Matching and Weighting Methods for Causal Inference

Kosuke Imai
Princeton University

Methods Workshop, Duke University
References to Relevant Papers

- “Covariate Balancing Propensity Score.” Working paper

All papers are available at [http://imai.princeton.edu/research](http://imai.princeton.edu/research)
Software Implementation

- Causal inference with regression: **Zelig: Everyone’s Statistical Software**
- Causal inference with matching: **MatchIt: Nonparametric Preprocessing for Parametric Causal Inference**
- Causal inference with propensity score: **CBPS: Covariate Balancing Propensity Score**
- Causal inference with fixed effects: **wfe: Weighted Fixed Effects Regressions for Causal Inference**

All software is available at

[http://imai.princeton.edu/software](http://imai.princeton.edu/software)
Matching and Weighting

- What is “matching”?
- Grouping observations based on their observed characteristics
  - 1. pairing
  - 2. subclassification
  - 3. subsetting

- What is “weighting”?
- Replicating observations based on their observed characteristics
- All types of matching are special cases with discrete weights

- What matching and weighting methods can do: flexible and robust causal modeling under selection on observables
- What they cannot do: eliminate bias due to unobserved confounding
Matching can be used for randomized experiments too!

- Randomization of treatment $\rightarrow$ unbiased estimates
- Improving efficiency $\rightarrow$ reducing variance
- Why care about efficiency? You care about your results!

- Randomized matched-pair design
- Randomized block design

Intuition: estimation uncertainty comes from pre-treatment differences between treatment and control groups

Mantra (Box, Hunter, and Hunter):

“Block what you can and randomize what you cannot”
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:
  
  \[
  \text{SATE} \equiv \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij}(1) - Y_{ij}(0))
  \]
  
  \[
  \text{PATE} \equiv \mathbb{E}(Y_{ij}(1) - Y_{ij}(0))
  \]

- 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 easier 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
For simplicity, assume equal cluster size, i.e., $n_j = n$ for all $j$

The difference-in-means estimator:

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

where $\bar{Y}_j \equiv \sum_{i=1}^{n_j} Y_{ij} / n_j$

Easy to show $\mathbb{E}(\hat{\tau} \mid \mathcal{O}) = \text{SATE}$ and thus $\mathbb{E}(\hat{\tau}) = \text{PATE}$

Exact population variance:

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

Intracluster correlation coefficient $\rho_t$:

$$
\text{Var}(\bar{Y}_j(t)) = \frac{\sigma_t^2}{n} \{ 1 + (n - 1) \rho_t \} \leq \sigma_t^2
$$
Cluster Standard Error

- Cluster robust “sandwich” variance estimator:

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

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

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

\[
\text{Finite Sample Bias} = - \left( \frac{\text{Var}(Y_j(1))}{m_1^2} + \frac{\text{Var}(Y_j(0))}{m_0^2} \right)
\]

- Bias vanishes asymptotically as \( m \to \infty \) with \( n \) fixed

- Implication: cluster standard errors by the unit of treatment assignment
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 non-randomly 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
Okay, but how should I match/block without the treatment group?

Goal: match/block well on powerful predictors of outcome (prognostic factors)

(Coarsened) Exact matching

Matching based on a similarity measure:

\[
\text{Mahalanobis distance} = \sqrt{(X_i - X_j)^\top \hat{\Sigma}^{-1} (X_i - X_j)}
\]

Could combine the two
Relative Efficiency of Matched-Pair Design (MPD)

- Compare with completely-randomized design
- Greater (positive) correlation within pair $\rightarrow$ greater efficiency
- PATE: MPD is between 1.8 and 38.3 times more efficient!
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?
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 via Zelig
Matching as Nonparametric Preprocessing

- **Reading**: Ho *et al.* Political Analysis (2007)
- 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 $\implies$ model dependence
- Non-parametric regression $\implies$ 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 covariates
- 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 \( \Gamma \)?
- Limitations of sensitivity analysis

Further Reading: P. Rosenbaum. *Observational Studies.*
The Role of Propensity Score

- The probability of receiving the treatment:
  \[ \pi(X_i) \equiv \Pr(T_i = 1 | X_i) \]

- The balancing property (no assumption):
  \[ T_i \perp \perp X_i | \pi(X_i) \]

- Exogeneity given the propensity score (under exogeneity given covariates):
  \[ (Y_i(1), Y_i(0)) \perp \perp T_i | \pi(X_i) \]

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

- Exact matching
- Mahalanobis distance matching: $\sqrt{(X_i - X_j)^\top \hat{\Sigma}^{-1} (X_i - X_j)}$
- 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

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., $\chi^2$, $F$, 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

Kosuke Imai (Princeton)
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, 

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

- $\bar{X}_{mt}$ and $\bar{X}_{mc}$ 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.
Recent 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
- Full matching: subclassification with variable strata size
- Genetic matching: maximize minimum $p$-value
- Coarsened exact matching: exact match on binned covariates
- SVM subsetting: find the largest, balanced subset for general treatment regimes
Inverse Propensity Score Weighting

- Matching is inefficient because it throws away data
- Matching is a special case of weighting
- Weighting by inverse propensity score (Horvitz-Thompson):

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

- Unstable when some weights are extremely small
- An improved weighting scheme:

\[
\frac{\sum_{i=1}^{n} \{ T_i Y_i / \hat{\pi}(X_i) \}}{\sum_{i=1}^{n} \{ T_i / \hat{\pi}(X_i) \}} - \frac{\sum_{i=1}^{n} \{ (1 - T_i) Y_i / (1 - \hat{\pi}(X_i)) \}}{\sum_{i=1}^{n} \{ (1 - T_i) / (1 - \hat{\pi}(X_i)) \}}
\]
Weighting Both Groups to Balance Covariates

- Balancing condition: $E \left\{ \frac{T_i X_i}{\pi(X_i)} - \frac{(1-T_i)X_i}{1-\pi(X_i)} \right\} = 0$
Weighting Control Group to Balance Covariates

- Balancing condition: \( \mathbb{E} \left\{ T_i X_i - \frac{\pi(X_i)(1-T_i)X_i}{1-\pi(X_i)} \right\} = 0 \)
Efficient Doubly-Robust Estimators

- The estimator by Robins et al.:

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

- Consistent if either the propensity score model or the outcome model is correct
- (Semiparametrically) Efficient
- **Further Reading**: Lunceford and Davidian (2004, *Stat. in Med.*)
Marginal Structural Models for Longitudinal Data

- Units $i = 1, \ldots, N$ and time $j = 1, \ldots, J$
- Eventual outcome $Y_i$ measured at time $J$
- Treatment and covariate history: $\bar{T}_{ij}$ and $\bar{X}_{ij}$
- Quantity of interest: (marginal) $\text{ATE} = \mathbb{E}\{Y_i(\bar{t})\}$
- Sequential ignorability assumption:
  \[ Y_i(t) \perp \perp T_{ij} \mid \bar{X}_{ij}, \bar{T}_{i,j-1} \]

- Inverse-probability-of-treatment weight:
  \[ w_i = \frac{1}{P(\bar{T}_{ij} \mid \bar{X}_{ij})} = \prod_{j=1}^{J} \frac{1}{P(T_{ij} \mid T_{i,j-1}, \bar{X}_{ij})} \]

- Stabilized weight: multiply $w_i$ by $P(\bar{T}_{ij})$
- Analysis: weighted regression of $Y_i$ on $\bar{T}_{ij}$

**Further Readings:** Robins et al. (2000), Blackwell (2013)
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
- Tautology: propensity score works only when you get it right!
- In fact, estimated propensity score works even better than true propensity score when the model is correct

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

- Propensity score methods can be sensitive to misspecification
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:
  - $X_{i1} = \exp(X_i^*/2)$
  - $X_{i2} = X_i^*/(1 + \exp(X_i^*) + 10)$
  - $X_{i3} = (X_{i1} X_i^*/25 + 0.6)^3$
  - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$

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 Great If the Model is Correct

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLM</td>
<td>True</td>
<td>GLM</td>
</tr>
<tr>
<td>(1) Both models correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>0.33</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>−0.13</td>
<td>−0.13</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−0.04</td>
<td>−0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.04</td>
<td>−0.04</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>0.01</td>
<td>−0.18</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>0.01</td>
<td>−0.05</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(2) Propensity score model correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>−0.32</td>
<td>−0.17</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>−0.27</td>
<td>−0.35</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−0.07</td>
<td>−0.07</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.07</td>
<td>−0.07</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>−0.02</td>
<td>−0.04</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−0.01</td>
<td>−0.01</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.01</td>
<td>−0.01</td>
</tr>
</tbody>
</table>
Weighting Estimators Are Sensitive to Misspecification

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3) Outcome model correct</td>
<td></td>
<td>GLM</td>
<td>True</td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>24.25</td>
<td>−0.18</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.70</td>
<td>−0.26</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.29</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.08</td>
<td>−0.10</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>41.14</td>
<td>−0.23</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>4.93</td>
<td>−0.02</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.94</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>(4) Both models incorrect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>30.32</td>
<td>−0.38</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.93</td>
<td>−0.09</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.13</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−7.46</td>
<td>0.37</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>101.47</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>5.16</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.95</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−48.66</td>
<td>0.08</td>
</tr>
</tbody>
</table>
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:

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

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

- Score condition is a particular covariate balancing condition!
Estimation and Inference

- **Just-identified CBPS:**
  - Find the values of model parameters that satisfy covariate balancing conditions in the sample
  - Method of moments: \( \# \text{ of parameters} = \# \text{ of balancing conditions} \)

- **Over-identified CBPS:**
  - \( \# \text{ of parameters} < \# \text{ of balancing conditions} \)
  - Generalized method of moments (GMM):
    
    \[
    \hat{\beta} = \arg\min_{\beta \in \Theta} \bar{g}_\beta(T, X)^\top \Sigma^{-1}_\beta \bar{g}_\beta(T, X)
    \]

    where

    \[
    \bar{g}_\beta(T, X) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)} \right) \]

    and \( \Sigma_\beta \) is the covariance of moment conditions
  - Enables misspecification test
<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLM</td>
<td>CBPS1</td>
<td>CBPS2</td>
</tr>
<tr>
<td>(1) Both models correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>0.33</td>
<td>2.06</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>-0.13</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>-0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>-0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>0.01</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(2) Propensity score model correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>-0.05</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>-0.13</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>-0.02</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>
### CBPS Makes Weighting Methods More Robust

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>GLM</td>
<td>CBPS1</td>
<td>CBPS2</td>
</tr>
<tr>
<td>(3) Outcome model correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>24.25</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.70</td>
<td>−1.37</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.29</td>
<td>−2.37</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.08</td>
<td>−0.10</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>41.14</td>
<td>−2.02</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>4.93</td>
<td>−1.39</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.94</td>
<td>−2.99</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>(4) Both models incorrect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>30.32</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.93</td>
<td>−1.26</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.13</td>
<td>−2.20</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−7.46</td>
<td>−2.59</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>101.47</td>
<td>−2.05</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>5.16</td>
<td>−1.44</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.95</td>
<td>−3.01</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−48.66</td>
<td>−3.59</td>
</tr>
</tbody>
</table>
CBPS Sacrifices Likelihood for Better Balance

Both Models Specified Correctly

Neither Model Specified Correctly

Log–Likelihood

Covariate Imbalance

Likelihood–Balance Tradeoff

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?
## 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:**

\[
Y_1 - \frac{1}{2} (Y_3 + Y_4)
\]

\[
Y_2 - \frac{1}{2} (Y_3 + Y_4)
\]

\[
\frac{1}{3} (Y_1 + Y_2 + Y_5) - Y_3
\]

\[
\frac{1}{3} (Y_1 + Y_2 + Y_5) - Y_4
\]

\[
Y_5 - \frac{1}{2} (Y_3 + Y_4)
\]
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:

\[
\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{Y}_{i}(1) - \hat{Y}_{i}(0) \right)
\]

where

\[
\hat{Y}_{i}(1) = \begin{cases} 
\frac{1}{\sum_{i' = 1}^{N} X_{i'}} \sum_{i' = 1}^{N} X_{i'} Y_{i'} & \text{if } X_i = 1 \\
\frac{1}{\sum_{i' = 1}^{N} \left(1 - X_{i'}\right)} \sum_{i' = 1}^{N} \left(1 - X_{i'}\right) Y_{i'} & \text{if } X_i = 0
\end{cases}
\]

\[
\hat{Y}_{i}(0) = \begin{cases} 
Y_i & \text{if } X_i = 1 \\
\frac{1}{\sum_{i' = 1}^{N} \left(1 - X_{i'}\right)} \sum_{i' = 1}^{N} \left(1 - X_{i'}\right) Y_{i'} & \text{if } X_i = 0
\end{cases}
\]

- 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:
  \[
  \hat{\beta}_{FE} \xrightarrow{p} \frac{\mathbb{E}\{\text{ATE}_i \sigma_i^2\}}{\mathbb{E}(\sigma_i^2)}
  \]
  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 \( \xrightarrow{\text{within-unit}} \) comparison
Mismatches in One-Way Fixed Effects Model

- T: treated observations
- C: control observations
- Circles: Proper matches
- Triangles: “Mismatches” → attenuation bias
Proposition 1

\[
\hat{\beta}_{FE} = \frac{1}{K} \left\{ \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \hat{Y}_{it}(1) - \hat{Y}_{it}(0) \right) \right\},
\]

\[
\hat{Y}_{it}(x) \quad = \quad \begin{cases} 
Y_{it} & \text{if } X_{it} = x \\
\frac{1}{T-1} \sum_{t' \neq t} Y_{it'} & \text{if } X_{it} = 1 - x 
\end{cases} \quad \text{for } x = 0, 1
\]

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

- \( K \): average proportion of proper matches across all observations
- More mismatches \( \Rightarrow \) 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
Proposition 2

The unadjusted matching estimator

\[ \hat{\beta}^M = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\hat{Y}_{it}(1) - \hat{Y}_{it}(0)) \]

where

\[ \hat{Y}_{it}(1) = \begin{cases} Y_{it} & \text{if } X_{it} = 1 \\ \frac{\sum_{t'=1}^{T} X_{it'} Y_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it} = 0 \end{cases} \quad \text{and} \quad \hat{Y}_{it}(0) = \begin{cases} \frac{\sum_{t'=1}^{T} (1-X_{it'}) Y_{it'}}{\sum_{t'=1}^{T} (1-X_{it'})} & \text{if } X_{it} = 1 \\ Y_{it} & \text{if } X_{it} = 0 \end{cases} \]

is equivalent to the weighted fixed effects model

\[ (\hat{\alpha}^M, \hat{\beta}^M) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (Y_{it} - \alpha_i - \beta X_{it})^2 \]

\[ W_{it} = \begin{cases} \frac{\sum_{t'=1}^{T} X_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it} = 1, \\ \frac{\sum_{t'=1}^{T} (1-X_{it'})}{\sum_{t'=1}^{T} (1-X_{it'})} & \text{if } X_{it} = 0. \end{cases} \]
Equal Weights

Treatment

Weights

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

0 0 0 0 0
1/2 1 0 0 0
1/2 0 0 0 0
0 0 0 0 1/2

1/2 0 0 0 0
## Different Weights

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>C T C C T</td>
<td>0 1 0 0 3/4</td>
</tr>
<tr>
<td>T C T T C</td>
<td>0 2/3 0 0 1</td>
</tr>
<tr>
<td>C C T C C</td>
<td>0 1/3 0 0 0</td>
</tr>
<tr>
<td>T T T T C</td>
<td>0 0 0 0 1/4</td>
</tr>
</tbody>
</table>

- 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 \( \Delta Y_{it} = Y_{it} - Y_{i,t-1}, \Delta X_{it} = X_{it} - X_{i,t-1} \)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>C T C C T</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>T C T C C</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>C C T C C</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>T T T C T</td>
<td>0 0 0 0</td>
</tr>
</tbody>
</table>
Mismatches in Two-Way FE Model

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

**Units**

**Triangles**: Two kinds of mismatches
- Same treatment status
- Neither same unit nor same time
Some mismatches can be eliminated
You can NEVER eliminate them all
Cross Section Analysis = Weighted Time FE Model

- Average Outcome
- treatment group
- control group
- counterfactual

Time t vs. time t+1
First Difference = Weighted **Unit** FE Model

![Graph showing the comparison between treatment and control groups over time](image)

- **Average Outcome**
  - Treatment group
  - Control group
  - Counterfactual

- **Time**: t, t+1

---

Kosuke Imai (Princeton)  
Matching and Weighting Methods  
Duke (January 18 – 19, 2013)  
51 / 57
What about Difference-in-Differences (DiD)?
General DiD = Weighted Two-Way (Unit and Time) FE

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

### Treatment

<table>
<thead>
<tr>
<th>C</th>
<th>T</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

### Weights

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Weights can be negative $\implies$ 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)
  - 162 countries, and 196,207 (dyad-year) observations
- **Year fixed effects model:** standard and weighted
  \[ \ln Y_{it} = \alpha_t + \beta X_{it} + \delta^T Z_{it} + \epsilon_{it} \]
  - \(X_{it}: \text{ Formal membership/Participant}\) (1) Both vs. One, (2) One vs. None, (3) Both vs. One/None
  - \(Z_{it}: 15 \text{ dyad-varying covariates (e.g., log product GDP)}\)
- **Year fixed effects:** standard, weighted, and first difference
- **Two-way fixed effects:** standard and difference-in-differences
## Empirical Results

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Membership</th>
<th>Year Fixed Effects</th>
<th>Dyad Fixed Effects</th>
<th>Year and Dyad Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both vs. Mix</td>
<td>Formal</td>
<td>0.004</td>
<td>−0.002</td>
<td>−0.048</td>
</tr>
<tr>
<td></td>
<td>(N=196,207)</td>
<td>(0.031)</td>
<td>(0.030)</td>
<td>(0.025)</td>
</tr>
<tr>
<td></td>
<td>White’s p-value</td>
<td>1.000</td>
<td>0.064</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Both vs. Mix</td>
<td>Participants</td>
<td>0.199</td>
<td>0.193</td>
</tr>
<tr>
<td></td>
<td>(N=196,207)</td>
<td>(0.034)</td>
<td>(0.035)</td>
<td>(0.031)</td>
</tr>
<tr>
<td></td>
<td>White’s p-value</td>
<td>0.998</td>
<td>0.000</td>
<td>0.102</td>
</tr>
<tr>
<td>Both vs. One</td>
<td>Formal</td>
<td>−0.006</td>
<td>−0.005</td>
<td>−0.034</td>
</tr>
<tr>
<td></td>
<td>(N=175,814)</td>
<td>(0.031)</td>
<td>(0.031)</td>
<td>(0.025)</td>
</tr>
<tr>
<td></td>
<td>White’s p-value</td>
<td>1.000</td>
<td>0.031</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Both vs. One</td>
<td>Participants</td>
<td>0.180</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>(N=187,651)</td>
<td>(0.035)</td>
<td>(0.036)</td>
<td>(0.031)</td>
</tr>
<tr>
<td></td>
<td>White’s p-value</td>
<td>0.999</td>
<td>0.000</td>
<td>0.086</td>
</tr>
<tr>
<td>One vs. None</td>
<td>Formal</td>
<td>0.007</td>
<td>0.046</td>
<td>−0.011</td>
</tr>
<tr>
<td></td>
<td>(N=109,702)</td>
<td>(0.053)</td>
<td>(0.056)</td>
<td>(0.041)</td>
</tr>
<tr>
<td></td>
<td>White’s p-value</td>
<td>0.276</td>
<td>0.058</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>One vs. None</td>
<td>Participants</td>
<td>0.163</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td>(N=70,298)</td>
<td>(0.072)</td>
<td>(0.079)</td>
<td>(0.062)</td>
</tr>
<tr>
<td></td>
<td>White’s p-value</td>
<td>0.046</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>covariates</td>
<td></td>
<td>dyad-varying covariates</td>
<td>year-varying covariates</td>
<td>no covariate</td>
</tr>
</tbody>
</table>
Concluding Remarks

- Matching methods do:
  - make causal assumptions transparent by identifying counterfactuals
  - make regression models robust by reducing model dependence

- But they 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

- Weighting methods generalize matching methods
  - Sensitive to propensity score model specification
  - Robust estimation of propensity score model

- Next methodological challenges for causal inference: temporal and spatial dynamics, networks effects