beta}_{1,\text{iv}} = \hat{\beta}_{1,\text{ils}} = \hat{\beta}_{1,\text{2sls}}$ holds under OLS models in both stages
- How about non-linear model? Two examples
- First example
  - i Non-linear 2nd stage, e.g., the outcome Y is a quadratic function of the treatment D: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \epsilon_i$
- Can we simply run a single first stage model, obtain an estimated (instrumented) $\hat{D}$ and plug it into the second stage?
  - i fit a linear 1st stage regression of D on Z, an fit an OLS of Y on $(1, \hat{D}_i, \hat{D}_i^2)$
- Unfortunately, the answer is NO. In fact, "forbidden" by MIT professor Jerry Hausman in 1975 (Angrist and Pischke, 2009)
- What is the correct way? Run a separate 1st stage regression for D and D^2, respectively, obtain separate estimates, $\hat{D}$ and $\widehat{D^2}$, respectively, and then plug these into second stage,

2SLS with non-linear models: Forbidden regressions

Angrist and Pischke, 2009, page 190

As a rule, naively plugging in first-stage fitted values in nonlinear models is a bad idea. This includes models with a nonlinear second stage as well as those where the CEF for the first stage is nonlinear. Suppose, for example, that you believe the causal relation between schooling and earnings is approximately quadratic but otherwise homogeneous (as in Card's (1995) structural model). In other words, the model of interest is

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho_1 \mathrm{s}_i + \rho_2 \mathrm{s}_i^2 + \eta_i. \qquad (4.6.5)$$

Given two instruments, it's easy enough to estimate (4.6.5), treating both s_i and s_i^2 as endogenous. In this case, there are two first-stage equations, one for s_i and one for s_i^2. Although you need at least two instruments for this to work, it's natural to use the original instrument and its square (unless the only instrument is a dummy, in which case you'll need a better idea).

You might be tempted, however, to work with a single first stage, say equation (4.6.2), and estimate the following second stage manually:

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho_1 \hat{\mathrm{s}}_i + \rho_2 \hat{\mathrm{s}}_i^2 + [\eta_i + \rho_1(\mathrm{s}_i - \hat{\mathrm{s}}_i) + \rho_2(\mathrm{s}_i^2 - \hat{\mathrm{s}}_i^2)].$$

This is a mistake, since $\hat{\mathrm{s}}_i$ can be correlated with $\mathrm{s}_i^2 - \hat{\mathrm{s}}_i^2$ while $\hat{\mathrm{s}}_i^2$ can be correlated with both $\mathrm{s}_i - \hat{\mathrm{s}}_i$ and $\mathrm{s}_i^2 - \hat{\mathrm{s}}_i^2$. In contrast, as long as X_i and Z_i are uncorrelated with η_i in (4.6.5) and you have enough instruments in Z_i, 2SLS estimation of (4.6.5) is straightforward.

2SLS with non-linear models: binary treatment and 2SRI

- A second example of the forbidden regression is with a non-linear 1st stage, e.g. when the treatment D is binary
- One might be tempted to fit a logistic model in 1st stage and plug the predicted $\hat{D}_i$ into the second stage, so-called two-stage predictor substitution (2SPS) approach
- This is also wrong and forbidden
- A correct way is the two-stage residual inclusion (2SRI):
  1. Stage 1: Fit a logistic model of D on Z: $logit(\Pr(D_i = 1|X, Z) \sim Z_i + X_i$ and obtain the residual in predicting D: $r_i = \hat{D}_i - D_i$
  2. Stage 2: Regress Y on treatment D, covariates X and the residuals from Stage 1: $Y_i \sim D_i + X_i + r_i$
- Terza, Basu, Rathouz (2008, J health Econ) showed: coefficient of D in Stage 2 of 2SRI is consistent for β_1 in the IV model 1, but 2SPS is not

IV/2SLS: open questions

- What if the true outcome model has interactions between X and treatment D, i.e. heterogeneous treatment effect? What is the estimand? How to estimate?
- What if the data is clustered, e.g. patients clustered in hospitals? Should we use random effects models in both stages? What is exactly the 2SLS estimator now? Is it still consistent?
- A key feature: the IV/2SLS approach, including causal estimand is tied with a specific outcome model, i.e. an OLS model with homogeneous treatment effects.
- Inflexible to extend to more complex settings
- Distinct from the model-free causal estimands in the potential outcome framework
- General question: is there model-free interpretation or formulation in terms of potential outcomes?

History of IV

- Earliest concept of IV is usually attributed to Philip G. Wright and/or his son Sewall (Appendix B of *The Tariff on Animal and Vegetable Oils*, 1928)
- IV has since became a central technique of modern econometrics
- Classic econometric formulation of IVs is in terms of structural equations and assumptions about the IV being uncorrelated with structural error terms
- In a series of landmark papers in 1990's, Angrist, Imbens and Rubin connected IV to the potential outcomes framework in causal inference
- In statistics literature, IV was later extended to principal stratification (Frangakis and Rubin, 2002) for handling general post-treatment confounding

IV: potential outcomes and assumptions

- Below we discuss IV using the potential outcome notation, focusing on the case of binary treatment and IV
- IV $Z = 0, 1$; treatment $D = 0, 1$; outcome Y; covariates X
- Potential outcomes: $D(z), Y(z, d)$
- Assumptions:
  - A1 SUTVA
  - A2 IV is positively correlated with treatment: $cor(Z_i, D_i) > 0$ (usually the higher correlation the better)
  - A3 IV is independent of unmeasured confounders (conditional on covariates X): $\{Y(z, d), D(z)\} \perp Z|X$, for all z, d
  - A4 Exclusion restriction (ER): the IV affects outcomes only through its effect on treatment received: $Y(z', d) = Y(z, d)$ for all units. Under ER, $Y(z, d) \equiv Y(d)$ for $z = 0, 1$
- For point identification of a causal effect, Angrist et al. (1996) imposed an additional assumption
  - A5 Monotonicity: $D_i(1) \geq D_i(0)$ for all i

Motivating Context: randomized experiments with noncompliance

Angrist, Imbens, Rubin (1996, JASA)

- Noncompliance: units take treatment different from the assigned one
- Random treatment assigned: Z_i
- Actual treatment: D_i
- Noncompliance: $Z_i \neq D_i$ for some units
- Noncompliance can arise because, e.g. side effects, perception of the effectiveness of the treatment or control
- Noncompliance is self-selected: breaks the initial randomization

Big aside on noncompliance

- Two types of noncompliance
  - One-sided compliance: the control group is restricted access to treatment, so that noncompliance is only on the treatment group. Example: trials on a new drug
  - Two-sided compliance: both groups have access to treatment. Example: randomized encouragement trials
- Two naive approach:
  - Per-protocol: discarding non-complying units $\{i : Z_i \neq D_i\}$
  - As-treated: ignoring the initial random assignment, comparing units per their actual treatment status
- Both approaches are invalid. Why?
  - Per-protocol: compliance is self-selected, the remaining subsample is not representative of the whole study population
  - As-treated: randomization is broken

Intention-to-Treat (ITT) Approach

- The standard analysis for randomized studies with noncompliance is called *Intention to Treat* (ITT)
- ITT: ignores observed compliance information and compares those assigned to treatment to those assigned to control
- ITT estimand: $\tau^{ITT} = \mathbb{E}[Y_i(1) - Y_i(0)]$, essentially ATE of the assigned treatment Z on outcome
- Estimator of the ITT effect:
  $\widehat{\tau}^{ITT} = \sum_i Y_i Z_i / \sum_i Z_i - \sum_i Y_i(1 - Z_i) / \sum_i (1 - Z_i)$
- Rationale: preserve the randomization
- ITT procedure gives a valid estimate of the causal effect of the assignment on outcome (effectiveness), but not the effect of the treatment received on outcome (efficacy)

Effectiveness and Efficacy

- Effectiveness: the effect of a treatment work in practice
- Efficacy: the effect of a treatment in ideal situations
- Example: In the clinical development of a vaccine, an efficacy study asks the question, “Does the vaccine work?" In contrast, an effectiveness study asks the question “Does vaccination help people?"
- Effectiveness is more of policy interest (population level); efficacy is more of clinical or scientific interest (individual level)
- Randomized experiments are usually designed to study efficacy, but noncompliance and other complications render this difficult

Instrumental Variable Approach to Noncompliance

Angrist, Imbens, and Rubin (1996, JASA)

- Z_i: assigned treatment; D_i: actual treatment (which might be different from Z_i); Y_i: outcome
- The treatment received D is post-assignment, therefore also has two potential outcomes: $D(z), z = 0, 1$
- Potential outcomes: $Y(z)$ for $z = 0, 1$ (omit the double index $Y(z, d)$ here for simplicity)
- Observed data: $Z_i, D_i = D(Z_i), Y_i = Y(Z_i)$
- The central idea: (i) random assignment is an IV; (ii) divide units into latent subgroups based on their compliance behavior
- Defining compliance type: $S_i = (D_i(0), D_i(1))$.
- S_i is different from the actual treatment received D_i

Compliance Types

- Four possible compliance types

		$D_i(0)$	
		0	1
$D_i(1)$	0	never-taker (NT)	defier (D)
	1	complier (C)	always-taker (AT)

- The true compliance type S is not observed on all units
- The observed cells of Z and D are mixture of different compliance types

Z	D	S
0	0	[C, NT]
0	1	[AT, D]
1	0	[NT, D]
1	1	[C, AT]

- Additional assumptions are required to identify the causal effects for each type.

ITT Effects

- A key observation: the compliance type S_i does not change according to the assignment Z_i. It can be viewed as a (latent) baseline characteristic
- Define ITT effects for each compliance type:

$$\tau_s^{ITT} = \mathbb{E}[Y_i(1) - Y_i(0)|S_i = s],$$

for $s = c, n, a, d$.

- The global ITT may be written as the weighted average of the ITT effects across the four subpopulations:

$$\tau_Y^{ITT} = \pi_c \tau_c^{ITT} + \pi_n \tau_n^{ITT} + \pi_a \tau_a^{ITT} + \pi_d \tau_d^{ITT}$$

where π_s is the proportion of units of type s and the treatment received was $D_i = d$

Re-exam IV Identification Assumptions

In the context of randomized experiments with noncompliance

- A1: SUTVA ✓
- A2: random assignment (i.e. IV) has a (strong) effect on the actual treatment $cov(Z_i, D_i) > 0$ ✓ (Note: this is different from $cov(Z_i, S_i)$, which is 0 here due to randomization of Z)
- A3: IV is randomized – hold by design ✓
- A4: ER, no direct effect of coin flip on outcome, i.e. ruling out placebo effect. Mostly reasonable.

$$Y_i(0) = Y_i(1), \quad \text{for all } i \in S_i = n, a$$

  Some subtle difference in ER between noncompliers (always-taker, never-taker) and compliers
- A5: Monotonicity $D_i(1) \geq D_i(0)$ for all i. Reasonable in most cases.

Identification of the Causal Effects

- The monotonicity of compliance rules out the existence of defiers, $\pi_d = 0$
- ER implies that $\tau_n^{ITT} = \tau_a^{ITT} = 0$; this is reasonable because for never-takers the assignment does not affect the receipt of the treatment
- ER and monotonicity allow the identification of the ITT effect for compliers

$$\tau_c^{ITT} = \tau_Y^{ITT} / \pi_c$$

- The global ITT may be viewed as a conservative estimate of the treatment effect: with $0 < \tau_c < 1$, we have $\tau^{ITT} < \tau_c^{ITT}$

Identify and estimate distribution of compliance types: two-sided noncompliance

- Under monotonicity
  - the units in the $(Z = 0, D = 1)$ must be always-takers
  - the units in the $(Z = 1, D = 0)$ must be never-takers
  - the units in the $(Z = 1, D = 1)$ can be either compliers or always-takers
  - the units in the $(Z = 0, D = 0)$ can be either compliers or never-takers

- Under randomization, the proportion of never-takers, compliers, and always-takers are the same between the two arms ($Z = 1$ and $Z = 0$)

Identify and estimate distribution of compliance types: two-sided noncompliance

- Combining monotonicity and randomization, we can nonparametrically identify the proportions of each compliance type from observed data

$$\begin{aligned}
\pi_a &= \Pr(D_i(0) = D_i(1) = 1) = \Pr(D_i = 1, Z_i = 0) = E[D_i|Z_i = 0] \\
\pi_n &= \Pr(D_i(0) = D_i(1) = 0) = \Pr(D_i = 0, Z_i = 1) = 1 - E[D_i|Z_i = 1] \\
\pi_c &= \Pr(D_i(0) = 0, D_i(1) = 1) = E[D_i|Z_i = 1] - E[D_i|Z_i = 0]
\end{aligned}$$

- We can use the moment counterpart of the above quantities, to get a moment estimator of π_s:

$$\begin{aligned}
\hat{\pi}_a &= N_{01}/(N_{01} + N_{00}) \\
\hat{\pi}_n &= N_{10}/(N_{11} + N_{10}) \\
\hat{\pi}_c &= 1 - \hat{\pi}_n - \hat{\pi}_a
\end{aligned}$$

where N_{zd} is the number of units in $Z = z, D = d$ cell; N_z is the number of units in arm $Z = z$

Identify and estimate distribution of compliance types

- One-sided noncompliance can be viewed as a special case: there is no always-takers: $\pi_a = 0$
- We can also define the ITT effect of Z on D:
  $\tau_D^{ITT} = E[D_i(1) - D_i(0)]$
- Under randomization:
  $\tau_D^{ITT} = E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = \pi_c$

Complier Average Causal Effect (CACE)

- The ITT effect of the compliers τ_c is also known as the Complier Average Causal Effect (CACE) or Local Average Treatment Effect (LATE, Imbens and Angrist (1994))

$$\tau^{CACE} \equiv \tau_c^{ITT} = E[Y_i(1) - Y_i(0)|S_i = c]$$

- Under A1-A5, CACE is identified as

$$\tau^{CACE} = \tau_Y^{ITT}/\pi_c = \tau_Y^{ITT}/\tau_D^{ITT} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$$

- CACE is a ratio of two causal effects (effect of Z on Y, and Z on D), equivalent to the IV estimand when the random assignment is viewed as an instrument

- To interpret CACE as the causal effect as the treatment received (efficacy), another ER assumption is often (implicitly) made: For all compliers, the effect of the random assignment is only through the effect of the treatment received.

Moment Estimates of CACE

- For one-sided noncompliance (no always-takers)

$$\widehat{\tau}^{CACE} = \frac{\sum_i Y_i Z_i / \sum_i Z_i - \sum_i Y_i (1 - Z_i) / \sum_i (1 - Z_i)}{\sum_i D_i Z_i / \sum_i Z_i}$$

- For two-sided noncompliance

$$\widehat{\tau}^{CACE} = \frac{\sum_i Y_i Z_i / \sum_i Z_i - \sum_i Y_i (1 - Z_i) / \sum_i (1 - Z_i)}{1 - \sum_i D_i (1 - Z_i) / \sum_i (1 - Z_i) - \sum_i (1 - D_i) Z_i / \sum_i Z_i}$$

- Standard errors can be obtained asymptotically or via bootstrap
- Without monotonicity or ER, one can still obtain nonparametric bounds for the effects, but the bounds are often too wide to be informative

Nonparametric identification of stratum average outcomes

- Under monotonicity:
  - the units in the $(Z = 0, D = 1)$ must be always-takers, and can nonparametrically identify the outcome

$$E[Y_i(1)|S_i = a] = E[Y_i|D_i = 1, Z_i = 0]$$

  - the units in the $(Z = 1, D = 0)$ must be never-takers, and can nonparametrically identify the outcome

$$E[Y_i(0)|S_i = n] = E[Y_i|D_i = 0, Z_i = 1]$$

  - the units in the $(Z = 1, D = 1)$ can be either compliers or always-takers

$$E[Y_i|D_i = 1, Z_i = 1] = \frac{\pi_c}{\pi_c + \pi_a} E[Y_i(1)|S_i = c] + \frac{\pi_a}{\pi_c + \pi_a} E[Y_i(1)|S_i = a]$$

  - the units in the $(Z = 0, D = 0)$ can be either compliers or never-takers

$$E[Y_i|D_i = 0, Z_i = 0] = \frac{\pi_c}{\pi_c + \pi_n} E[Y_i(0)|S_i = c] + \frac{\pi_n}{\pi_c + \pi_n} E[Y_i(0)|S_i = n]$$

- In combination, we identify $E[Y_i(0)|S_i = c]$ and $E[Y_i(1)|S_i = c]$

Extrapolating to the Full Population

- We can learn from these averages whether there is any evidence of heterogeneity in outcomes by compliance status, by comparing
  - the pair of average outcomes of $Y_i(0)$: $E[Y_i(0)|S_i = n]$ vs. $E[Y_i(0)|S_i = c]$, and
  - the pair of average outcomes of $Y_i(1)$: $E[Y_i(1)|S_i = a]$ and $E[Y_i(1)|S_i = c]$

- If compliers, never-takers and always-takers are found to be substantially different, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types

Example: Vitamin A Supplement

Sommer and Zeger, 1991, Stat in Med

- Goal: Study the effect of vitamin A supplements on infant mortality in Indonesia
- The vitamin supplements were randomly assigned to villages, but some of the individuals in villages assigned to the treatment group failed to receive them
- None of the individuals assigned to the control group received the supplements
- So noncompliance is one-sided
- Outcome Y: survival of the infant (binary)
- Z, D are binary

Example: Vitamin A Supplement

Sommer and Zeger (1991), SIM

Table 23.1: Sommer–Zeger Vitamin Supplement Data

Compliance Type	Assignment $Z_{\text{obs},i}$	Vitamin Supplements $W_{\text{obs},i}$	Survival $Y_{\text{obs},i}$	Number of Units (Total 23,682)
co or nc	0	0	0	74
co or nc	0	0	1	11,514
nc	1	0	0	34
nc	1	0	1	2385
co	1	1	0	12
co	1	1	1	9663

- $\widehat{\tau}_y^{ITT} = (34+12)/(34+12+2385+9663) - 74/(74+11514) = -0.00258$
- $\widehat{\tau}_d^{ITT} = (12+9663)/(34+2385+12+9663) = 0.79998$
- $\widehat{\tau}^{CACE} = \widehat{\tau}_y^{ITT}/\widehat{\tau}_d^{ITT} = -0.00258/0.79998 = -0.00323$

Connection to IV estimation

- Vytlacil (2002, Econometrica) showed: Assumptions A1-A5 are equivalent to a nonparametric version of the latent index model in economics

$$\begin{aligned} D_i^* &= \alpha_0 + \alpha_1 Z_i + \varepsilon_{i1} \\ Y_i &= \beta_0 + \beta_1 D_i + \varepsilon_{i2} \end{aligned}$$

  where $D_i^* = \mathbf{1}\{D_i > 0\}$, and Z_i independent of $\varepsilon_1, \varepsilon_2$

- Here D^* is the latent index, interpreted as the expected net utility of choosing treatment A versus B
- Most IV used in economics are not in the context of randomized experiments; instead, IV is usually a natural experiment.
- In econometrics, estimation is usually conducted via two-stage least square 2SLS
- An excellent review of IV for causal inference: Baiocchi, Cheng, Small (2014, SIM)

Sources of IV in Economic Studies

The biggest challenge in using IV methods is finding a good IV.
Several common sources of IVs in economic studies

- The half or quarter of the year of birth:
  - When one was born in the year is largely randomized by nature, and does not affect later income directly.
  - It does directly affects at what age you goes to school first, may creates one year of difference in the year of school entrance.
  - due to the compulsory education requirement, it can create one year difference in education, which in turn affects income or other labor outcomes
- Tax
  - Goal: Study the effect of smoking (Z) on health (Y)
  - IV: tobacco tax
  - Reasoning: tobacco tax rate is controlled by government, it does not directly affect one's health. But it affects the price of tobacco, thus in turn affects how much one smokes, which affects one's health.
- More examples in Angrist and Pischke (2008, Mostly Harmless Econometrics)

Sources of IV in Health Studies

- Several common sources of IVs in health studies
  - Randomized encouragement design
  - Distance to specialty care provider (the McClellan example)
  - Calendar time: Variations in the use of one treatment versus another over time could result from changes in guidelines; changes in formularies or reimbursement policies; changes in physician preference. Challenge: how to take care of natural time trend in outcome and treatment?
  - Insurance plan
  - Preference-based IVs
  - Genes: Mendelian Randomization

Preference-based IVs

- Idea: Find naturally occurring variation in medical practice patterns at the level of geographic region, hospital, or individual physician; and then use whether the region/hospital/individual physician has a high or low use of treatment A (compared with treatment B) as the IV.
- Potential problems:
  - preference-based IVs may have a direct effect on the outcome
  - preference-based IVs often involve clustered data. Related research and guideline (e.g. estimand, model, se) is largely lacking

Genes as IV: Mendelian randomization

- Goal: study the causal effect of some exposure on health outcome, e.g. smoking on lung cancer, or blood pressure on stoke
- The exposure (phenotype) is usually confounded
- IV: polymorphism of some genes (genetic variants)
- Reasoning: the assortment of genes from parents to offspring is random, i.e. genetic variants are randomly assigned conditional on a parent's genes
- Potential problems: (i) affects of a single allele is often too tiny, multiple weak IVs; (ii) unmeasured confounding through population stratification; (iii) genetic linkage; (iv) potential direct effect through pleiotropy
- Active research area in genetic epidemiology and statistics

Extensions and Open Questions

- In practice, often IV and treatment is not binary, can be multiple and continuous IVs
- From the 2SLS perspective, extension to these settings seems to be straightforward: change a single binary IV to multiple or continuous IV and then perform 2SLS
- But problematic from a causal inference perspective:
  - does the 2SLS estimate still have a causal interpretation?
  - how to extend to complex settings like clustered data or heterogenous treatment effect?
- Even with the Angrist, Imbens and Rubin approach to binary IV and binary treatment case, 2SLS is not efficient because it does not utilize the mixture (of compliance type) structure of the causal formulation (Imbens and Rubin, 1997; Hirano et al. 2003)

References

Angrist JD, Pischke JS. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American statistical Association. 1995 Jun 1;90(430):431-42.

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American statistical Association, 91(434), 444-455.

Baiocchi, M., Cheng, J., Small, D. S. (2014). Instrumental variable methods for causal inference. Statistics in medicine, 33(13), 2297-2340.

Bound, J., D. Jaeger, and R. Baker, (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. Journal of the American Statistical Association, 90, 443-450.

Imbens, G. W. (2014). Instrumental Variables: An Econometrician's Perspective. Statistical Science, 29(3), 323-358.

Imbens, G., and J. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica, 61(2), 467-476.

Heckman, J. J. and Vytlacil, E. J. (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc. Natn. Acad. Sci. USA, 96, 4730-4734.

References

McClellan M, McNeil B, Newhouse J. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. Journal of the American Medical Association 1994; 272(11):859-866.

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

Sommer, A., Zeger, S. L. (1991). On estimating efficacy from clinical trials. Statistics in medicine, 10(1), 45-52.

Terza JV, Basu A, Rathouz PJ. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of health economics. 27(3):531-43.

Vytlacil E. Independence, monotonicity, and latent indexmodels: an equivalence result. Econometrica 2002; 70:331-341.

Wooldridge, J., (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

Imbens, G. W., Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. The annals of statistics, 305-327.