

Causal Inference

Kosuke Imai

Princeton University

October 27, 2010

Readings

- (Required) Lecture notes (available at Blackboard)
- (Required) Imbens, G. and J. Wooldridge. (2009). “Recent Developments in the Econometrics of Program Evaluation” *Journal of Economic Literature*, Vol. 47, pp. 5 – 86.
- (Required) Ho, D. E., K. Imai, G. King, and E. A. Stuart. (2007). “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis*, Vol. 15, pp. 199–236.
- (Suggested) Angrist, J. and Krueger, A. (2001). “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal of Economic Perspectives*, Vol. 15, pp. 69–85.
- (Suggested) Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press. Chapters 4–5.
- (Suggested) Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press. Chapters 1–7.

What We Learned about Causal Inference in POL 572

- Statistical analysis of randomized experiments: Neyman's analysis
- Instrumental variables: statistical analysis of randomized experiments with noncompliance
- Causal mediation analysis: structural equation modeling
- Statistical analysis of fuzzy and sharp regression discontinuity design: use of regression and instrumental variables

- Good causal inference = good design + good analysis

Regression and Causal Inference

- Recall the notation: treatment T_i , potential outcomes $Y_i(t)$, observed outcome Y_i , pre-treatment covariates X_i
- Regression function:

$$\mu(t, x) = \mathbb{E}(Y_i \mid T_i = t, X_i = x)$$

- The Average Treatment Effect (ATE):

$$\tau \equiv \mathbb{E}(Y_i(1) - Y_i(0))$$

- $\tau \stackrel{?}{=} \int \{\mathbb{E}(Y_i \mid T_i = 1, X_i = x) - \mathbb{E}(Y_i \mid T_i = 0, X_i = x)\} dF_{X_i}(x)$
- Recall the distinction between estimand and estimator
- Reminder: please report quantities of interest such as ATE rather than model parameters such as coefficients

Identification vs. Statistical Inference

- Identification: How much can you learn about the estimand if you had an infinite amount of data?
- Statistical Inference: How much can you learn about the estimand from a finite sample?
- Identification precedes statistical inference
- *Partial* (sharp bounds) vs. *Point* identification (point estimates)
- *Parametric* vs. *Nonparametric* identification analysis
- Key questions:
 - 1 What can be learned without making any assumption other than the ones which we know are satisfied by the design?
 - 2 What is a minimum set of assumptions required for the point identification of an estimand?
 - 3 Can we characterize the identification region if we relax some or all of these assumptions?
- **Law of Decreasing Credibility** (Manski): The credibility of inference decreases with the strength of the assumptions maintained

Identification of the Average Treatment Effect

- Assumption 1: Overlap (i.e., no extrapolation)

$$0 < \Pr(T_i = 1 \mid X_i = x) < 1 \text{ for any } x \in \mathcal{X}$$

- Assumption 2: Ignorability (exogeneity, unconfoundedness, no omitted variable, selection on observables, etc.)

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i = x \text{ for any } x \in \mathcal{X}$$

- Under these assumptions:

$$\tau = \mathbb{E}\{\mu(1, X_i) - \mu(0, X_i)\}$$

τ is identified since $F(X_i)$ is identified

- Regression-based Estimator: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \{\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\}$
- Delta method is pain, but the Quasi-Bayesian simulation is easy (Zelig does this for you for many models)

Nonparametric Identification Analysis

- In randomized experiments, these assumptions are satisfied
- In observational studies, their validity is in doubt
- What happens if the assumptions are violated?
- Identification analysis without assuming a model
- Decomposition of asymptotic bias:

$$\begin{aligned} & \mathbb{E}(Y_i(0) \mid T_i = 1) - \mathbb{E}\{\mu(0, X_i)\} \\ = & \int_{S_1 \setminus S} \mathbb{E}(Y_i(0) \mid T_i = 1, X_i = x) dF_{X_i \mid T_i=1}(x) \\ & - \int_{S_0 \setminus S} \mathbb{E}(Y_i(0) \mid T_i = 0, X_i = x) dF_{X_i \mid T_i=0}(x) \\ + & \int_S \mathbb{E}(Y_i(0) \mid T_i = 0, X_i = x) d\{F_{X_i \mid T_i=1}(x) - F_{X_i \mid T_i=0}(x)\} \\ + & \int_S \{\mathbb{E}(Y_i(0) \mid T_i = 1, X_i = x) - \mathbb{E}(Y_i(0) \mid T_i = 0, X_i = x)\} dF_{X_i \mid T_i=1}(x) \end{aligned}$$

Sharp Bounds

- Manski-Robins bounds of $\tau(X_i)$:

$$\begin{aligned} & \{a_1 - \mathbb{E}(Y_i | T_i = 0, X_i)\}\{1 - e(X_i)\} + \{\mathbb{E}(Y_i | T_i = 1, X_i) - b_0\}e(X_i), \\ & \{b_1 - \mathbb{E}(Y_i | T_i = 0, X_i)\}\{1 - e(X_i)\} + \{\mathbb{E}(Y_i | T_i = 1, X_i) - a_0\}e(X_i) \end{aligned}$$

where $e(X_i) = \Pr(T_i = 1 | X_i)$ and $-\infty < a_t \leq Y_i(t) \leq b_t < \infty$

- Special case: $a_t = 0$ and $b_t = 1$:

$$\begin{aligned} & [-\Pr(Y_i = 0 | T_i = 1, X_i)e(X_i) - \Pr(Y_i = 1 | T_i = 0, X_i)\{1 - e(X_i)\}], \\ & \Pr(Y_i = 1 | T_i = 1, X_i)e(X_i) + \Pr(Y_i = 0 | T_i = 0, X_i)\{1 - e(X_i)\}] \end{aligned}$$

- The width of the bounds is 1: “A glass is half empty/full”
- Monotone treatment selection:

$$\begin{aligned} & \{\mathbb{E}(Y_i | T_i = 1, X_i) \Pr(T_i = 1 | X_i) + a \Pr(T_i = 0 | X_i) - \mathbb{E}(Y_i | X_i), \\ & \mathbb{E}(Y_i | X_i) - \mathbb{E}(Y_i | T_i = 0, X_i) \Pr(T_i = 0 | X_i) - a \Pr(T_i = 1 | X_i)\}. \end{aligned}$$

Matching as Nonparametric Preprocessing

- Assume exogeneity holds
- Need to model $\mathbb{E}(Y_i | T_i, X_i)$
- Non-parametric regression – **curse of dimensionality**
- Parametric regression – functional-form/distributional assumptions
- Preprocess the data so that treatment and control groups are similar to each other in terms of the observed pre-treatment covariates
- Goal of matching: achieve **balance**

$$\tilde{F}(X | T = 1) = \tilde{F}(X | T = 0)$$

where $\tilde{F}(\cdot)$ is the *empirical* distribution

- Exact matching: impossible in most cases
- Maximize balance via matching
- Parametric adjustment for remaining imbalance
- Minimal role of statistical models; reduced **model dependence**

The Role of Propensity Score

- The probability of receiving the treatment:

$$\pi(X_i) \equiv \Pr(T_i = 1 \mid X_i)$$

- The balancing property under exogeneity:

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

- Exogeneity given the propensity score:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid \pi(X_i)$$

- Dimension reduction
- But, true propensity score is unknown
- Possible to extend it to non-binary treatment

Methods to Improve Covariate Balance

- **Matching:** Each treated unit is paired with a similar control unit based on the pre-treatment covariates.
- **Subclassification:** Treated and control units are grouped to form subclasses based on the pre-treatment covariates so that within each subclass treated units are similar to control units.
- **Weighting:** Weight each observation within the treated or control groups by the inverse of the probability of being in that group.

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{\pi}(X_i)} \right)$$

or

$$\frac{\sum_{i=1}^n \{T_i Y_i / \hat{\pi}(X_i)\}}{\sum_{i=1}^n \{T_i / \hat{\pi}(X_i)\}} - \frac{\sum_{i=1}^n \{(1 - T_i) Y_i / (1 - \hat{\pi}(X_i))\}}{\sum_{i=1}^n \{(1 - T_i) / (1 - \hat{\pi}(X_i))\}}$$

- The goal of all three methods is to improve balance

Common Matching Methods

- Mahalanobis distance matching

$$\sqrt{(X_i - X_j)^\top \tilde{\Sigma}^{-1} (X_i - X_j)}$$

- Propensity score matching
- One-to-one, one-to-many matching
- Caliper matching
- Subclassification on propensity score
- Optimal/Genetic matching
- Matching with and without replacement

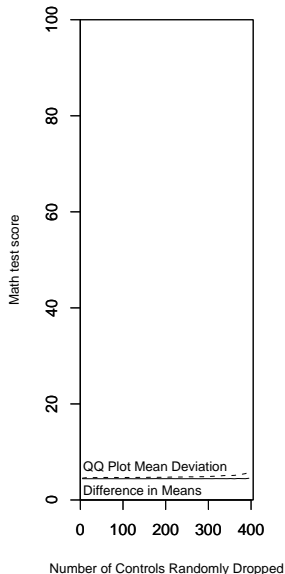
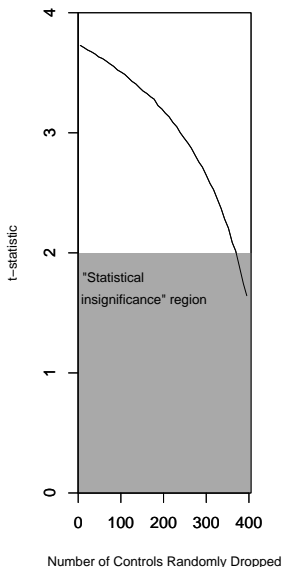
- Which matching method to choose?
- Whatever gives you the “best” balance!

How to Check Balance

- Success of matching method depends on the resulting balance
- How should one assess the balance of matched data?
- Ideally, compare the joint distribution of all covariates for the matched treatment and control groups
- In practice, this is impossible when X is high-dimensional
- Check various lower-dimensional summaries; (standardized) mean difference, variance ratio, empirical CDF, etc.
- Frequent use of **balance test**
 - t test for difference in means for each variable of X
 - other test statistics; e.g., χ^2 , F , Kolmogorov-Smirnov tests
 - statistically insignificant test statistics as a justification for the adequacy of the chosen matching method and/or a stopping rule for maximizing balance

An Illustration of Balance Test Fallacy

- School Dropout Demonstration Assistance Program.
- Treatment: school “restructuring” programs.
- Outcome: dropout rates.
- We look at the baseline math test score.
- “Silly” matching algorithm: randomly selects control units to discard.



Problems with Hypothesis Tests as Stopping Rules

- Balance test is a function of both balance and statistical power
- The more observations dropped, the less power the tests have
- t -test is affected by factors other than balance,

$$\frac{\sqrt{n_m}(\bar{X}_{mt} - \bar{X}_{mc})}{\sqrt{\frac{s_{mt}^2}{r_m} + \frac{s_{mc}^2}{1-r_m}}}$$

- \bar{X}_{mt} and \bar{X}_{mc} are the sample means
 - s_{mt}^2 and s_{mc}^2 are the sample variances
 - n_m is the total number of remaining observations
 - r_m is the ratio of remaining treated units to the total number of remaining observations
-
- Balance is a characteristic of sample rather than population
 - Even in experiments, (pre-randomization) matching is preferred

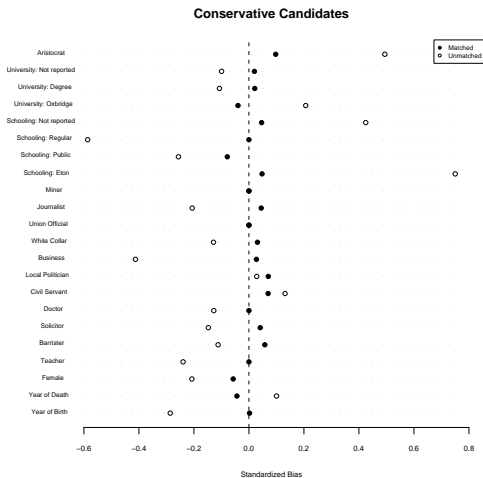
Recent Developments of Matching Methods

- The main problem of matching: balance checking
- Propensity score tautology
- Skip balance checking altogether
- Specify desired degree of balance before matching
- Simple implementation: exact restrictions on key confounders
- Fine matching
- Coarsened exact matching
- Synthetic matching

An Empirical Example

- “Value of political power” by Eggers and Hainmueller (APSR)

Figure 3: Covariate Balance Before and After Matching



Double Robustness Property

- Why care about propensity score?
- Propensity score model specification can be difficult when X_i is high-dimensional
- Doubly-robust estimator:

$$\hat{\tau}_{DR} \equiv \left\{ \frac{1}{n} \sum_{i=1}^n \hat{\mu}(1, X_i) + \frac{1}{n} \sum_{i=1}^n \frac{T_i(Y_i - \hat{\mu}(1, X_i))}{\hat{\pi}(X_i)} \right\} \\ - \left\{ \frac{1}{n} \sum_{i=1}^n \hat{\mu}(0, X_i) + \frac{1}{n} \sum_{i=1}^n \frac{(1 - T_i)(Y_i - \hat{\mu}(0, X_i))}{1 - \hat{\pi}(X_i)} \right\}$$

- Consistent if either the propensity score model or the outcome model is correct
- Efficient if both models are correct

Sensitivity Analysis

- Idea: How large a departure from the key assumption must occur for the conclusions to no longer hold?
- Example: parametric sensitivity analysis (Imbens)

$$Y_i(T_i) \mid T_i, X_i, U_i \stackrel{\text{indep.}}{\sim} \mathcal{N}(\tau T_i + X_i^\top \beta + \delta U_i, \sigma^2)$$
$$\Pr(T_i = 1 \mid X_i, U_i) = \frac{\exp(X_i^\top \gamma + \alpha U_i)}{1 + \exp(X_i^\top \gamma + \alpha U_i)}$$

where U_i is an unobserved binary variable with $p = \Pr(U_i = 1)$

- Sensitivity parameters (p, α, δ) : (p, δ) is easy to interpret
- What about α ? Odds ratio

$$\gamma = \frac{\Pr(T_i = 1 \mid X_i, U_i = 1) / \Pr(T_i = 0 \mid X_i, U_i = 1)}{\Pr(T_i = 1 \mid X_i, U_i = 0) / \Pr(T_i = 0 \mid X_i, U_i = 0)} = \exp(\alpha)$$

R^2 Interpretation

- An alternative interpretation by Imbens:

$$R_Y^2(\alpha, \delta, \rho) = \frac{\tilde{R}_Y^2(\alpha, \delta, \rho) - \tilde{R}_Y^2(0, 0, \rho)}{1 - \tilde{R}_Y^2(0, 0, \rho)}$$

$$R_T^2(\alpha, \delta, \rho) = \frac{\tilde{R}_T^2(\alpha, \delta, \rho) - \tilde{R}_T^2(0, 0, \rho)}{1 - \tilde{R}_T^2(0, 0, \rho)}$$

where $\tilde{R}_Y^2(\alpha, \delta, \rho) = 1 - \hat{\sigma}(\alpha, \delta, \rho) / \text{var}(Y_i)$ and
 $\tilde{R}_T^2(\alpha, \delta, \rho) = \{\hat{\gamma}(\alpha, \delta, \rho)^\top \Sigma_X \hat{\gamma}(\alpha, \delta, \rho) + \alpha^2 \rho(1 - \rho)\} / \{\hat{\gamma}(\alpha, \delta, \rho)^\top \Sigma_X \hat{\gamma}(\alpha, \delta, \rho) + \alpha^2 \rho(1 - \rho) + \pi^2/3\}$

- Beyond parametric sensitivity analysis
- Nonparametric sensitivity analysis: e.g., Rosenbaum's Γ in matched studies

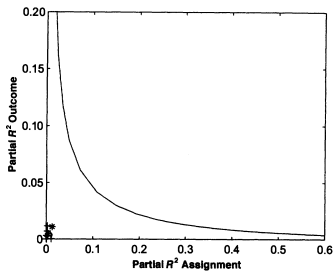


FIGURE 1. LALONDE EXPERIMENTAL SAMPLE

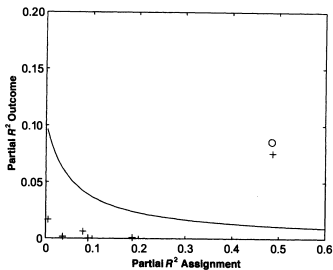


FIGURE 3. LALONDE NONEXPERIMENTAL GAIN SAMPLE

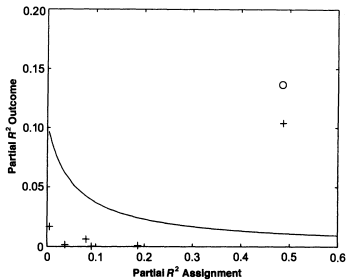


FIGURE 2. LALONDE NONEXPERIMENTAL SAMPLE

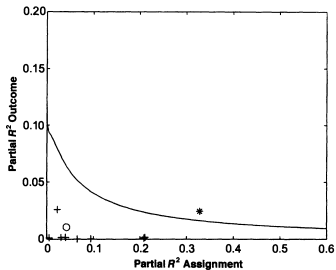


FIGURE 4. LALONDE RESTRICTED SAMPLE

Repeated Observations and Causal Inference

- How to make causal inference with repeated measures?
- Consider a case of two periods: $t = 0, 1$
- Observed outcomes: Y_{0i} and Y_{1i}
- Group indicator: $G_i = 1$ (treatment group), $G_i = 0$ (control group)
- Binary treatment (given only in time $t = 1$): $Z_{ti} = tG_i$
- Potential outcomes: $Y_{0i} = Y_{0i}(0)$ and $Y_{1i} = Y_{1i}(G_i)$
- Pre-treatment (and/or time-invariant) covariates: $X_i = X_{0i}$
- Quantities of interest:

$$\tau_{ATE} \equiv \mathbb{E}(Y_{1i}(1) - Y_{1i}(0)),$$

$$\tau_{ATT} \equiv \mathbb{E}(Y_{1i}(1) - Y_{1i}(0) \mid G_i = 1).$$

Exogeneity with Repeated Measures

- Treatment is “randomized” given X_i and Y_{0i} ,

$$(Y_{1i}(1), Y_{1i}(0)) \perp\!\!\!\perp Z_{1i} \mid (X_i = x, Y_{0i} = y_0).$$

- Overlap of support condition:

$$0 < \Pr(Z_{1i} = 1 \mid X_i = x, Y_{0i} = y_0) < 1$$

- For ATT, we only need

$$Y_{1i}(0) \perp\!\!\!\perp Z_{1i} \mid (X_i = x, Y_{0i} = y_0)$$

- Correspond to the **lagged dependent variable model**:

$$Y_{1i}(Z_{1i}) = \alpha + \beta Z_{1i} + \gamma^\top X_i + \delta Y_{0i} + \epsilon_i,$$

where $\mathbb{E}(\epsilon_i \mid Z_{1i}, X_i, Y_{0i}) = 0$

Difference-in-Differences Models

- Exogeneity assumption may be too strong
- An alternative: Difference-in-differences (DID)

$$\tau_{\text{DID}} = \mathbb{E}(Y_{1i} - Y_{0i} \mid G_i = 1) - \mathbb{E}(Y_{1i} - Y_{0i} \mid G_i = 0)$$

- Also called, **change-score**, before-and-after models
- What is the causal interpretation?
- The key identifying assumption,

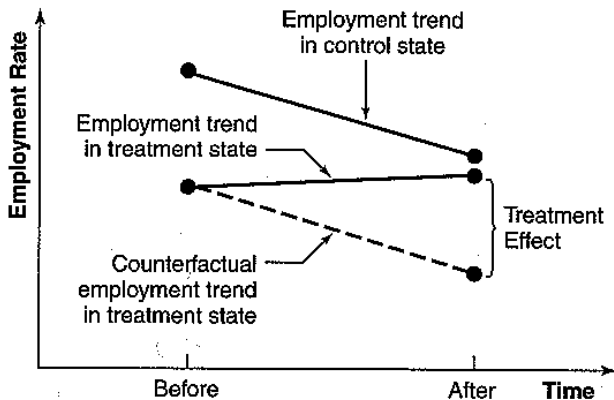
$$\mathbb{E}(Y_{1i}(0) - Y_{0i}(0) \mid G_i = 1, X_i) = \mathbb{E}(Y_{1i}(0) - Y_{0i}(0) \mid G_i = 0, X_i)$$

- Under this assumption,

$$\text{ATT} = \tau_{\text{DID}} \neq \text{ATE}$$

A Graphical Illustration of DID

- Remove unit specific effects and time trend by differencing



Justification based on the Linear Regression

- Model:

$$Y_{ti}(Z_{ti}) = \alpha_i + \beta Z_{ti} + \gamma t + \delta X_i + \epsilon_{ti}$$

- That is,

$$Y_{0i}(0) = \alpha_i + \delta X_i + \epsilon_{0i},$$

$$Y_{1i}(0) = \alpha_i + \gamma + \delta X_i + \epsilon_{1i},$$

$$Y_{1i}(1) = \alpha_i + \beta + \gamma + \delta X_i + \epsilon_{1i}$$

- The key assumption:

$$\mathbb{E}(Y_{1i}(0) - Y_{0i}(0) \mid G_i = g, X_i) = \gamma \quad \text{or} \quad \mathbb{E}(\epsilon_{1i} - \epsilon_{0i} \mid G_i, X_i) = 0$$

- But, allows for the correlation between Z_{1i} and α_i as well as between ϵ_{ti} and α_i

Lagged Dependent Variable or DID Models?

- DID model relies on the linearity assumption
- What happens if you log-transform variables?

$$\begin{aligned} & \mathbb{E}(\log Y_{1i}(0) - \log Y_{0i}(0) \mid G_i = 1, X_i) \\ & \neq \mathbb{E}(\log Y_{1i}(0) - \log Y_{0i}(0) \mid G_i = 0, X_i) \end{aligned}$$

- Thus, better to rewrite the assumption as, for $t = 0, 1$,

$$Y_{ti}(0) \perp\!\!\!\perp Z_{1i} \mid X_i = x,$$

- Note the difference with exogeneity:

$$Y_{1i}(0) \perp\!\!\!\perp Z_{1i} \mid (X_i = x, Y_{0i} = y_0)$$

- The two assumptions are not nested
- Can you justify not adjusting for the observed difference in Y_{0i} ?
- If you match exactly on Y_{0i} , no difference between two models

Concluding Remarks

- Distinction between associational and causal relationships
- Do not just report coefficients: e.g., calculate ATE/ATT
- Causal inference in observational studies requires an additional assumption
- Ignorability: regression, matching, doubly-robust estimator
- Ignorability may be difficult to defend: endogeneity!
 - 1 Conduct randomized experiments in the field or the lab
 - 2 Search for natural experiments
 - Exogenous variations to the treatment
 - Instrumental variables
 - Regression discontinuity design (see POL 572 slides)
 - 3 Sensitivity analysis: how robust is your finding?
- Credibility of untestable assumptions
- Tradeoff between internal and external validity