The Statistics of Causal Inference in the Social Sciences

Political Science 236A
Statistics 239A

Jasjeet S. Sekhon

UC Berkeley

October 15, 2010

Robustness of Randomization Inference

- Randomization inference is often more robust than alternatives
- An example of this is when method of moment estimators become close to unidentified in finite samples
- An illustration is provide by the Wald (1940) estimator, which is often used to estimate treatment effects with instruments

What is IV Used For?

- IV used for causal inference
- The most compelling common example: estimate the average treatment effect when there is one-way non-compliance in an experiment.
- Assumptions are weak in this case. In most other examples, the behavioral assumptions are strong.

Experimental Data

- If assignment to treatment is randomized, the inference problem is straightforward because the two groups are from the same population: $\{Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i\}$.
- The Average Treatment Effect (ATE) can be estimated:

$$
\begin{aligned}
\bar{\tau} &\equiv \bar{Y}_{i1} - \bar{Y}_{i0} \\
\bar{\tau} &= E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) \\
&= E(Y_i|T_i = 1) - E(Y_i|T_i = 0)
\end{aligned}
$$

Observational Data: Selection on Observables

- With observational data, the treatment and control groups are not drawn from the same population.
- Progress can be made if we assume that the two groups are comparable once we condition on observable covariates denoted by X_i.
- This is the conditional independence assumption:

  $$\{Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i | X_i\},$$

  the reasonableness of this assumption depends on the design.

Instrumental Variables (IV) as an Alternative

- IV methods solve the problem of missing or unknown control IF the instruments are valid
- Simple example under the unnecessary assumption of constant effects. This assumption is only used to simply the presentation here:

$$\alpha = Y_{1i} - Y_{0i}$$
$$Y_{0i} = \beta + \epsilon_i,$$

where $\beta \equiv \mathbb{E}[Y_{0i}]$

- When we do permutation inference, we will assume under the null, unlike the Wald estimator, that the potential outcomes are fixed.

Towards the Wald Estimator I

- The potential outcomes model can now be written:

$$Y_i = \beta + \alpha T_i + \epsilon_i, \quad (1)$$

  But T_i is likely correlated with ϵ_i.
- Suppose: a third variable Z_i which is correlated with T_i, but is unrelated to Y for any other reason—i.e., $Y_{0i} \perp\!\!\!\perp Z_i$ and $\mathbb{E}[\epsilon_i \mid Z_i] = 0$.
- Z_i is said to be an IV or "an instrument" for the causal effect of T on Y

Towards the Wald Estimator II

- Suppose that Z_i is dichotomous (0,1). Then

$$\alpha = \frac{\mathbb{E}[Y_i \mid Z = 1] - \mathbb{E}[Y_i \mid Z = 0]}{\mathbb{E}[T_i \mid Z = 1] - \mathbb{E}[T_i \mid Z = 0]} \tag{2}$$

- The sample analog of this equation is called the Wald estimator, since it first appear in Wald (1940) on errors-in-variables problems.
- More general versions for continuous, multi-valued, or multiple instruments.
- Problem: what if in finite samples the denominator in Eq 2 is close to zero?

Conditions for a Valid Instrument

- Instrument Relevance:

  $$\mathrm{cov}(Z, T) \neq 0$$

- Instrument Exogeneity:

  $$\mathrm{cov}(Z, \epsilon) = 0$$

- These conditions ensure that the part of X that is correlated with Z only contains exogenous variation
- Instrument relevance is testable
- Instrument exogeneity is NOT testable. It must be true by design

Weak Instruments

- A weak instrument is one where the denominator in Eq 2 is close to zero.
- This poses two distinct problems:
  - if the instrument is extremely weak, it may provide little or no useful information
  - commonly used statistical methods for IV do not accurately report this lack of information

Two Stage Least Squares Estimator (TSLS)

$$
\begin{aligned}
Y &= \beta T + \epsilon \qquad (3) \\
T &= Z\gamma + \upsilon, \qquad (4)
\end{aligned}
$$

where Y, ϵ, υ are $N \times 1$, and Z is $N \times K$, where K is the number of instruments

- Note we do not assume that $\mathbb{E}(T, \epsilon) = 0$, which is the central problem
- We assume instead $\mathbb{E}(\upsilon \mid Z) = 0$, $\mathbb{E}(\upsilon, \epsilon) = 0$, $\mathbb{E}(Z, \epsilon) = 0$, $\mathbb{E}(Z, T) \neq 0$

TSLS Estimator

$$\hat{\beta}_{iv} = (T'P_zT)^{-1}T'P_zY, \tag{5}$$

where $P_z = Z(Z'Z)^{-1}Z'$, the projection matrix for Z. It can be shown that:

- $\text{plim } \hat{\beta}_{ols} = \beta + \frac{\sigma_{T,\epsilon}}{\sigma_T^2}$
- $\text{plim } \hat{\beta}_{iv} = \beta + \frac{\sigma_{\hat{T},\epsilon}}{\sigma_T^2}$

Quarter of Birth and Returns to Schooling

- Angrist and Krueger (1991) want to estimate the causal effect of education
- This is difficult, so they propose a way to estimate the causal effect of compulsory school attendance on earnings
- Every state has a minimum number δ of years of schooling that all students must have
- But, the laws are written in terms of the age at which a student can leave school
- Because birth dates vary, individual students are required to attend between δ and $\delta + 1$ years of schooling

Quarter of Birth

- The treatment T is years of education
- The instrument Z is quarter of birth (exact birth date to be precise)
- The outcome Y are earnings

The Data

- Census data is used
- 329,509 men born between 1930 and 1939
- Observe: years of schooling, birth date, earnings in 1980
- mean number of years of education is 12.75
- Example instrument: being born in the fourth quarter of the year, which is 1 for 24.5% of sample

How Much Information Is There?

- If the number of years of education is regressed on this quarter-of-birth indicator, the least squares regression coefficient is 0.092 with standard error 0.013
- if log-earnings are regressed on the quarter-of-birth indicator, the coefficient is 0.0068 with standard error 0.0027, so being born in the fourth quarter is associated with about $\frac{2}{3}$% higher earnings.
- Wald estimator: $\dfrac{0.0068}{0.092} \approx 0.074$

Estimates Using the Wald Estimator

Estimate	95% lower	95% upper
Simple Wald Estimator		
0.074	0.019	0.129
Multivariate TSLS Estimator		
0.074	0.058	0.090

Problem

- replace the actual quarter-of-birth variable by a randomly generated instrument that carries no information because it is unrelated to years of education.
- As first reported by Bound, Jaeger, and Baker (1995), the TSLS estimate incorrectly suggests that the data are informative, indeed, very informative when there are many instruments.
- The 95% confidence interval: 0.042, 0.078
- But the true estimate is 0!

Method

- There are S strata with n_s units in stratum s and N subjects in total
- Y_{Csi} is the control ($T = 0$) potential outcome for unit i in strata s
- Y_{tsi} is the potential outcome for unit i in strata s if the unit received treatment $t \neq 0$
- Simplifying assumption (not necessary), effect is proportional:
  $Y_{tsi} - Y_{Csi} = \beta t$.
- t is years of education beyond the minimum that are required by law

Instruments

- In stratum s there is a preset, sorted, fixed list of n_s instrument settings h_{sj}, $j = 1, \cdots, n_s$, where $h_{sj} \leq h_{s,j+1}$ $\forall$ s, j
- $\mathbf{h} = (h_{11}, h_{12}, \cdots, h_{1,n_1}, h_{2,1}, \cdots, h_{S,n_s})^T$
- Instrument settings in **h** are randomly permuted within strata
- Assignment of instrument settings, **z**, is **z=ph** where **p** is a stratified permutation matrix—i.e., an $N \times N$ block diagonal matrix with S blocks, $\mathbf{p}_1, \cdots, \mathbf{p}_S$.
- Block $\mathbf{p_s}$ is an $n_s \times n_s$ permutation matrix—i.e., $\mathbf{p}_s$ is a matrix of 0s and 1s s.t. each row and column sum to 1

Permutations

- Let Ω be the set of all stratified permutation matrices **p**, so Ω is a set containing $|\Omega| = \prod_{s=1}^{S} n_s!$ matrices, where $|\Omega|$ denotes the number of elements of the set Ω
- Pick a random **P** from Ω where $Pr(P = p) = \dfrac{1}{|\Omega|}$ for each $\mathbf{p} \in \Omega$
- Then $\mathbf{Z} = \mathbf{Ph}$ is a random permutation of **h** within strata, so the ith unit in stratum s receives instrument setting Z_{si}

Outcomes

- For each $\mathbf{z}$ there is a $t_{si\mathbf{z}}$ for each unit who then has an outcome $Y_{Csi} + \beta t_{si\mathbf{z}}$
- T_{si} is the dose, treatment value, for unit i in stratum s so $T_{si} = t_{si\mathbf{z}}$
- Let Y_{si} be the response for this unit, so $Y_{si} = Y_{Csi} + \beta T_{si}$
- Write $\mathbf{T} = (T_{11}, \cdots, T_{S,n_s}^{T})$ and $\mathbf{Y} = (Y_{11}, \cdots, Y_{S,n_s}^{T})$

Hypothesis Testing I

- We wish to test $H_0 : \beta = \beta_0$
- Let $\mathbf{q}(\cdot)$ be a method of scoring response such as their ranks within strata
- Let $\rho(\mathbf{Z})$ be some way of scoring the instrument settings such that $\rho(\mathbf{ph}) = \mathbf{p}\rho(\mathbf{h})$ for each $\mathbf{p} \in \Omega$
- The test statistic is $U = \mathrm{q}(\mathrm{Y} - \beta_0(T))^T \rho(\mathrm{Z})$
- For appropriate scores, U can be Wilcoxon's stratified rank sum statistic, the Hodges-Lehmann aligned rank statistic, the stratified Spearman rank correlation, etc

Hypothesis Testing II

- If H_0 were true, $\mathbf{Y} - \beta_0 \mathbf{T} = \mathbf{Y}_C$ would be fixed, not varying with $\mathbf{Z}$: $\mathbf{q}(\mathbf{Y} - \beta_0 \mathbf{T}) = \mathbf{q}(\mathbf{Y}_C) = \mathbf{q}$ would also be fixed
- If the null is false, $\mathbf{Y} - \beta_0 \mathbf{T} = \mathbf{Y}_C + (\beta - \beta_0)\mathbf{T}$ will be related to the dose $\mathbf{T}$ and related to $\mathbf{Z}$
- Our test amounts to looking for an absence of a relationship between $\mathbf{Y} - \beta_0 \mathbf{T}$ and $\mathbf{Z}$.

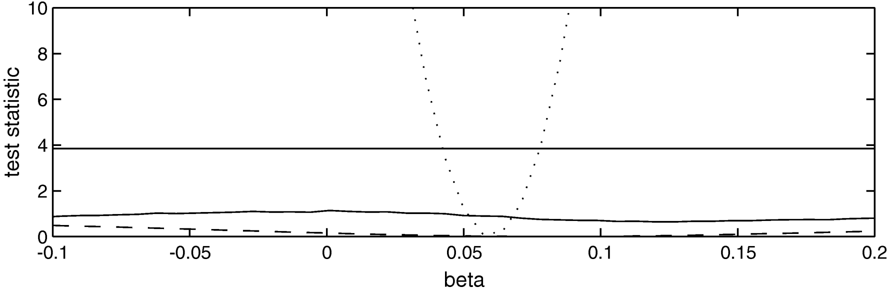
Exact Test

- An exact test computes $\mathbf{q}(\mathbf{Y} - \beta_0 \mathbf{T})$, which is the fixed value $\mathbf{q} = \mathbf{q}(\mathbf{Y}_C)$, in which case $U = \mathbf{q}^T \mathbf{P} \rho(\mathbf{H})$
- The chance that $U \leq u$ under H_0 is the proportion of $\mathbf{p} \in \Omega$ that $\mathbf{q}^T \mathbf{P} \rho(\mathbf{H}) \leq u$

Comparison of instrumental variable estimates with uninformative data

Procedure	95% lower	95% upper
TSLS	0.042	0.078
Permute ranks	-1	1
Permute log-earnings	-1	1

Results with Uninformative Instrument



Dotted line is TSLS; solid line is randomization test using ranks; and dashed line is randomization test using the full observed data

References

This treatment is based on:

- Imbens and Rosenbaum (2005): "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Series A, vol 168(1), 109–126.
- Angrist and Krueger (1991): "Does compulsory school attendance affect earnings?" *QJE* 1991; 106: 979–1019.
- Bound et al. (1995): "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Regressors is Weak," *JASA* 90, June 1995, 443–450.

Angrist, J and AB Krueger. 1991. “Does compulsory school attendance affect earnings?” *Quarterly Journal of Economics* 106: 979–1019.

Bound, J., D. Jaeger, and R. Baker. 1995. “Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Regressors is Weak.” *Journal of the American Statistical Association* 90: 443–450.

Imbens, Guido W. and Paul Rosenbaum. 2005. “Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education.” *Journal of the Royal Statistical Society, Series A* 168: 109–126.