Statistics and Causal Inference

Kosuke Imai

Princeton University

June 2012
Empirical Implications of Theoretical Models (EITM)
Summer Institute

Three Modes of Statistical Inference

1. Descriptive Inference: summarizing and exploring data
   - Inferring “ideal points” from rollcall votes
   - Inferring “topics” from texts and speeches
   - Inferring “social networks” from surveys
2. Predictive Inference: forecasting out-of-sample data points
   - Inferring future state failures from past failures
   - Inferring population average turnout from a sample of voters
   - Inferring individual level behavior from aggregate data
3. Causal Inference: predicting counterfactuals
   - Inferring the effects of ethnic minority rule on civil war onset
   - Inferring *why* incumbency status affects election outcomes
   - Inferring whether the lack of war among democracies can be attributed to regime types

What is "Identification"?

- Inference: Learn about what you do not observe (*parameters*) from what you do observe (*data*)
- Identification: How much can we learn about parameters from infinite amount of data?

- Ambiguity vs. Uncertainty
- Identification assumptions vs. Statistical assumptions
- Point identification vs. Partial identification

- FURTHER READING: C. F. Manski. (2007). *Identification for Prediction and Decision*. Harvard University Press.

What is Causal Inference?

- Comparison between factual and counterfactual
- Incumbency effect:
  What would have been the election outcome if a candidate were not an incumbent?
- Resource curse thesis:
  What would have been the GDP growth rate without oil?
- Democratic peace theory:
  Would the two countries have escalated crisis in the same situation if they were both autocratic?
- FURTHER READING: Holland, P. (1986). Statistics and causal inference. (with discussions) *Journal of the American Statistical Association*, Vol. 81: 945–960.

Defining Causal Effects

- Units: $i = 1, \ldots, n$
- "Treatment": $T_i = 1$ if treated, $T_i = 0$ otherwise
- *Observed* outcome: Y_i
- Pre-treatment covariates: X_i
- Potential outcomes: $Y_i(1)$ and $Y_i(0)$ where $Y_i = Y_i(T_i)$

Voters	Contact	Turnout		Age	Party ID
i	T_i	$Y_i(1)$	$Y_i(0)$	X_i	X_i
1	1	1	?	20	D
2	0	?	0	55	R
3	0	?	1	40	R
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	0	?	62	D

- Causal effect: $Y_i(1) - Y_i(0)$

The Key Assumptions

- No simultaneity (different from endogeneity)
- No interference between units: $Y_i(T_1, T_2, \ldots, T_n) = Y_i(T_i)$
- Potential violations:
  1. spill-over effects
  2. carry-over effects
- Cluster randomized experiments as a solution (more later)
- Stable Unit Treatment Value Assumption (SUTVA): no interference + "the same version" of the treatment

- Potential outcome is thought to be fixed: data cannot distinguish fixed and random potential outcomes
- But, potential outcomes across units have a distribution
- Observed outcome is random because the treatment is random

- Multi-valued treatment: more potential outcomes for each unit

Causal Effects of Immutable Characteristics

- “No causation without manipulation” (Holland, 1986)
- Immutable characteristics; gender, race, age, etc.
- What does the causal effect of gender mean?

- Causal effect of having a female politician on policy outcomes (Chattopadhyay and Duflo, 2004 *QJE*)
- Causal effect of having a discussion leader with certain preferences on deliberation outcomes (Humphreys *et al*. 2006 *WP*)
- Causal effect of a job applicant’s gender/race on call-back rates (Bertrand and Mullainathan, 2004 *AER*)

Average Treatment Effects

- Sample Average Treatment Effect (SATE):

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

- Population Average Treatment Effect (PATE):

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

- Population Average Treatment Effect for the Treated (PATT):

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$$

- Causal heterogeneity: Zero ATE doesn't mean zero effect for everyone!
- Other quantities: Conditional ATE, Quantile Treatment Effects, etc.

Classical Randomized Experiments

- Units: $i = 1, \dots, n$
- May constitute a simple random sample from a population
- Treatment: $T_i \in \{0, 1\}$
- Outcome: $Y_i = Y_i(T_i)$
- Complete randomization of the treatment assignment
- Exactly n_1 units receive the treatment
- $n_0 = n - n_1$ units are assigned to the control group
- Assumption: for all $i = 1, \dots, n$, $\sum_{i=1}^{n} T_i = n_1$ and

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i, \quad \Pr(T_i = 1) = \frac{n_1}{n}$$

- Estimand $=$ SATE or PATE
- Estimator $=$ Difference-in-means:

$$\hat{\tau} \equiv \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i) Y_i$$

Estimation of Average Treatment Effects

- Key idea (Neyman 1923): Randomness comes from treatment assignment (plus sampling for PATE) alone
- Design-based (randomization-based) rather than model-based
- Statistical properties of $\hat{\tau}$ based on design features

- Define $\mathcal{O} \equiv \{Y_i(0), Y_i(1)\}_{i=1}^n$
- Unbiasedness (over repeated treatment assignments):

$$\begin{aligned}\mathbb{E}(\hat{\tau} \mid \mathcal{O}) &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(T_i \mid \mathcal{O})Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\{1-\mathbb{E}(T_i \mid \mathcal{O})\}Y_i(0) \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \;=\; \text{SATE}\end{aligned}$$

- Over repeated sampling: $\mathbb{E}(\hat{\tau}) = \mathbb{E}(\mathbb{E}(\hat{\tau} \mid \mathcal{O})) = \mathbb{E}(\text{SATE}) = \text{PATE}$

Relationship with Regression

- The model: $Y_i = \alpha + \beta T_i + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$
- Equivalence: least squares estimate $\hat{\beta}$ =Difference in means

- Potential outcomes representation:

$$Y_i(T_i) = \alpha + \beta T_i + \epsilon_i$$

- Constant additive unit causal effect: $Y_i(1) - Y_i(0) = \beta$ for all i
- $\alpha = \mathbb{E}(Y_i(0))$

- A more general representation:

$$Y_i(T_i) = \alpha + \beta T_i + \epsilon_i(T_i) \quad \text{where} \quad \mathbb{E}(\epsilon_i(t)) = 0$$

- $Y_i(1) - Y_i(0) = \beta + \epsilon_i(1) - \epsilon_i(0)$
- $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$
- $\alpha = \mathbb{E}(Y_i(0))$ as before

Bias of Model-Based Variance

- The design-based perspective: use Neyman's exact variance
- What is the bias of the model-based variance estimator?
- Finite sample bias:

$$\begin{aligned} \text{Bias} &= \mathbb{E}\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^n (T_i - \overline{T}_n)^2}\right) - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right) \\ &= \frac{(n_1 - n_0)(n-1)}{n_1 n_0 (n-2)}(\sigma_1^2 - \sigma_0^2) \end{aligned}$$

- Bias is zero when $n_1 = n_0$ or $\sigma_1^2 = \sigma_0^2$
- In general, bias can be negative or positive and does not asymptotically vanish

Robust Standard Error

- Suppose $\text{Var}(\epsilon_i \mid T) = \sigma^2(T_i) \neq \sigma^2$
- Heteroskedasticity consistent robust variance estimator:

$$\widehat{\text{Var}((\hat{\alpha}, \hat{\beta}) \mid T)} = \left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 x_i x_i^\top\right) \left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1}$$

where in this case $x_i = (1, T_i)$ is a column vector of length 2

- Model-based justification: asymptotically valid in the presence of heteroskedastic errors
- Design-based evaluation:

$$\text{Finite Sample Bias} = -\left(\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{n_0^2}\right)$$

- Bias vanishes asymptotically

Cluster Randomized Experiments

- Units: $i = 1, 2, \ldots, n_j$
- Clusters of units: $j = 1, 2, \ldots, m$
- Treatment at cluster level: $T_j \in \{0, 1\}$
- Outcome: $Y_{ij} = Y_{ij}(T_j)$
- Random assignment: $(Y_{ij}(1), Y_{ij}(0)) \perp\!\!\!\perp T_j$
- Estimands at unit level:

$$
\begin{aligned}
\text{SATE} &\equiv \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij}(1) - Y_{ij}(0)) \\
\text{PATE} &\equiv \mathbb{E}(Y_{ij}(1) - Y_{ij}(0))
\end{aligned}
$$

- Random sampling of clusters and units

Merits and Limitations of CREs

- Interference between units within a cluster is allowed
- Assumption: No interference between units of different clusters
- Often easy to implement: Mexican health insurance experiment

- Opportunity to estimate the spill-over effects
- D. W. Nickerson. Spill-over effect of get-out-the-vote canvassing within household (*APSR*, 2008)

- Limitations:
  1. A large number of possible treatment assignments
  2. Loss of statistical power

Design-Based Inference

- For simplicity, assume equal cluster size, i.e., $n_j = n$ for all j
- The difference-in-means estimator:

$$\hat{\tau} \equiv \frac{1}{m_1}\sum_{j=1}^{m} T_j \overline{Y}_j - \frac{1}{m_0}\sum_{j=1}^{m}(1 - T_j)\overline{Y}_j$$

  where $\overline{Y}_j \equiv \sum_{i=1}^{n_j} Y_{ij}/n_j$
- Easy to show $\mathbb{E}(\hat{\tau} \mid \mathcal{O}) = \text{SATE}$ and thus $\mathbb{E}(\hat{\tau}) = \text{PATE}$
- Exact population variance:

$$\text{Var}(\hat{\tau}) = \frac{\text{Var}(\overline{Y_j(1)})}{m_1} + \frac{\text{Var}(\overline{Y_j(0)})}{m_0}$$

- Intracluster correlation coefficient ρ_t:

$$\text{Var}(\overline{Y_j(t)}) = \frac{\sigma_t^2}{n}\{1 + (n-1)\rho_t\} \leq \sigma_t^2$$

Cluster Standard Error

- Cluster robust variance estimator:

$$\widehat{\mathrm{Var}((\hat{\alpha}, \hat{\beta}) \mid T)} = \left(\sum_{j=1}^{m} X_j^\top X_j\right)^{-1} \left(\sum_{j=1}^{m} X_j^\top \hat{\epsilon}_j \hat{\epsilon}_j^\top X_j\right) \left(\sum_{j=1}^{m} X_j^\top X_j\right)^{-1}$$

  where in this case $X_j = [1\, T_j]$ is an $n_j \times 2$ matrix and $\hat{\epsilon}_j = (\hat{\epsilon}_{1j}, \ldots, \hat{\epsilon}_{n_j j})$ is a column vector of length n_j

- Design-based evaluation (assume $n_j = n$ for all j):

$$\text{Finite Sample Bias} = -\left(\frac{\mathbb{V}(\overline{Y_j(1)})}{m_1^2} + \frac{\mathbb{V}(\overline{Y_j(0)})}{m_0^2}\right)$$

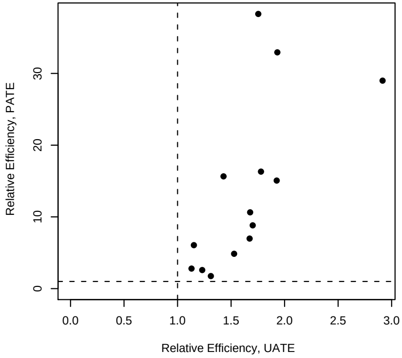
- Bias vanishes asymptotically as $m \to \infty$ with n fixed
- **Implication:** cluster standard errors by the unit of treatment assignment

Example: Seguro Popular de Salud (SPS)

- Evaluation of the Mexican universal health insurance program
- Aim: "provide social protection in health to the 50 million uninsured Mexicans"
- A key goal: reduce out-of-pocket health expenditures
- Sounds obvious but not easy to achieve in developing countries
- Individuals must affiliate in order to receive SPS services
- 100 health clusters nonrandomly chosen for evaluation
- Matched-pair design: based on population, socio-demographics, poverty, education, health infrastructure etc.
- "Treatment clusters": encouragement for people to affiliate
- Data: aggregate characteristics, surveys of $32,000$ individuals

Relative Efficiency of Matched-Pair Design (MPD)

- Compare with completely-randomized design
- Greater (positive) correlation within pair $\rightarrow$ greater efficiency
- UATE: MPD is between 1.1 and 2.9 times more efficient
- PATE: MPD is between 1.8 and 38.3 times more efficient!



Methodological Challenges

- Even randomized experiments often require sophisticated statistical methods
- Deviation from the protocol:
    1. Spill-over, carry-over effects
    2. Noncompliance
    3. Missing data, measurement error
- Beyond the average treatment effect:
    1. Treatment effect heterogeneity
    2. Causal mechanisms
- Getting more out of randomized experiments:
    1. Generalizing experimental results
    2. Deriving individualized treatment rules

Challenges of Observational Studies

- Randomized experiments vs. Observational studies
- Tradeoff between internal and external validity
  - **Endogeneity**: selection bias
  - Generalizability: sample selection, Hawthorne effects, realism
- Statistical methods cannot replace good research design
- "Designing" observational studies
  - Natural experiments (haphazard treatment assignment)
  - Examples: birthdays, weather, close elections, arbitrary administrative rules and boundaries
- "Replicating" randomized experiments
- Key Questions:
  1. Where are the counterfactuals coming from?
  2. Is it a credible comparison?

A Close Look at Fixed Effects Regression

- Fixed effects models are a primary workhorse for causal inference
- Used for stratified experimental and observational data
- Also used to adjust for unobservables in observational studies:
  - "Good instruments are hard to find ..., so we'd like to have other tools to deal with unobserved confounders. This chapter considers ... strategies that use data with a time or cohort dimension to control for unobserved but fixed omitted variables"
    (Angrist & Pischke, *Mostly Harmless Econometrics*)
  - "fixed effects regression can scarcely be faulted for being the bearer of bad tidings" (Green *et al.*, *Dirty Pool*)
- Common claim: Fixed effects models are superior to matching estimators because the latter can only adjust for observables
- **Question:** What are the exact causal assumptions underlying fixed effects regression models?

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.)

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

- Conditional expectation function: $\mu(t, x) = \mathbb{E}(Y_i(t) \mid T_i = t, X_i = x)$
- 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 simulation is easy (Zelig)

Matching and Regression in Cross-Section Settings

Units	**1**	**2**	**3**	**4**	**5**
Treatment status	**T**	**T**	**C**	**C**	**T**
Outcome	Y_1	Y_2	Y_3	Y_4	Y_5

- Estimating the Average Treatment Effect (ATE) via matching:

$$\begin{array}{rcl} Y_1 & - & \frac{1}{2}(Y_3 + Y_4) \\ Y_2 & - & \frac{1}{2}(Y_3 + Y_4) \\ \frac{1}{3}(Y_1 + Y_2 + Y_5) & - & Y_3 \\ \frac{1}{3}(Y_1 + Y_2 + Y_5) & - & Y_4 \\ Y_5 & - & \frac{1}{2}(Y_3 + Y_4) \end{array}$$

Matching Representation of Simple Regression

- Cross-section simple linear regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Binary treatment: $X_i \in \{0, 1\}$
- Equivalent matching estimator:

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y_i(1)} - \widehat{Y_i(0)} \right)$$

where

$$\widehat{Y_i(1)} = \begin{cases} Y_i & \text{if } X_i = 1 \\ \frac{1}{\sum_{i'=1}^{N} X_{i'}} \sum_{i'=1}^{N} X_{i'} Y_{i'} & \text{if } X_i = 0 \end{cases}$$

$$\widehat{Y_i(0)} = \begin{cases} \frac{1}{\sum_{i'=1}^{N} (1 - X_{i'})} \sum_{i'=1}^{N} (1 - X_{i'}) Y_{i'} & \text{if } X_i = 1 \\ Y_i & \text{if } X_i = 0 \end{cases}$$

- Treated units matched with the average of non-treated units

One-Way Fixed Effects Regression

- Simple (one-way) FE model:

$$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$$

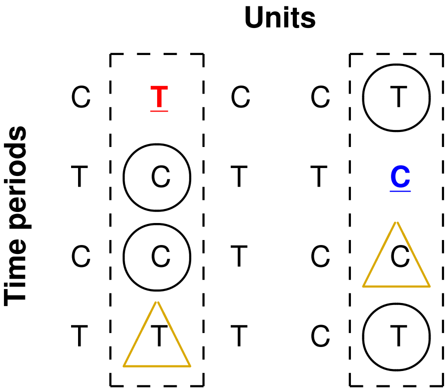
- Commonly used by applied researchers:
  - Stratified randomized experiments (Duflo *et al.* 2007)
  - Stratification and matching in observational studies
  - Panel data, both experimental and observational
- $\hat{\beta}_{FE}$ may be biased for the ATE even if X_{it} is exogenous within each unit
- It converges to the weighted average of conditional ATEs:

$$\hat{\beta}_{FE} \xrightarrow{p} \frac{\mathbb{E}\{\mathrm{ATE}_i \, \sigma_i^2\}}{\mathbb{E}(\sigma_i^2)}$$

where $\sigma_i^2 = \sum_{t=1}^{T}(X_{it} - \overline{X}_i)^2/T$

- How are counterfactual outcomes estimated under the FE model?
- Unit fixed effects $\Longrightarrow$ within-unit comparison

Mismatches in One-Way Fixed Effects Model



- T: treated observations
- C: control observations
- **Circles**: Proper matches
- **Triangles**: “Mismatches” $\Longrightarrow$ attenuation bias

Matching Representation of Fixed Effects Regression

Proposition 1

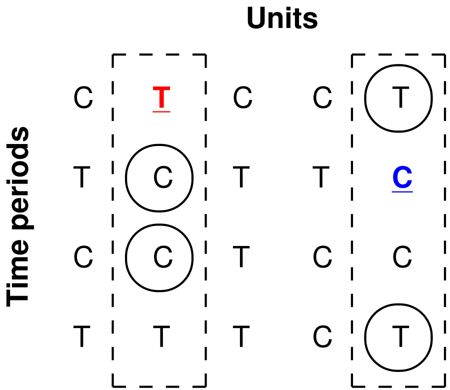
$$\hat{\beta}^{FE} = \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)\right\},$$

$$\widehat{Y_{it}(x)} = \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{T-1}\sum_{t' \neq t} Y_{it'} & \text{if } X_{it} = 1 - x \end{cases} \quad \textit{for} \quad x = 0, 1$$

$$K = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it} \cdot \frac{1}{T-1}\sum_{t' \neq t}(1 - X_{it'}) + (1 - X_{it}) \cdot \frac{1}{T-1}\sum_{t' \neq t} X_{it'}\right\}.$$

- K: average proportion of proper matches across all observations
- More mismatches $\Longrightarrow$ larger adjustment
- Adjustment is required except very special cases
- “Fixes” attenuation bias but this adjustment is not sufficient
- Fixed effects estimator is a special case of matching estimators

Unadjusted Matching Estimator



- Consistent if the treatment is exogenous within each unit
- Only equal to fixed effects estimator if heterogeneity in either treatment assignment or treatment effect is non-existent

Unadjusted Matching = **Weighted** FE Estimator

Proposition 2

The unadjusted matching estimator

$$\hat{\beta}^M = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)$$

where

$$\widehat{Y_{it}(1)} = \begin{cases} Y_{it} & \text{if } X_{it} = 1 \\ \frac{\sum_{t'=1}^{T} X_{it'} Y_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it} = 0 \end{cases} \quad \text{and} \quad \widehat{Y_{it}(0)} = \begin{cases} \frac{\sum_{t'=1}^{T}(1-X_{it'})Y_{it'}}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if } X_{it} = 1 \\ Y_{it} & \text{if } X_{it} = 0 \end{cases}$$

is equivalent to the weighted fixed effects model

$$(\hat{\alpha}^M, \hat{\beta}^M) = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \alpha_i - \beta X_{it})^2$$

$$W_{it} \equiv \begin{cases} \frac{T}{\sum_{t'=1}^{T} X_{it'}} & \text{if} \quad X_{it} = 1, \\ \frac{T}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if} \quad X_{it} = 0. \end{cases}$$

Equal Weights

Treatment					Weights				
C	**T**	C	C	T	0	**1**	0	0	$\frac{1}{2}$
T	C	T	T	**C**	0	$\frac{1}{2}$	0	0	**1**
C	C	T	C	C	0	$\frac{1}{2}$	0	0	0
T	T	T	C	T	0	0	0	0	$\frac{1}{2}$

Different Weights

Treatment					Weights				
C	T	C	C	T	0	1	0	0	$\frac{3}{4}$
T	C	T	T	C	0	$\frac{2}{3}$	0	0	1
C	C	T	C	C	0	$\frac{1}{3}$	0	0	0
T	T	T	C	T	0	0	0	0	$\frac{1}{4}$

- Any within-unit matching estimator leads to weighted fixed effects regression with particular weights
- We derive regression weights given *any* matching estimator for various quantities (ATE, ATT, etc.)

First Difference = Matching = Weighted One-Way FE

- $\Delta Y_{it} = \beta \Delta X_{it} + \epsilon_{it}$ where $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\Delta X_{it} = X_{it} - X_{i,t-1}$

Treatment

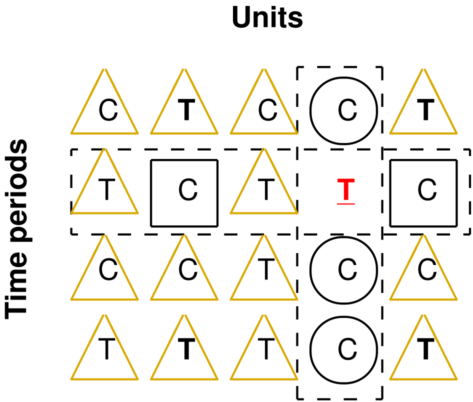
C	T	C	C	T
T	C	T	T	C
C	C	T	C	C
T	T	T	C	T

Weights

0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0

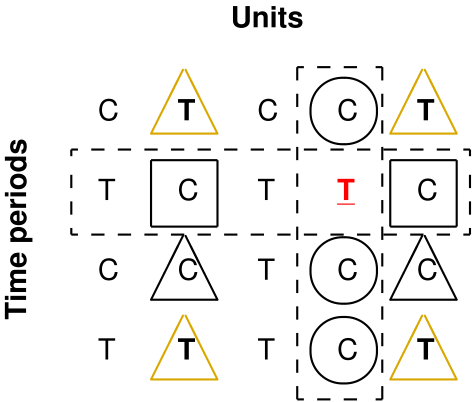
Mismatches in Two-Way FE Model

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it}$$



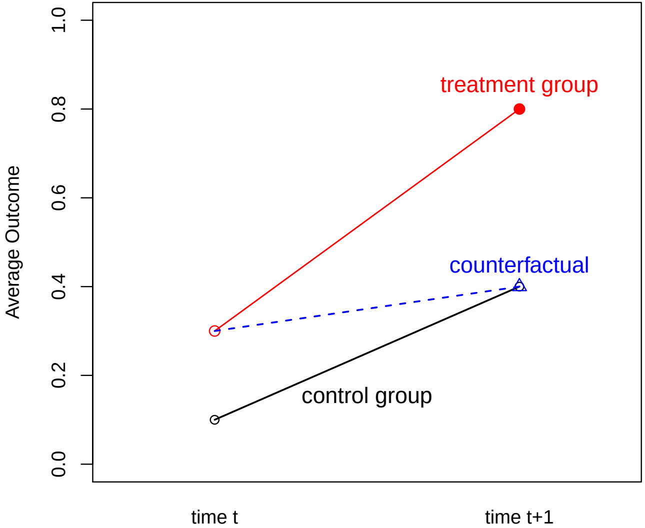
- **Triangles**: Two kinds of mismatches
  - Same treatment status
  - Neither same unit nor same time

Mismatches in Weighted Two-Way FE Model

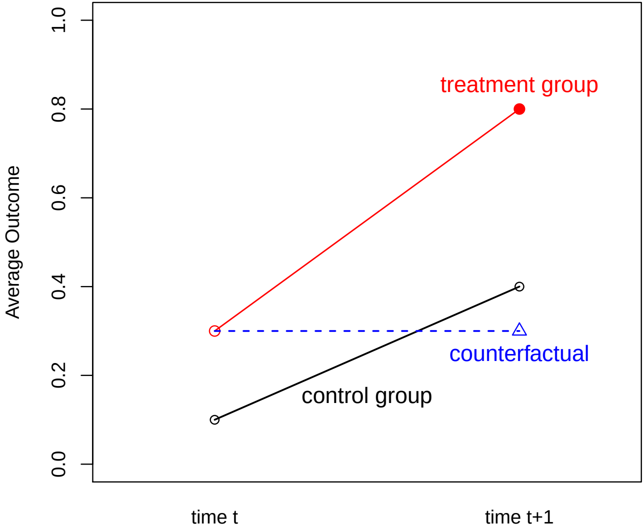


- Some mismatches can be eliminated
- You can NEVER eliminate them all

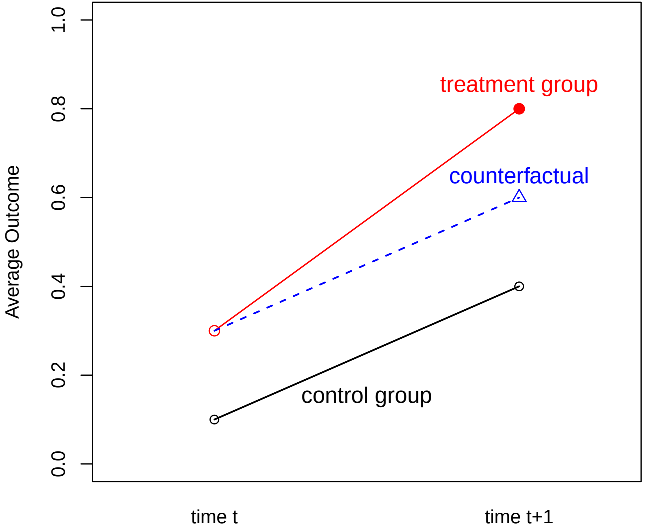
Cross Section Analysis = Weighted **Time** FE Model



First Difference = Weighted **Unit** FE Model



What about Difference-in-Differences (DiD)?



- 2×2: standard two-way fixed effects
- General setting: Multiple time periods, repeated treatments

Treatment

C	**T**	C	C	**T**
T	C	T	**T**	C
C	C	T	C	C
T	**T**	T	C	**T**

Weights

0	0	0	0	0
0	$\frac{1}{2}$	0	**1**	$\frac{1}{2}$
0	$\frac{1}{2}$	0	1	$\frac{1}{2}$
0	0	0	0	0

- Weights can be negative $\Longrightarrow$ the method of moments estimator
- Fast computation is available

Effects of GATT Membership on International Trade

1. Controversy
   - Rose (2004): No effect of GATT membership on trade
   - Tomz et al. (2007): Significant effect with non-member participants

2. The central role of fixed effects models:
   - Rose (2004): one-way (year) fixed effects for dyadic data
   - Tomz *et al.* (2007): two-way (year and dyad) fixed effects
   - Rose (2005): "I follow the profession in placing most confidence in the fixed effects estimators; I have no clear ranking between country-specific and country pair-specific effects."
   - Tomz *et al.* (2007): "We, too, prefer FE estimates over OLS on both theoretical and statistical ground"

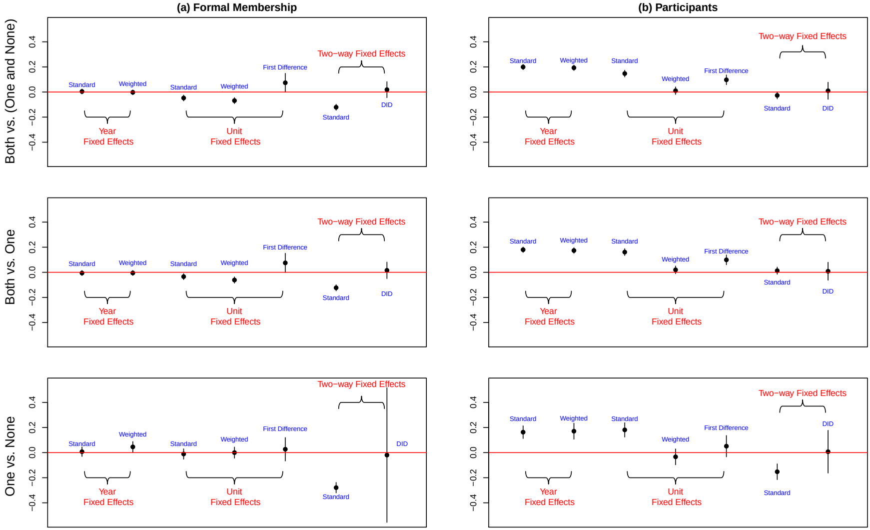
Data and Methods

- Data
  - Data set from Tomz et al. (2007)
  - Effect of GATT: 1948 – 1994
  - 162 countries, and 196,207 (dyad-year) observations
- Year fixed effects model: standard and weighted

$$\ln Y_{it} = \alpha_t + \beta X_{it} + \delta^\top Z_{it} + \epsilon_{it}$$

  - X_{it}: *Formal* membership/*Participant* (1) Both vs. One, (2) One vs. None, (3) Both vs. One/None
  - Z_{it}: 15 dyad-varying covariates (e.g., log product GDP)
- Year fixed effects: standard, weighted, and first difference
- Two-way fixed effects: standard and difference-in-differences

Empirical Results



Matching as Nonparametric Preprocessing

- Assume exogeneity holds: matching does NOT solve endogeneity
- Need to model $\mathbb{E}(Y_i \mid T_i, X_i)$
- Parametric regression – functional-form/distributional assumptions $\Longrightarrow$ model dependence
- Non-parametric regression $\Longrightarrow$ curse of dimensionality
- Preprocess the data so that treatment and control groups are similar to each other w.r.t. the observed pre-treatment covariates

- Goal of matching: achieve balance $=$ independence between T and X
- "Replicate" randomized treatment w.r.t. observed covaraites
- Reduced model dependence: minimal role of statistical modeling

Sensitivity Analysis

- Consider a simple pair-matching of treated and control units
- Assumption: treatment assignment is "random"
- Difference-in-means estimator

- Question: How large a departure from the key (untestable) assumption must occur for the conclusions to no longer hold?
- Rosenbaum's sensitivity analysis: for any pair j,

$$\frac{1}{\Gamma} \leq \frac{\Pr(T_{1j}=1)/\Pr(T_{1j}=0)}{\Pr(T_{2j}=1)/\Pr(T_{2j}=0)} \leq \Gamma$$

- Under ignorability, $\Gamma = 1$ for all j
- How do the results change as you increase Γ?
- Limitations of sensitivity analysis
- FURTHER READING: P. Rosenbaum. *Observational Studies*.

The Role of Propensity Score

- The probability of receiving the treatment:

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

- The balancing property:

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

- Exogeneity given the propensity score (under exogeneity given covariates):

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

- Dimension reduction
- But, true propensity score is unknown: propensity score tautology (more later)

Classical Matching Techniques

- Exact matching
- Mahalanobis distance matching: $\sqrt{(X_i - X_j)^\top \widetilde{\Sigma}^{-1}(X_i - X_j)}$
- Propensity score matching
- One-to-one, one-to-many, and subclassification
- Matching with caliper

- Which matching method to choose?
- Whatever gives you the "best" balance!
- Importance of substantive knowledge: propensity score matching with exact matching on key confounders

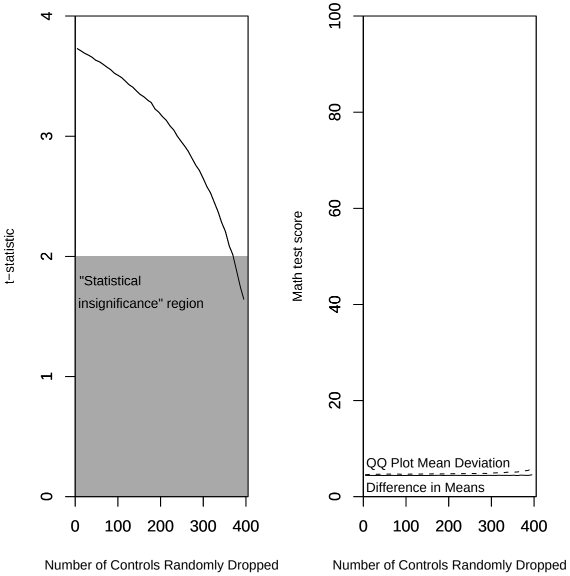
- FURTHER READING: Rubin (2006). *Matched Sampling for Causal Effects* (Cambridge UP)

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



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,

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

  - $\overline{X}_{mt}$ and $\overline{X}_{mc}$ are the sample means
  - s^2_{mt} and s^2_{mc} 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

Advances in Matching Methods

- The main problem of matching: balance checking
- Skip balance checking all together
- Specify a balance metric and optimize it

- Optimal matching: minimize sum of distances
- Genetic matching: maximize minimum p-value
- Coarsened exact matching: exact match on binned covariates
- SVM matching: find the largest, balanced subset

Inverse Propensity Score Weighting

- Matching is inefficient because it throws away data
- Weighting by inverse propensity score

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

- An improved weighting scheme:

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

- Unstable when some weights are extremely small

Efficient Doubly-Robust Estimators

- The estimator by Robins *et al.* :

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

- Consistent if either the propensity score model or the outcome model is correct
- (Semiparametrically) Efficient
- FURTHER READING: Lunceford and Davidian (2004, *Stat. in Med.*)

Propensity Score Tautology

- Propensity score is unknown
- Dimension reduction is purely theoretical: must model T_i given X_i
- Diagnostics: covariate balance checking
- In practice, adhoc specification searches are conducted
- Model misspecification is always possible

- Theory (Rubin *et al.*): ellipsoidal covariate distributions
  $\Longrightarrow$ equal percent bias reduction
- Skewed covariates are common in applied settings

- Propensity score methods can be sensitive to misspecification

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified
- Setup:
  - 4 covariates X_i^*: all are *i.i.d.* standard normal
  - Outcome model: linear model
  - Propensity score model: logistic model with linear predictors
  - Misspecification induced by measurement error:
    - $X_{i1} = \exp(X_{i1}^*/2)$
    - $X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)$
    - $X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3$
    - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$
- Weighting estimators to be evaluated:
  1. Horvitz-Thompson
  2. Inverse-probability weighting with normalized weights
  3. Weighted least squares regression
  4. Doubly-robust least squares regression

Weighting Estimators Do Fine If the Model is Correct

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(1) Both models correct					
$n = 200$	HT	-0.01	0.68	13.07	23.72
	IPW	-0.09	-0.11	4.01	4.90
	WLS	0.03	0.03	2.57	2.57
	DR	0.03	0.03	2.57	2.57
$n = 1000$	HT	-0.03	0.29	4.86	10.52
	IPW	-0.02	-0.01	1.73	2.25
	WLS	-0.00	-0.00	1.14	1.14
	DR	-0.00	-0.00	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	-0.32	-0.17	12.49	23.49
	IPW	-0.27	-0.35	3.94	4.90
	WLS	-0.07	-0.07	2.59	2.59
	DR	-0.07	-0.07	2.59	2.59
$n = 1000$	HT	0.03	0.01	4.93	10.62
	IPW	-0.02	-0.04	1.76	2.26
	WLS	-0.01	-0.01	1.14	1.14
	DR	-0.01	-0.01	1.14	1.14

Weighting Estimators Are Sensitive to Misspecification

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(3) Outcome model correct					
$n = 200$	HT	24.72	0.25	141.09	23.76
	IPW	2.69	-0.17	10.51	4.89
	WLS	-1.95	0.49	3.86	3.31
	DR	0.01	0.01	2.62	2.56
$n = 1000$	HT	69.13	-0.10	1329.31	10.36
	IPW	6.20	-0.04	13.74	2.23
	WLS	-2.67	0.18	3.08	1.48
	DR	0.05	0.02	4.86	1.15
(4) Both models incorrect					
$n = 200$	HT	25.88	-0.14	186.53	23.65
	IPW	2.58	-0.24	10.32	4.92
	WLS	-1.96	0.47	3.86	3.31
	DR	-5.69	0.33	39.54	3.69
$n = 1000$	HT	60.60	0.05	1387.53	10.52
	IPW	6.18	-0.04	13.40	2.24
	WLS	-2.68	0.17	3.09	1.47
	DR	-20.20	0.07	615.05	1.75

Smith and Todd (2005, *J. of Econometrics*)

- LaLonde (1986; *Amer. Econ. Rev.*):
  - Randomized evaluation of a job training program
  - Replace experimental control group with another non-treated group
  - Current Population Survey and Panel Study for Income Dynamics
  - Many evaluation estimators didn't recover experimental benchmark
- Dehejia and Wahba (1999; *J. of Amer. Stat. Assoc.*):
  - Apply propensity score matching
  - Estimates are close to the experimental benchmark
- Smith and Todd (2005):
  - Dehejia & Wahba (DW)'s results are sensitive to model specification
  - They are also sensitive to the selection of comparison sample

Propensity Score Matching Fails Miserably

- One of the most difficult scenarios identified by Smith and Todd:
  - LaLonde experimental sample rather than DW sample
  - Experimental estimate: \$886 (s.e. $=$ 488)
  - PSID sample rather than CPS sample
- Evaluation bias:
  - Conditional probability of being in the experimental sample
  - Comparison between experimental control group and PSID sample
  - "True" estimate $=0$
  - Logistic regression for propensity score
  - One-to-one nearest neighbor matching with replacement

Propensity score model	Estimates
Linear	-835
	(886)
Quadratic	-1620
	(1003)
Smith and Todd (2005)	-1910
	(1004)

Covariate Balancing Propensity Score

- Recall the dual characteristics of propensity score
  1. Conditional probability of treatment assignment
  2. Covariate balancing score
- Implied moment conditions:
  1. Score equation:

$$\mathbb{E}\left\{\frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)}\right\} = 0$$

  2. Balancing condition:

$$\mathbb{E}\left\{\frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)}\right\} = 0$$

where $\widetilde{X}_i = f(X_i)$ is any vector-valued function

Generalized Method of Moments (GMM) Framework

- Over-identification: more moment conditions than parameters
- GMM (Hansen 1982):

$$\hat{\beta}_{\text{GMM}} = \underset{\beta \in \Theta}{\operatorname{argmin}} \ \bar{g}_\beta(T,X)^\top \Sigma_\beta(T,X)^{-1} \bar{g}_\beta(T,X)$$

where

$$\bar{g}_\beta(T,X) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\begin{pmatrix} \frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)} \\ \frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)} \end{pmatrix}}_{g_\beta(T_i, X_i)}$$

- "Continuous updating" GMM estimator with the following Σ:

$$\Sigma_\beta(T,X) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(g_\beta(T_i,X_i) g_\beta(T_i,X_i)^\top \mid X_i)$$

- Newton-type optimization algorithm with MLE as starting values

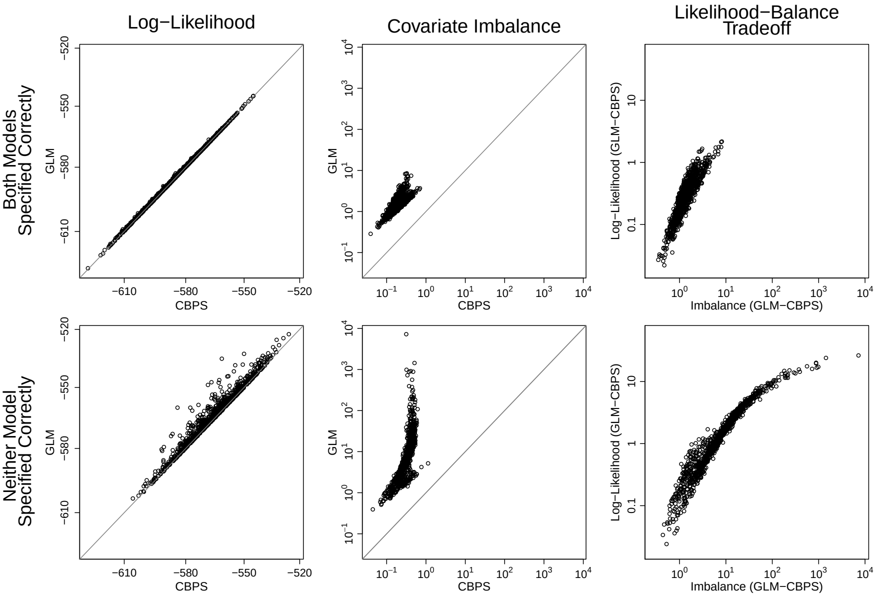
Revisiting Kang and Schafer (2007)

		Bias				RMSE			
Sample size	Estimator	GLM	Balance	CBPS	True	GLM	Balance	CBPS	True
(1) Both models correct									
$n = 200$	HT	−0.01	2.02	0.73	0.68	13.07	4.65	4.04	23.72
	IPW	−0.09	0.05	−0.09	−0.11	4.01	3.23	3.23	4.90
	WLS	0.03	0.03	0.03	0.03	2.57	2.57	2.57	2.57
	DR	0.03	0.03	0.03	0.03	2.57	2.57	2.57	2.57
$n = 1000$	HT	−0.03	0.39	0.15	0.29	4.86	1.77	1.80	10.52
	IPW	−0.02	0.00	−0.03	−0.01	1.73	1.44	1.45	2.25
	WLS	−0.00	−0.00	−0.00	−0.00	1.14	1.14	1.14	1.14
	DR	−0.00	−0.00	−0.00	−0.00	1.14	1.14	1.14	1.14
(2) Propensity score model correct									
$n = 200$	HT	−0.32	1.88	0.55	−0.17	12.49	4.67	4.06	23.49
	IPW	−0.27	−0.12	−0.26	−0.35	3.94	3.26	3.27	4.90
	WLS	−0.07	−0.07	−0.07	−0.07	2.59	2.59	2.59	2.59
	DR	−0.07	−0.07	−0.07	−0.07	2.59	2.59	2.59	2.59
$n = 1000$	HT	0.03	0.38	0.15	0.01	4.93	1.75	1.79	10.62
	IPW	−0.02	−0.00	−0.03	−0.04	1.76	1.45	1.46	2.26
	WLS	−0.01	−0.01	−0.01	−0.01	1.14	1.14	1.14	1.14
	DR	−0.01	−0.01	−0.01	−0.01	1.14	1.14	1.14	1.14

CBPS Makes Weighting Methods Work Better

		Bias				RMSE			
Sample size	Estimator	GLM	Balance	CBPS	True	GLM	Balance	CBPS	True
(3) Outcome model correct									
$n = 200$	HT	24.72	0.33	-0.47	0.25	141.09	4.55	3.70	23.76
	IPW	2.69	-0.71	-0.80	-0.17	10.51	3.50	3.51	4.89
	WLS	-1.95	-2.01	-1.99	0.49	3.86	3.88	3.88	3.31
	DR	0.01	0.01	0.01	0.01	2.62	2.56	2.56	2.56
$n = 1000$	HT	69.13	-2.14	-1.55	-0.10	1329.31	3.12	2.63	10.36
	IPW	6.20	-0.87	-0.73	-0.04	13.74	1.87	1.80	2.23
	WLS	-2.67	-2.68	-2.69	0.18	3.08	3.13	3.14	1.48
	DR	0.05	0.02	0.02	0.02	4.86	1.16	1.16	1.15
(4) Both models incorrect									
$n = 200$	HT	25.88	0.39	-0.41	-0.14	186.53	4.64	3.69	23.65
	IPW	2.58	-0.71	-0.80	-0.24	10.32	3.49	3.50	4.92
	WLS	-1.96	-2.01	-2.00	0.47	3.86	3.88	3.88	3.31
	DR	-5.69	-2.20	-2.18	0.33	39.54	4.22	4.23	3.69
$n = 1000$	HT	60.60	-2.16	-1.56	0.05	1387.53	3.11	2.62	10.52
	IPW	6.18	-0.87	-0.72	-0.04	13.40	1.86	1.80	2.24
	WLS	-2.68	-2.69	-2.70	0.17	3.09	3.14	3.15	1.47
	DR	-20.20	-2.89	-2.94	0.07	615.05	3.47	3.53	1.75

CBPS Sacrifices Likelihood for Better Balance



Revisiting Smith and Todd (2005)

- Evaluation bias: "true" bias $= 0$
- CBPS improves propensity score matching across specifications and matching methods
- However, specification test rejects the null

	1-to-1 Nearest Neighbor			Optimal 1-to-N Nearest Neighbor		
Specification	GLM	Balance	CBPS	GLM	Balance	CBPS
Linear	−835	−559	−302	−885	−257	−38
	(886)	(898)	(873)	(435)	(492)	(488)
Quadratic	−1620	−967	−1040	−1270	−306	−140
	(1003)	(882)	(831)	(406)	(407)	(392)
Smith & Todd	−1910	−1040	−1313	−1029	−672	−32
	(1004)	(860)	(800)	(413)	(387)	(397)

Standardized Covariate Imbalance

- Covariate imbalance in the (Optimal 1–to–N) matched sample
- Standardized difference-in-means

	Linear			Quadratic			Smith & Todd		
	GLM	Balance	CBPS	GLM	Balance	CBPS	GLM	Balance	CBPS
Age	−0.060	−0.035	−0.063	−0.060	−0.035	−0.063	−0.031	0.035	−0.013
Education	−0.208	−0.142	−0.126	−0.208	−0.142	−0.126	−0.262	−0.168	−0.108
Black	−0.087	0.005	−0.022	−0.087	0.005	−0.022	−0.082	−0.032	−0.093
Married	0.145	0.028	0.037	0.145	0.028	0.037	0.171	0.031	0.029
High school	0.133	0.089	0.174	0.133	0.089	0.174	0.189	0.095	0.160
74 earnings	−0.090	0.025	0.039	−0.090	0.025	0.039	−0.079	0.011	0.019
75 earnings	−0.118	0.014	0.043	−0.118	0.014	0.043	−0.120	−0.010	0.041
Hispanic	0.104	−0.013	0.000	0.104	−0.013	0.000	0.061	0.034	0.102
74 employed	0.083	0.051	−0.017	0.083	0.051	−0.017	0.059	0.068	0.022
75 employed	0.073	−0.023	−0.036	0.073	−0.023	−0.036	0.099	−0.027	−0.098
Log-likelihood	−326	−342	−345	−293	−307	−297	−295	−231	−296
Imbalance	0.507	0.264	0.312	0.544	0.304	0.300	0.515	0.359	0.383

Extensions to Other Causal Inference Settings

- Propensity score methods are widely applicable
- This means that CBPS is also widely applicable
- Potential extensions:
    1. Non-binary treatment regimes
    2. Causal inference with longitudinal data
    3. Generalizing experimental estimates
    4. Generalizing instrumental variable estimates
- All of these are situations where balance checking is difficult

Concluding Remarks

- Matching methods do:
  - make causal assumptions transparent by identifying counterfactuals
  - make regression models robust by reducing model dependence
- Matching methods cannot solve endogeneity
- Only good research design can overcome endogeneity
- Recent advances in matching methods
  - directly optimize balance
  - the same idea applied to propensity score
- Next methodological challenges: panel data
  - Fixed effects regression assumes no carry-over effect
  - They do not model dynamic treatment regimes

Coping with Endogeneity in Observational Studies

- Selection bias in observational studies
- Two research design strategies:
  1. Find a plausibly exogenous treatment
  2. Find a plausibly exogenous instrument
- A valid instrument satisfies the following conditions
  1. Exogenously assigned – no confounding
  2. It monotonically affects treatment
  3. It affects outcome only through treatment – no direct effect
- Challenge: plausibly exogenous instruments with no direct effect tends to be weak

Partial Compliance in Randomized Experiments

- Unable to force all experimental subjects to take the (randomly) assigned treatment/control
- Intention-to-Treat (ITT) effect $\neq$ treatment effect
- Selection bias: self-selection into the treatment/control groups

- Political information bias: effects of campaign on voting behavior
- Ability bias: effects of education on wages
- Healthy-user bias: effects of exercises on blood pressure

- Encouragement design: randomize the encouragement to receive the treatment rather than the receipt of the treatment itself

Potential Outcomes Notation

- Randomized encouragement: $Z_i \in \{0, 1\}$
- Potential treatment variables: $(T_i(1), T_i(0))$
  1. $T_i(z) = 1$: would receive the treatment if $Z_i = z$
  2. $T_i(z) = 0$: would not receive the treatment if $Z_i = z$
- Observed treatment receipt indicator: $T_i = T_i(Z_i)$
- Observed and potential outcomes: $Y_i = Y_i(Z_i, T_i(Z_i))$
- Can be written as $Y_i = Y_i(Z_i)$

- No interference assumption for $T_i(Z_i)$ and $Y_i(Z_i, T_i)$
- Randomization of encouragement:

$$(Y_i(1), Y_i(0), T_i(1), T_i(0)) \perp\!\!\!\perp Z_i$$

- But $(Y_i(1), Y_i(0)) \not\perp\!\!\!\perp T_i \mid Z_i = z$, i.e., selection bias

Principal Stratification Framework

- Imbens and Angrist (1994, *Econometrica*); Angrist, Imbens, and Rubin (1996, *JASA*)
- Four principal strata (latent types):
  - compliers $(T_i(1), T_i(0)) = (1, 0)$,
  - non-compliers $\begin{cases} \textit{always}-\textit{takers} & (T_i(1), T_i(0)) = (1, 1), \\ \textit{never}-\textit{takers} & (T_i(1), T_i(0)) = (0, 0), \\ \textit{defiers} & (T_i(1), T_i(0)) = (0, 1) \end{cases}$
- Observed and principal strata:

	$Z_i = 1$	$Z_i = 0$
$T_i = 1$	Complier/Always-taker	Defier/Always-taker
$T_i = 0$	Defier/Never-taker	Complier/Never-taker

Instrumental Variables and Causality

- Randomized encouragement as an instrument for the treatment
- Two additional assumptions
  1. Monotonicity: No defiers

$$T_i(1) \geq T_i(0) \quad \text{for all } i.$$

  2. Exclusion restriction: Instrument (encouragement) affects outcome only through treatment

$$Y_i(1, t) = Y_i(0, t) \quad \text{for } t = 0, 1$$

  Zero ITT effect for always-takers and never-takers

- ITT effect decomposition:

$$\begin{aligned} \text{ITT} &= \text{ITT}_c \times \Pr(\text{compliers}) + \text{ITT}_a \times \Pr(\text{always} - \text{takers}) \\ &\quad + \text{ITT}_n \times \Pr(\text{never} - \text{takers}) \\ &= \text{ITT}_c \Pr(\text{compliers}) \end{aligned}$$

IV Estimand and Interpretation

- IV estimand:

$$
\begin{aligned}
\mathrm{ITT}_c &= \frac{\mathrm{ITT}}{\Pr(\text{compliers})} \\
&= \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(T_i \mid Z_i = 1) - \mathbb{E}(T_i \mid Z_i = 0)} \\
&= \frac{\mathrm{Cov}(Y_i, Z_i)}{\mathrm{Cov}(T_i, Z_i)}
\end{aligned}
$$

- $\mathrm{ITT}_c =$ Complier Average Treatment Effect (CATE)
- Local Average Treatment Effect (LATE)
- CATE $\neq$ ATE unless ATE for noncompliers equals CATE
- Different encouragement (instrument) yields different compliers
- Debate among Deaton, Heckman, and Imbens in *J. of Econ. Lit.*

Violation of IV Assumptions

- Violation of exclusion restriction:

$$\text{Large sample bias} \quad = \quad \text{ITT}_{\text{noncomplier}} \frac{\Pr(\text{noncomplier})}{\Pr(\text{complier})}$$

- Weak encouragement (instruments)
- Direct effects of encouragement; failure of randomization, alternative causal paths

- Violation of monotonicity:

$$\text{Large sample bias} \quad = \quad \frac{\{\text{CATE} + \text{ITT}_{\text{defier}}\} \Pr(\text{defier})}{\Pr(\text{complier}) - \Pr(\text{defier})}$$

- Proportion of defiers
- Heterogeneity of causal effects

An Example: Testing Habitual Voting

- Gerber *et al.* (2003) *AJPS*
- Randomized encouragement to vote in an election
- Treatment: turnout in the election
- Outcome: turnout in the next election

- Monotonicity: Being contacted by a canvasser would *never* discourage anyone from voting
- Exclusion restriction: being contacted by a canvasser in this election has no effect on turnout in the next election other than through turnout in this election
- CATE: Habitual voting for those who would vote if and only if they are contacted by a canvasser in this election

Multi-valued Treatment

- Angrist and Imbens (1995, *JASA*)
- Two stage least squares regression:

$$\begin{aligned} T_i &= \alpha_2 + \beta_2 Z_i + \eta_i, \\ Y_i &= \alpha_3 + \gamma T_i + \epsilon_i. \end{aligned}$$

- Binary encouragement and binary treatment,
  - $\hat{\gamma} = \widehat{\text{CATE}}$ (no covariate)
  - $\hat{\gamma} \xrightarrow{P} \text{CATE}$ (with covariates)
- Binary encouragement multi-valued treatment
- Monotonicity: $T_i(1) \geq T_i(0)$
- Exclusion restriction: $Y_i(1, t) = Y_i(0, t)$ for each $t = 0, 1, \ldots, K$

- Estimator

$$\hat{\gamma}_{TSLS} \xrightarrow{P} \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(T_i, Z_i)} = \frac{\mathbb{E}(Y_i(1) - Y_i(0))}{\mathbb{E}(T_i(1) - T_i(0))}$$

$$= \sum_{k=0}^{K} \sum_{j=k+1}^{K} w_{jk} \mathbb{E}\left(\frac{Y_i(1) - Y_i(0)}{j-k} \,\middle|\, T_i(1) = j, T_i(0) = k \right)$$

where w_{jk} is the weight, which sums up to one, defined as,

$$w_{jk} = \frac{(j-k)\Pr(T_i(1) = j, T_i(0) = k)}{\sum_{k'=0}^{K} \sum_{j'=k'+1}^{K} (j'-k')\Pr(T_i(1) = j', T_i(0) = k')}.$$

- Easy interpretation under the constant additive effect assumption for every complier type
- Assume encouragement induces at most only one additional dose
- Then, $w_k = \Pr(T_i(1) = k, T_i(0) = k-1)$

Partial Identification of the ATE

- Balke and Pearl (1997, *JASA*)
- Randomized binary encouragement, Z_i
- Binary treatment, $T_i = T_i(Z_i)$
- Suppose exclusion restriction holds
- Binary outcome, $Y_i = Y_i(T_i, Z_i) = Y_i^*(T_i)$
- 16 Latent types defined by $(Y_i(1), Y_i(0), T_i(1), T_i(0))$

$$q(y_1, y_0, t_1, t_0) \equiv \Pr(Y_i^*(1) = y_1, Y_i^*(0) = y_0, T_i(1) = t_1, T_i(0) = t_0)$$

- ATE

$$\begin{aligned} & \mathbb{E}(Y_i^*(1) - Y_i^*(0)) \\ = \; & \sum_{y_0} \sum_{t_1} \sum_{t_0} q(1, y_0, t_1, t_0) - \sum_{y_1} \sum_{t_1} \sum_{t_0} q(y_1, 1, t_1, t_0) \end{aligned}$$

Derivation of Sharp Bounds

- Data generating mechanism implies

$$
\begin{aligned}
\Pr(Y_i = y, T_i = 1 \mid Z_i = 1) &= \sum_{y_0} \sum_{t_0} q(y, y_0, 1, t_0) \\
\Pr(Y_i = y, T_i = 0 \mid Z_i = 1) &= \sum_{y_1} \sum_{t_0} q(y_1, y, 0, t_0) \\
\Pr(Y_i = y, T_i = 1 \mid Z_i = 0) &= \sum_{y_0} \sum_{t_1} q(y, y_0, t_1, 1) \\
\Pr(Y_i = y, T_i = 0 \mid Z_i = 0) &= \sum_{y_1} \sum_{t_1} q(y_1, y, t_1, 0).
\end{aligned}
$$

- Monotonicity (optional): $q(y_1, y_0, 0, 1) = 0$
- Obtain sharp bounds via linear programming algorithms
- Bounds are sometimes informative

Fuzzy Regression Discontinuity Design

- Sharp regression discontinuity design: $T_i = \mathbf{1}\{X_i \geq c\}$
- What happens if we have noncompliance?
- Forcing variable as an instrument: $Z_i = \mathbf{1}\{X_i \geq c\}$
- Potential outcomes: $T_i(z)$ and $Y_i(z, t)$

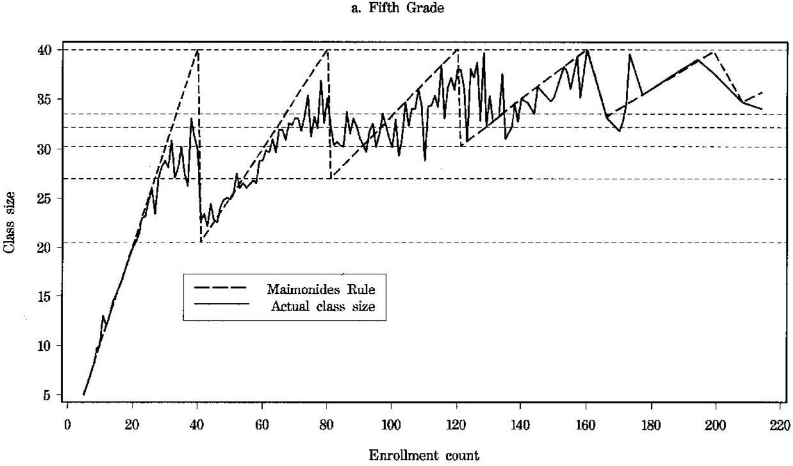
- Monotonicity: $T_i(1) \geq T_i(0)$
- Exclusion restriction: $Y_i(0, t) = Y_i(1, t)$
- $\mathbb{E}(T_i(z) \mid X_i = x)$ and $\mathbb{E}(Y_i(z, T_i(z)) \mid X_i = x)$ are continuous in x
- Estimand: $\mathbb{E}(Y_i(1, T_i(1)) - Y_i(0, T_i(0)) \mid Complier, X_i = c)$
- Estimator:

$$\frac{\lim_{x \downarrow c} \mathbb{E}(Y_i \mid X_i = x) - \lim_{x \uparrow c} \mathbb{E}(Y_i \mid X_i = x)}{\lim_{x \downarrow c} \mathbb{E}(T_i \mid X_i = x) - \lim_{x \uparrow c} \mathbb{E}(T_i \mid X_i = x)}$$

- Disadvantage: external validity

An Example: Class Size Effect (Angrist and Lavy)

- Effect of class-size on student test scores
- Maimonides' Rule: Maximum class size $= 40$



Concluding Remarks

- Instrumental variables in randomized experiments: dealing with partial compliance
- Additional (untestable) assumptions are required
  1. partial identification
  2. sensitivity analysis
- ITT vs. CATE

- Instrumental variables in observational studies: dealing with selection bias
- Validity of instrumental variables requires rigorous justification
- Tradeoff between internal and external validity