PS207 Quantitative Causal Inference

Selection on Observables

Matto Mildenberger

UC Santa Barbara

Special thanks to Chad Hazlett (UCLA) for select slides, used with permission

Observational Studies

- Randomization allows unbiased estimation of $\mathbb{E}[Y_{0i}|D_i = 1]$ via $\mathbb{E}[Y_{0i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}|D_i = 0]$ via $\mathbb{E}[Y_{1i}|D_i = 1]$.

Observational Studies

- Randomization allows unbiased estimation of $\mathbb{E}[Y_{0i}|D_i = 1]$ via $\mathbb{E}[Y_{0i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}|D_i = 0]$ via $\mathbb{E}[Y_{1i}|D_i = 1]$.

  In terms of variables that might influence the outcome, we often refer to this as "balancing" observed and unobserved influences on potential outcomes to prevent confounding

Observational Studies

- Randomization allows unbiased estimation of $\mathbb{E}[Y_{0i}|D_i = 1]$ via $\mathbb{E}[Y_{0i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}|D_i = 0]$ via $\mathbb{E}[Y_{1i}|D_i = 1]$.

  In terms of variables that might influence the outcome, we often refer to this as "balancing" observed and unobserved influences on potential outcomes to prevent confounding

- When we cannot randomize, we do observational studies: adjust for observed covariates and hope that unobservables are balanced

  Back to potential outcomes, we're hoping that after adjustment $\mathbb{E}[Y_{0i}|D_i = 1] \approx \mathbb{E}[Y_{0i}|D_i = 0])$ etc.

Observational Studies

- Randomization allows unbiased estimation of $\mathbb{E}[Y_{0i}|D_i = 1]$ via $\mathbb{E}[Y_{0i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}|D_i = 0]$ via $\mathbb{E}[Y_{1i}|D_i = 1]$.

  In terms of variables that might influence the outcome, we often refer to this as "balancing" observed and unobserved influences on potential outcomes to prevent confounding

- When we cannot randomize, we do observational studies: adjust for observed covariates and hope that unobservables are balanced

  Back to potential outcomes, we're hoping that after adjustment $\mathbb{E}[Y_{0i}|D_i = 1] \approx \mathbb{E}[Y_{0i}|D_i = 0])$ etc.

The rest of the course: assessing and improving observational studies

Observational Studies

- Randomization allows unbiased estimation of $\mathbb{E}[Y_{0i}|D_i = 1]$ via $\mathbb{E}[Y_{0i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}|D_i = 0]$ via $\mathbb{E}[Y_{1i}|D_i = 1]$.

  In terms of variables that might influence the outcome, we often refer to this as "balancing" observed and unobserved influences on potential outcomes to prevent confounding

- When we cannot randomize, we do observational studies: adjust for observed covariates and hope that unobservables are balanced

  Back to potential outcomes, we're hoping that after adjustment $\mathbb{E}[Y_{0i}|D_i = 1] \approx \mathbb{E}[Y_{0i}|D_i = 0])$ etc.

The rest of the course: assessing and improving observational studies

- What is your strategy for identifying the missing potential outcomes from observational data?
- What assumptions does it involve? Can we weaken these?
- Are they credible? Which of them can we test?

The Good, the Bad, and the Ugly

Treatments, Covariates, Outcomes

- Randomized Experiment: Well-defined treatment, clear distinction between covariates and outcomes, control of assignment mechanism

The Good, the Bad, and the Ugly

Treatments, Covariates, Outcomes

- Randomized Experiment: Well-defined treatment, clear distinction between covariates and outcomes, control of assignment mechanism
- Better Observational Study: Well-defined treatment, clear distinction between covariates and outcomes, precise knowledge of assignment mechanism

  Can convincingly answer the following question: Why do two units who are identical on measured covariates receive different treatments?

The Good, the Bad, and the Ugly

Treatments, Covariates, Outcomes

- Randomized Experiment: Well-defined treatment, clear distinction between covariates and outcomes, control of assignment mechanism
- Better Observational Study: Well-defined treatment, clear distinction between covariates and outcomes, precise knowledge of assignment mechanism

  Can convincingly answer the following question: Why do two units who are identical on measured covariates receive different treatments?

- Poorer Observational Study: Hard to say when treatment began or what the treatment really is. Distinction between covariates and outcomes is blurred. No precise knowledge of assignment mechanism

The Good, the Bad, and the Ugly

How were treatments assigned?

- Randomized Experiment: Random assignment

The Good, the Bad, and the Ugly

How were treatments assigned?

- Randomized Experiment: Random assignment
- Better Observational Study: Assignment is not random, but circumstances for the study were chosen so that treatment seems haphazard, or at least not obviously related to potential outcomes (sometimes we refer to these as natural or quasi-experiments)

The Good, the Bad, and the Ugly

How were treatments assigned?

- Randomized Experiment: Random assignment
- Better Observational Study: Assignment is not random, but circumstances for the study were chosen so that treatment seems haphazard, or at least not obviously related to potential outcomes (sometimes we refer to these as natural or quasi-experiments)
- Poorer Observational Study: No attention given to assignment process, units self-select into treatment based on potential outcomes

The Good, the Bad, and the Ugly

Were treated and controls comparable?

- Randomized Experiment: Balance table for observables.
- Better Observational Study: Balance table for observables. Ideally sensitivity analysis for unobservables.
- Poorer Observational Study: No direct assessment of comparability is presented.

The Good, the Bad, and the Ugly

Eliminating plausible alternatives to treatment effects?

- Randomized Experiment: List plausible alternatives, if any. Design attempts to avoid or shed light on possible alternatives (e.g. placebos, multiple treatments, monitoring of treatment, multiple measures). Report on potential attrition and non-compliance,

The Good, the Bad, and the Ugly

Eliminating plausible alternatives to treatment effects?

- Randomized Experiment: List plausible alternatives, if any. Design attempts to avoid or shed light on possible alternatives (e.g. placebos, multiple treatments, monitoring of treatment, multiple measures). Report on potential attrition and non-compliance,

- Better Observational Study: List plausible alternatives and study design includes features that shed light on these alternatives (e.g. multiple control groups, longitudinal covariate data, etc.). Rules out some alternatives, examines sensitivitiy. Requires more work than in experiment since there are usually many more alternatives.

The Good, the Bad, and the Ugly

Eliminating plausible alternatives to treatment effects?

- Randomized Experiment: List plausible alternatives, if any. Design attempts to avoid or shed light on possible alternatives (e.g. placebos, multiple treatments, monitoring of treatment, multiple measures). Report on potential attrition and non-compliance,

- Better Observational Study: List plausible alternatives and study design includes features that shed light on these alternatives (e.g. multiple control groups, longitudinal covariate data, etc.). Rules out some alternatives, examines sensitivitiy. Requires more work than in experiment since there are usually many more alternatives.

- Poorer Observational Study: Alternatives are mentioned in discussion section of the paper or not at all. Over-sells causal claim.

Good Observational Studies

Design features we can use to handle unobservables:

- Design comparisons so that unobservables are likely to be balanced (e.g.sub-samples, groups where treatment assignment was haphazard, etc.)

Good Observational Studies

Design features we can use to handle unobservables:

- Design comparisons so that unobservables are likely to be balanced (e.g.sub-samples, groups where treatment assignment was haphazard, etc.)
- Unobservables may differ, but use comparisons that can handle imbalanced time-invariant unobservables

Good Observational Studies

Design features we can use to handle unobservables:

- Design comparisons so that unobservables are likely to be balanced (e.g.sub-samples, groups where treatment assignment was haphazard, etc.)
- Unobservables may differ, but use comparisons that can handle imbalanced time-invariant unobservables
- Instrumental variables, sharp discontinuities, and fuzzy discontinuities...if applied correctly

Good Observational Studies

Design features we can use to handle unobservables:

- Design comparisons so that unobservables are likely to be balanced (e.g.sub-samples, groups where treatment assignment was haphazard, etc.)
- Unobservables may differ, but use comparisons that can handle imbalanced time-invariant unobservables
- Instrumental variables, sharp discontinuities, and fuzzy discontinuities...if applied correctly
- Multiple control groups that are know to differ on unobservables

Good Observational Studies

Design features we can use to handle unobservables:

- Design comparisons so that unobservables are likely to be balanced (e.g.sub-samples, groups where treatment assignment was haphazard, etc.)
- Unobservables may differ, but use comparisons that can handle imbalanced time-invariant unobservables
- Instrumental variables, sharp discontinuities, and fuzzy discontinuities...if applied correctly
- Multiple control groups that are know to differ on unobservables
- Sensitivity analysis and bounds

Outline

1. Identification under Conditional Ignorability
2. Estimation by Subclassification
3. Matching
   - Matching in X
   - Measuring Distance
   - Balance
   - Variance Estimation
   - Matching Functions
   - Example: Blattman and Annan (2009)

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$
- Quantities of interest:

$$\begin{aligned} \text{ATE:} \quad \tau_{ATE} &\equiv \mathbb{E}[Y_i(1) - Y_i(0)] \\ \text{ATT:} \quad \tau_{ATT} &\equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \end{aligned}$$

Question: Can we identify τ_{ATE} and τ_{ATT} when D_i is not randomized?

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$
- Quantities of interest:

$$\begin{aligned} \text{ATE:} \quad \tau_{ATE} &\equiv \mathbb{E}[Y_i(1) - Y_i(0)] \\ \text{ATT:} \quad \tau_{ATT} &\equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \end{aligned}$$

  Question: Can we identify τ_{ATE} and τ_{ATT} when D_i is not randomized?

- Pre-treatment covariates: $X_i = [X_{i1}, ..., X_{iK}]^\top \in \mathcal{X}$
  - Predetermined and causally precedent with respect to D_i
  - Examples: Sex, race, age, etc.
  - X_i may be correlated with both D_i and $Y_i(d)$, thereby confounding the causal relationship
  - Excludes correlates that are potentially affected by D_i (post-treatment covariates)

Conditional Ignorability

Recall that randomized experiments work because:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$$

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

Read: Among units with same values of X_i, D_i is "as-if" randomly assigned.

Conditional Ignorability

Recall that randomized experiments work because:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$$

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

Read: Among units with same values of X_i, D_i is "as-if" randomly assigned.

Assumption: Common Support

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

Read: For any value of X_i, unit could have received treatment or control

ATE, ATT, ATC Identified under SOO

Intuition: Within strata of X, you have an experiment

ATE, ATT, ATC Identified under SOO

Intuition: Within strata of X, you have an experiment

Proof: for ATE, τ, we have

ATE, ATT, ATC Identified under SOO

Intuition: Within strata of X, you have an experiment

Proof: for ATE, τ, we have

Part 1. Identifiability of $\tau(X)$:

$$\begin{aligned}\mathbb{E}[Y_{1i} - Y_{0i}|X_i = x] &= \mathbb{E}[Y_{1i}|X_i = x, D_i = 1] - \mathbb{E}[Y_{0i}|X_i = x, D_i = 0] \\ &= \mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0] \\ &= \mathbb{E}[\hat{\tau}|X_i]\end{aligned}$$

ATE, ATT, ATC Identified under SOO

Intuition: Within strata of X, you have an experiment

Proof: for ATE, τ, we have

Part 1. Identifiability of $\tau(X)$:

$$\begin{aligned}\mathbb{E}[Y_{1i} - Y_{0i}|X_i = x] &= \mathbb{E}[Y_{1i}|X_i = x, D_i = 1] - \mathbb{E}[Y_{0i}|X_i = x, D_i = 0] \\ &= \mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0] \\ &= \mathbb{E}[\hat{\tau}|X_i]\end{aligned}$$

Part 2. Common support gets you back to τ:

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= \mathbb{E}\Big[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]\Big] \quad \text{Why?} \\ &= \int \Big(\mathbb{E}[Y_i|D_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, X_i = x]\Big)p(x)dX \\ &= \mathbb{E}\Big[\mathbb{E}[\hat{\tau}|X_i]\Big] = \mathbb{E}[\hat{\tau}]\end{aligned}$$

Identification of ATT

By the similar logic, τ_{ATT} is also identified under the conditional ignorability and common support assumptions:

$$\tau_{ATT} = \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

Identification of ATT

By the similar logic, τ_{ATT} is also identified under the conditional ignorability and common support assumptions:

$$\tau_{ATT} = \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

Proof is similar:

$$\begin{aligned}\tau_{ATT} &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, D_i = 1] \quad \text{What is outer } \mathbb{E} \text{ over?}\end{aligned}$$

Identification of ATT

By the similar logic, τ_{ATT} is also identified under the conditional ignorability and common support assumptions:

$$\tau_{ATT} \quad = \quad \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

Proof is similar:

$$\begin{aligned}
\tau_{ATT} &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, D_i = 1] \quad \text{What is outer } \mathbb{E} \text{ over?} \\
&= \int \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x, D_i = 1] p(x \mid D_i = 1) dx \\
&= \int \{\mathbb{E}[Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[Y_i \mid X_i = x, D_i = 0]\} p(x \mid D_i = 1) dx \\
&= \mathbb{E}[\hat{\tau}(x) \mid D_i = 1].
\end{aligned}$$

Is $\tau_{ATE} = \tau_{ATT}$ when CI holds?

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?
- Consider our formula for ATE, except we condition on M_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{M_i}[\mathbb{E}[Y_i \mid D_i = 1, M_i] - \mathbb{E}[Y_i \mid D_i = 0, M_i]]$$

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?
- Consider our formula for ATE, except we condition on M_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{M_i}[\mathbb{E}[Y_i \mid D_i = 1, M_i] - \mathbb{E}[Y_i \mid D_i = 0, M_i]]$$

- Because M_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$M_i = D_i M_i(1) + (1 - D_i) M_i(0)$$

Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{M_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, M_i(1)]] - \mathbb{E}_{M_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, M_i(0)]] \neq \tau_{ATE}$$

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?
- Consider our formula for ATE, except we condition on M_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{M_i}[\mathbb{E}[Y_i \mid D_i = 1, M_i] - \mathbb{E}[Y_i \mid D_i = 0, M_i]]$$

- Because M_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$M_i = D_i M_i(1) + (1 - D_i) M_i(0)$$

Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{M_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, M_i(1)]] - \mathbb{E}_{M_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, M_i(0)]] \neq \tau_{ATE}$$

- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(M_i) = f(M_i(1)) = f(M_i(0))$.
This would be true if

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?
- Consider our formula for ATE, except we condition on M_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{M_i}[\mathbb{E}[Y_i \mid D_i = 1, M_i] - \mathbb{E}[Y_i \mid D_i = 0, M_i]]$$

- Because M_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$M_i = D_i M_i(1) + (1 - D_i) M_i(0)$$

Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{M_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, M_i(1)]] - \mathbb{E}_{M_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, M_i(0)]] \neq \tau_{ATE}$$

- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(M_i) = f(M_i(1)) = f(M_i(0))$.
  This would be true if
  - D_i has no effect on M_i...

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?
- Consider our formula for ATE, except we condition on M_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{M_i}[\mathbb{E}[Y_i \mid D_i = 1, M_i] - \mathbb{E}[Y_i \mid D_i = 0, M_i]]$$

- Because M_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$M_i = D_i M_i(1) + (1 - D_i) M_i(0)$$

Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{M_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, M_i(1)]] - \mathbb{E}_{M_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, M_i(0)]] \neq \tau_{ATE}$$

- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(M_i) = f(M_i(1)) = f(M_i(0))$.
  This would be true if
  - D_i has no effect on M_i...
  - but then we would not be concerned with controlling for M_i!

Revisiting Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, M_i?
- Consider our formula for ATE, except we condition on M_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{M_i}[\mathbb{E}[Y_i \mid D_i = 1, M_i] - \mathbb{E}[Y_i \mid D_i = 0, M_i]]$$

- Because M_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$M_i = D_i M_i(1) + (1 - D_i) M_i(0)$$

Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{M_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, M_i(1)]] - \mathbb{E}_{M_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, M_i(0)]] \neq \tau_{ATE}$$

- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(M_i) = f(M_i(1)) = f(M_i(0))$.
  This would be true if
  - D_i has no effect on M_i...
  - but then we would not be concerned with controlling for M_i!
- Better to think of post-treatment outcomes of potential mediators, a topic we will come back to.

Estimation

We have now *identified* causal effects under conditional ignorability.

Two challenges remain:

1. How do we actually do the conditioning needed for estimation under SOO/CI?
2. How do we evaluate the credibility of the SOO/CI assumptions?

Estimation

We have now *identified* causal effects under conditional ignorability.

Two challenges remain:

1. How do we actually do the conditioning needed for estimation under SOO/CI?
2. How do we evaluate the credibility of the SOO/CI assumptions?

These two questions are ubiquitous in causal inference work. We will spend the next few sessions alternating between these.

Estimation

We have now *identified* causal effects under conditional ignorability.

Two challenges remain:

1. How do we actually do the conditioning needed for estimation under SOO/CI?
2. How do we evaluate the credibility of the SOO/CI assumptions?

These two questions are ubiquitous in causal inference work. We will spend the next few sessions alternating between these.

Estimation

We have now *identified* causal effects under conditional ignorability.

Two challenges remain:

1. How do we actually do the conditioning needed for estimation under SOO/CI?
2. How do we evaluate the credibility of the SOO/CI assumptions?

These two questions are ubiquitous in causal inference work. We will spend the next few sessions alternating between these.

Let's start with the estimation problem under SOO. We will explore:

- sub-classification
- matching: with and without propensity score
- re-weighting: with and without propensity score
- regression, model-based imputation

Estimation

We have now *identified* causal effects under conditional ignorability.

Two challenges remain:

1. How do we actually do the conditioning needed for estimation under SOO/CI?
2. How do we evaluate the credibility of the SOO/CI assumptions?

These two questions are ubiquitous in causal inference work. We will spend the next few sessions alternating between these.

Let's start with the estimation problem under SOO. We will explore:

- sub-classification
- matching: with and without propensity score
- re-weighting: with and without propensity score
- regression, model-based imputation

For now: the most natural, sub-classification

Outline

1. Identification under Conditional Ignorability
2. Estimation by Subclassification
3. Matching
   - Matching in X
   - Measuring Distance
   - Balance
   - Variance Estimation
   - Matching Functions
   - Example: Blattman and Annan (2009)

Identification Results for Discrete Covariates

If X_i is discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

Identification Results for Discrete Covariates

If X_i is discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

So τ_{ATE} can be calculated by:

(1) Group units into strata (or cells) defined by the values of X_i.

(2) For each stratum, calculate difference in means of Y_i between the treated and untreated.

(3) Calculate the weighted average of (2), with weights equal to the proportions of units in the strata.

Identification Results for Discrete Covariates

If X_i is discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

So τ_{ATE} can be calculated by:

(1) Group units into strata (or cells) defined by the values of X_i.

(2) For each stratum, calculate difference in means of Y_i between the treated and untreated.

(3) Calculate the weighted average of (2), with weights equal to the proportions of units in the strata.

τ_{ATT} can be calculated similarly:

(1) $\cdot$ (2) Same as (1) $\cdot$ (2) for ATE.

(3) Calculate the weighted average of (2), with weights equal to the proportions of units in the strata within the treatment group.

Subclassification Estimators

This gives us the subclassification estimators:

$$\hat{\tau}_{ATE} = \sum_{j=1}^{M} \left\{ \overline{Y}_{1j} - \overline{Y}_{0j} \right\} \frac{n_j}{n}$$

$$\hat{\tau}_{ATT} = \sum_{j=1}^{M} \left\{ \overline{Y}_{1j} - \overline{Y}_{0j} \right\} \frac{n_{1j}}{n_1}$$

where
$$\begin{cases} M = \text{\# of strata} \\ n_j = \text{\# of units in cell } j \\ n_{1j} = \text{\# of treated units in cell } j \\ \overline{Y}_{dj} = \text{mean outcome for units with } D_i = d \text{ in cell } j \end{cases}$$

Example: Smoking and Mortality (Cochran 1968)

TABLE 1

DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Example: Smoking and Mortality (Cochran 1968)

TABLE 2

MEAN AGES, YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?

$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

$15 \cdot (29/40) + 35 \cdot (9/40) + 50 \cdot (2/40) = 21.2$

Smoking and Mortality (Cochran (1968))

TABLE 3

ADJUSTED DEATH RATES USING 3 AGE GROUPS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What would the (unadjusted) difference in mean death rates for smokers versus non-smokers be?

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What would the (unadjusted) difference in mean death rates for smokers versus non-smokers be? 6

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What would the (unadjusted) difference in mean death rates for smokers versus non-smokers be? 6

What is the subclassification estimator for the ATE of smoking?

$$\hat{\tau}_{ATE} = (28 - 24) \cdot \frac{10}{20} + (22 - 16) \cdot \frac{10}{20} = 5$$

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATE of smoking on death rate?

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATE of smoking on death rate?

Not identified! (because of the lack of common support)

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATT of smoking on death rate?

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATT of smoking on death rate?

$$\begin{aligned}\hat{\tau}_{ATT} &= (28-22)\cdot\frac{3}{10}+(21-16)\cdot\frac{3}{10}+(23-17)\cdot\frac{4}{10}\\ &= 5.1\end{aligned}$$

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

- Take the age-adjustment for example. Under what assumption do we get a causal effect estimate of the type of smoking on death rate?

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

- Take the age-adjustment for example. Under what assumption do we get a causal effect estimate of the type of smoking on death rate?
- Do you believe it? What's a counter-example?

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

- Take the age-adjustment for example. Under what assumption do we get a causal effect estimate of the type of smoking on death rate?
- Do you believe it? What's a counter-example?
- Can we test whether the conditional ignorability assumption is valid?

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

- Take the age-adjustment for example. Under what assumption do we get a causal effect estimate of the type of smoking on death rate?
- Do you believe it? What's a counter-example?
- Can we test whether the conditional ignorability assumption is valid?
- We'll talk more about sensitivity analyses and other approaches to bolster this type of analysis

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

- Take the age-adjustment for example. Under what assumption do we get a causal effect estimate of the type of smoking on death rate?
- Do you believe it? What's a counter-example?
- Can we test whether the conditional ignorability assumption is valid?
- We'll talk more about sensitivity analyses and other approaches to bolster this type of analysis

Did we need any model? Did we estimate any parameters?

Food for thought

The primary difficulty with this approach is believing the assumptions.

Estimation can be tricky but comes second.

- Take the age-adjustment for example. Under what assumption do we get a causal effect estimate of the type of smoking on death rate?
- Do you believe it? What's a counter-example?
- Can we test whether the conditional ignorability assumption is valid?
- We'll talk more about sensitivity analyses and other approaches to bolster this type of analysis

Did we need any model? Did we estimate any parameters?

Why do you think sub-classification is of limited use in most real applications?

Summary

- Causal inference in observational studies often rests on this "SOO" (or CI) assumption

Summary

- Causal inference in observational studies often rests on this "SOO" (or CI) assumption
- A useful intuition: "find strata of X in which you think an experiment is occurring"

Summary

- Causal inference in observational studies often rests on this "SOO" (or CI) assumption
- A useful intuition: "find strata of X in which you think an experiment is occurring"
- Goal is to approximate a randomized experiment within subgroups

Summary

- Causal inference in observational studies often rests on this "SOO" (or CI) assumption
- A useful intuition: "find strata of X in which you think an experiment is occurring"
- Goal is to approximate a randomized experiment within subgroups
- Plausibility of your conditional ignorability: can argue that variation in treatment status within strata of X is random?

Summary

- Causal inference in observational studies often rests on this "SOO" (or CI) assumption
- A useful intuition: "find strata of X in which you think an experiment is occurring"
- Goal is to approximate a randomized experiment within subgroups
- Plausibility of your conditional ignorability: can argue that variation in treatment status within strata of X is random?
- Do not control for post-treatment covariates!
- If you have a small number of discrete covariates, ATE can be estimated nonparametrically via subclassification

Summary

- Causal inference in observational studies often rests on this "SOO" (or CI) assumption
- A useful intuition: "find strata of X in which you think an experiment is occurring"
- Goal is to approximate a randomized experiment within subgroups
- Plausibility of your conditional ignorability: can argue that variation in treatment status within strata of X is random?
- Do not control for post-treatment covariates!
- If you have a small number of discrete covariates, ATE can be estimated nonparametrically via subclassification

Next up: more estimation strategies, same assumptions. Matching, weighting, regression.

Outline

1 Identification under Conditional Ignorability

2 Estimation by Subclassification

3 Matching
- Matching in X
- Measuring Distance
- Balance
- Variance Estimation
- Matching Functions
- Example: Blattman and Annan (2009)

SOO: Identification and Estimation

Recall the SOO identification assumption:

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

SOO: Identification and Estimation

Recall the SOO identification assumption:

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

SOO: Identification and Estimation

Recall the SOO identification assumption:

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

A direct approach would be:

- estimate effects $\tau(X)$ for each stratum or level of X
- average these, weighting over some distribution, e.g. $p(X)$ or $p(X|D = 1)$.

SOO: Identification and Estimation

Recall the SOO identification assumption:

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

A direct approach would be:

- estimate effects $\tau(X)$ for each stratum or level of X
- average these, weighting over some distribution, e.g. $p(X)$ or $p(X|D = 1)$.

Some methods do this quite literally:

- sub-classification
- matching

Matching is Not an Identification Strategy

ATT Estimation by Matching

- For each treated unit i with covariates X_i, you would like to estimate $\tau_i = Y_{1i} - Y_{0i}$.

ATT Estimation by Matching

- For each treated unit i with covariates X_i, you would like to estimate $\tau_i = Y_{1i} - Y_{0i}$.
- For treated units you observe Y_{1i}, but where to get Y_{0i}?

ATT Estimation by Matching

- For each treated unit i with covariates X_i, you would like to estimate $\tau_i = Y_{1i} - Y_{0i}$.
- For treated units you observe Y_{1i}, but where to get Y_{0i}?
- Matching: borrow it from control unit with (nearly) the same X_i

ATT Estimation by Matching

- For each treated unit i with covariates X_i, you would like to estimate $\tau_i = Y_{1i} - Y_{0i}$.
- For treated units you observe Y_{1i}, but where to get Y_{0i}?
- Matching: borrow it from control unit with (nearly) the same X_i
- So estimator is:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the closest value to X_i among the untreated observations.

ATT Estimation by Matching

- For each treated unit i with covariates X_i, you would like to estimate $\tau_i = Y_{1i} - Y_{0i}$.
- For treated units you observe Y_{1i}, but where to get Y_{0i}?
- Matching: borrow it from control unit with (nearly) the same X_i
- So estimator is:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the closest value to X_i among the untreated observations.

We can also use the average of M closest matches:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)}, \right) \right\}$$

ATT Estimation by Matching

- For each treated unit i with covariates X_i, you would like to estimate $\tau_i = Y_{1i} - Y_{0i}$.
- For treated units you observe Y_{1i}, but where to get Y_{0i}?
- Matching: borrow it from control unit with (nearly) the same X_i
- So estimator is:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the closest value to X_i among the untreated observations.

We can also use the average of M closest matches:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)}, \right) \right\}$$

Does SOO assumption guarantee that this gets you the *ATT*?

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?

$\widehat{\tau}_{ATT} = 1/3 \cdot (6 - 9) + 1/3 \cdot (1 - 0) + 1/3 \cdot (0 - 9) = -3.7$

Common Distance Metrics

What do we mean by "closeness" in X when it is multi-dimensional?

Covariate vectors for i, j: $X_i = [X_i^{(1)}, X_i^{(2)}, ..., X_i^{(k)}]^\top$ and $X_j = [X_j^{(1)}, X_j^{(2)}, ..., X_j^{(k)}]^\top$

Common Distance Metrics

What do we mean by "closeness" in X when it is multi-dimensional?

Covariate vectors for i, j: $X_i = [X_i^{(1)}, X_i^{(2)}, ..., X_i^{(k)}]^\top$ and $X_j = [X_j^{(1)}, X_j^{(2)}, ..., X_j^{(k)}]^\top$

Some common options:

1. Euclidean distance:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)^\top (X_i - X_j)}$$
$$= [(X_i^{(1)} - X_j^{(1)})^2 + ... + (X_i^{(P)} - X_j^{(P)})^2]^{\frac{1}{2}}$$

Common Distance Metrics

What do we mean by "closeness" in X when it is multi-dimensional?

Covariate vectors for i, j: $X_i = [X_i^{(1)}, X_i^{(2)}, ..., X_i^{(k)}]^\top$ and $X_j = [X_j^{(1)}, X_j^{(2)}, ..., X_j^{(k)}]^\top$

Some common options:

1. Euclidean distance:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)^\top (X_i - X_j)}$$
$$= [(X_i^{(1)} - X_j^{(1)})^2 + ... + (X_i^{(P)} - X_j^{(P)})^2]^{\frac{1}{2}}$$

2. "StataD" (`nnmatch`)/default in `Match()` for `R`: rescaled Euclidean

$$StataD(X_i, X_j) = \sqrt{(X_i - X_j)^\top diag(\Sigma_X^{-1})(X_i - X_j)}$$

where Σ is the Variance-Covariance-Matrix. Invariant to rescaling of X

Common Distance Metrics

3. Mahalanobis distance:

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)^\top \Sigma^{-1} (X_i - X_j)}$$

where Σ is the Variance-Covariance-Matrix. Invariant to rescaling and rotations of X

Common Distance Metrics

3. Mahalanobis distance:

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)^\top \Sigma^{-1} (X_i - X_j)}$$

where Σ is the Variance-Covariance-Matrix. Invariant to rescaling and rotations of X

4. GeneticD (`GenMatch`):

$$GeneticD(X_i, X_j) = \sqrt{(X_i - X_j)^\top (S^{-1/2})^\top \, W \, S^{-1/2} (X_i - X_j)}$$

where W is a $(P \times P)$ positive definite weight matrix with zeros in off-diagonals, controlling "variable importance"

Mahalanobis Distance: Example

	X_1	X_2
Treated	0	0
Control A	2	2
Control B	1.8	0

$$X_T = \begin{pmatrix} 0 & 0 \end{pmatrix}^\top \quad X_A = \begin{pmatrix} 2 & 2 \end{pmatrix}^\top \quad X_B = \begin{pmatrix} 1.8 & 0 \end{pmatrix}^\top$$

$$\Sigma = \begin{pmatrix} & X^{(1)} & X^{(2)} \\ X^{(1)} & 1 & .9 \\ X^{(2)} & .9 & 1 \end{pmatrix}$$

Which control is closer?

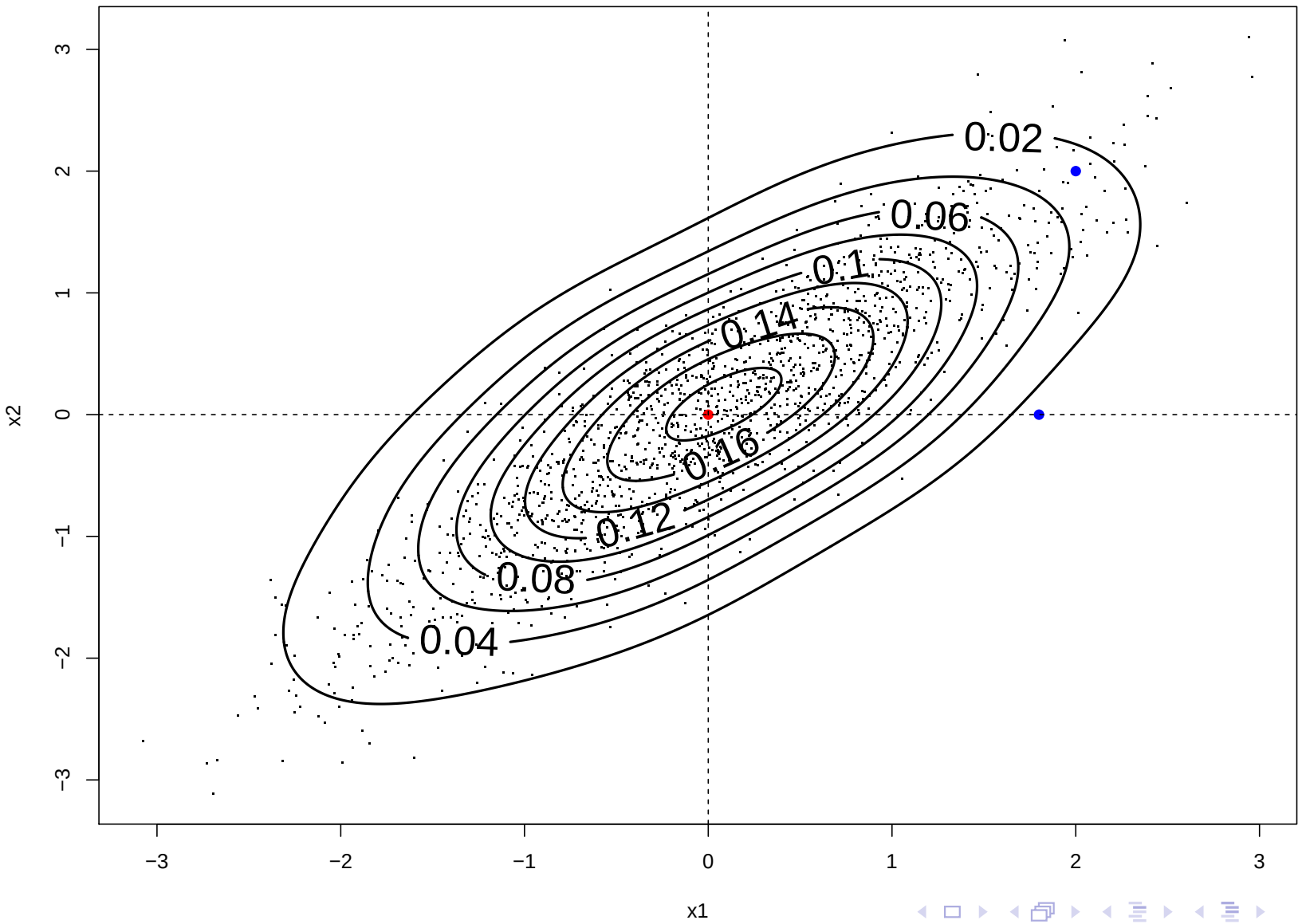
Mahalanobis Distance

$$X_T = \begin{pmatrix} 0 & 0 \end{pmatrix}^\top \quad X_A = \begin{pmatrix} 2 & 2 \end{pmatrix}^\top \quad X_B = \begin{pmatrix} 1.8 & 0 \end{pmatrix}^\top \quad \Sigma = \begin{pmatrix} & X^{(1)} & X^{(2)} \\ X^{(1)} & 1 & .9 \\ X^{(2)} & .9 & 1 \end{pmatrix}$$

$$\begin{aligned} MD(X_A, X_T) &= \sqrt{(X_A - X_T)^\top \Sigma^{-1} (X_A - X_T)} \\ &= \sqrt{[\begin{pmatrix} 2 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 0 \end{pmatrix}] \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}^{-1} [\begin{pmatrix} 2 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 0 \end{pmatrix}]^\top} \\ &= \sqrt{\begin{bmatrix} 2 & 2 \end{bmatrix} \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} \begin{bmatrix} 2 & 2 \end{bmatrix}^\top} \\ &= 4.2 \\ MD(X_B, X_T) &= \sqrt{\begin{bmatrix} 1.8 & 0 \end{bmatrix} \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} \begin{bmatrix} 1.8 & 0 \end{bmatrix}^\top} \\ &= 17 \end{aligned}$$

With $StataD(X_A, X_T) = \sqrt{(X_i - X_T)^\top diag(\Sigma_X^{-1})(X_i - X_T)}$ we find $StataD(X_A, X_T) = 84$ and $StataD(X_B, X_T) = 17$ since correlation is ignored.

Mahalanobis Distance



Local Methods and the Curse of Dimensionality

Big Problem: the volume increases exponentially when adding extra dimensions

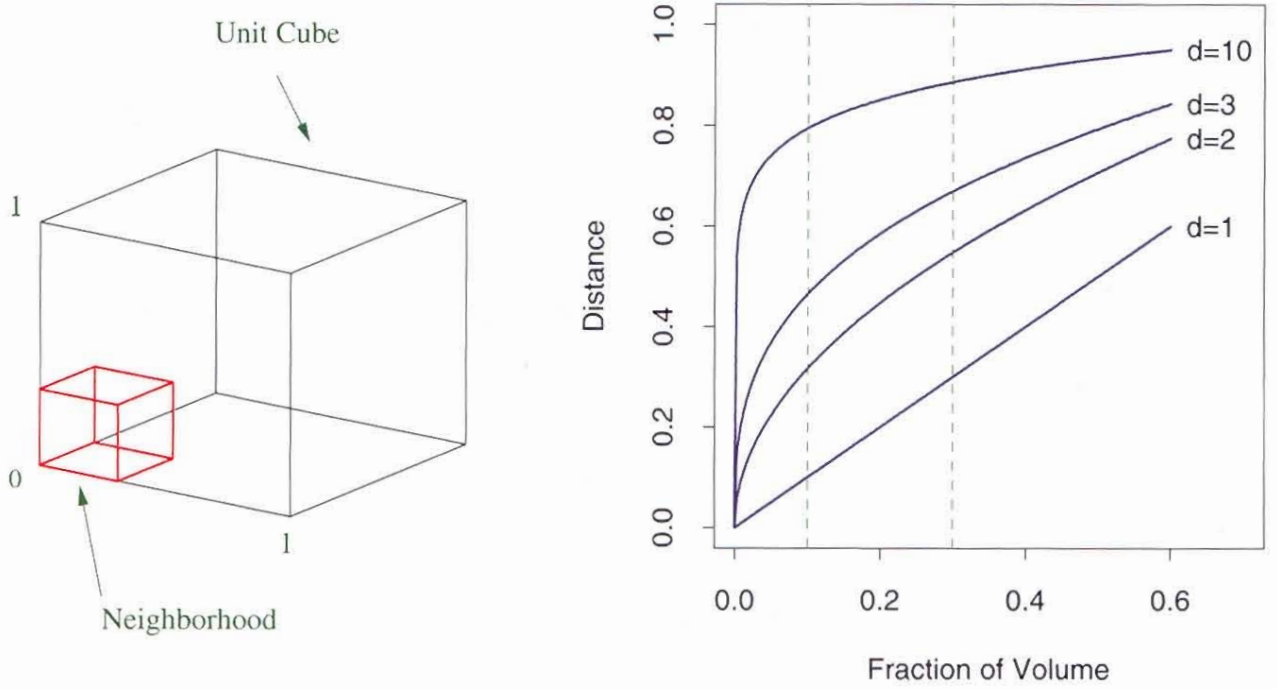


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Matching with Bias Correction

Weakness of matching: discrepancy=$||X_i - X_{j(1)}||_d > 0$

Matching with Bias Correction

Weakness of matching: discrepancy=$||X_i - X_{j(1)}||_d > 0$

With multiple continuous variables, this becomes problematic

- if X predicts treatment, then this discrepancy is not just random noise

Matching with Bias Correction

Weakness of matching: discrepancy=$||X_i - X_{j(1)}||_d > 0$

With multiple continuous variables, this becomes problematic

- if X predicts treatment, then this discrepancy is not just random noise
- does not average away quickly: not $\sqrt{N}$ consistent (Abadie and Imbens, 2005)

Matching with Bias Correction

Weakness of matching: discrepancy=$||X_i - X_{j(1)}||_d > 0$

With multiple continuous variables, this becomes problematic

- if X predicts treatment, then this discrepancy is not just random noise
- does not average away quickly: not $\sqrt{N}$ consistent (Abadie and Imbens, 2005)

Compromise: adjust for discrepancy:

- On average, Y_{0j} differs from Y_{0i} by $\mathbb{E}[Y_0|X_i] - \mathbb{E}[Y_0|X_{j(i)}]$
- Consider population regression function $\mu_0(X) = \mathbb{E}[Y_0|X]$
- Let's estimate $\hat{\mu}_0$ by OLS regression of Y on X among controls

$$\hat{\mu}_0(X) = \beta_0 + \beta_1 X$$

Matching with Bias Correction

So subtract off estimated discrepancy on Y_0 from each pair (Abadie & Imbens, 2005)

$$\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})),$$

- these "bias-corrected" matching estimators are an improvement even if $\hat{\mu}_0$ is misspecified
- the large sample distribution of this estimator (for the case of matching with replacement) is roughly normal.

In R: `Match(Y,Tr, X,BiasAdjust = TRUE)`

Bias Adjustment with Matched Data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	8

What is $\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$?

Bias Adjustment with Matched Data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	1	1	10
4		0	0	2
5		9	0	3
6		1	0	8

What is $\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$?

Estimate $\widehat{\mu}_0(x) = \beta_0 + \beta_1 x = 5 - .4x$.

Bias Adjustment with Matched Data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	1	1	10
4		0	0	2
5		9	0	3
6		1	0	8

What is $\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$?

Estimate $\widehat{\mu}_0(x) = \beta_0 + \beta_1 x = 5 - .4x$. Now plug in:

$$\begin{aligned} \widehat{\tau}_{ATT} &= 1/3\{((6-9) - (\widehat{\mu}_0(3) - \widehat{\mu}_0(3))) \\ &+ ((1-0) - (\widehat{\mu}_0(1) - \widehat{\mu}_0(2))) \\ &+ ((0-1) - (\widehat{\mu}_0(10) - \widehat{\mu}_0(8)))\} \\ &= -0.86 \end{aligned}$$

(Unadjusted: $1/3((6-9) + (1-0) + (0-1)) = -1$)

Choices when Matching

- With or Without Replacement?

Choices when Matching

- With or Without Replacement?
- How many matches?

Choices when Matching

- With or Without Replacement?
- How many matches?
- Which Matching Algorithm?
  - genetic matching, kernel Matching, full matching
  - coarsened exact matching
  - matching as pre-processing
  - propensity score matching

Choices when Matching

- With or Without Replacement?
- How many matches?
- Which Matching Algorithm?
  - genetic matching, kernel Matching, full matching
  - coarsened exact matching
  - matching as pre-processing
  - propensity score matching
- Good rule: use whatever gives you the best balance! Checking balance is important to get a sense for how much extrapolation is needed
  - should check balance on interactions and higher moments
- With insufficient overlap, all adjustment methods are problematic because we have rely on a model to impute missing potential outcomes.

Checking balance: Theory

- Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample

Checking balance: Theory

- Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
- In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)

Checking balance: Theory

- Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
- In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)
- Balance tests are often used (e.g. t-test, F-test, KS test)

Checking balance: Theory

- Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
- In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)
- Balance tests are often used (e.g. t-test, F-test, KS test)
- Note that balance tests can be misleading in a matching context: you can make everything insignificant by simply dropping lots of observations — make sure this is not happening (see Hartman & Hidalgo)

Checking balance: Theory

- Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
- In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)
- Balance tests are often used (e.g. t-test, F-test, KS test)
- Note that balance tests can be misleading in a matching context: you can make everything insignificant by simply dropping lots of observations — make sure this is not happening (see Hartman & Hidalgo)

Workflow:

- Estimate $\rightarrow$ Check Balance $\rightarrow$ Re-estimate $\rightarrow$ Check Balance $\rightarrow \cdots$ (ad infinitum until you get a good balance)
- Is this data snooping? No, because inference remains blind to Y

Kolmogorov-Smirnov (KS) Test

- The KS test is used to test whether two random variables are sampled from the same distribution

Kolmogorov-Smirnov (KS) Test

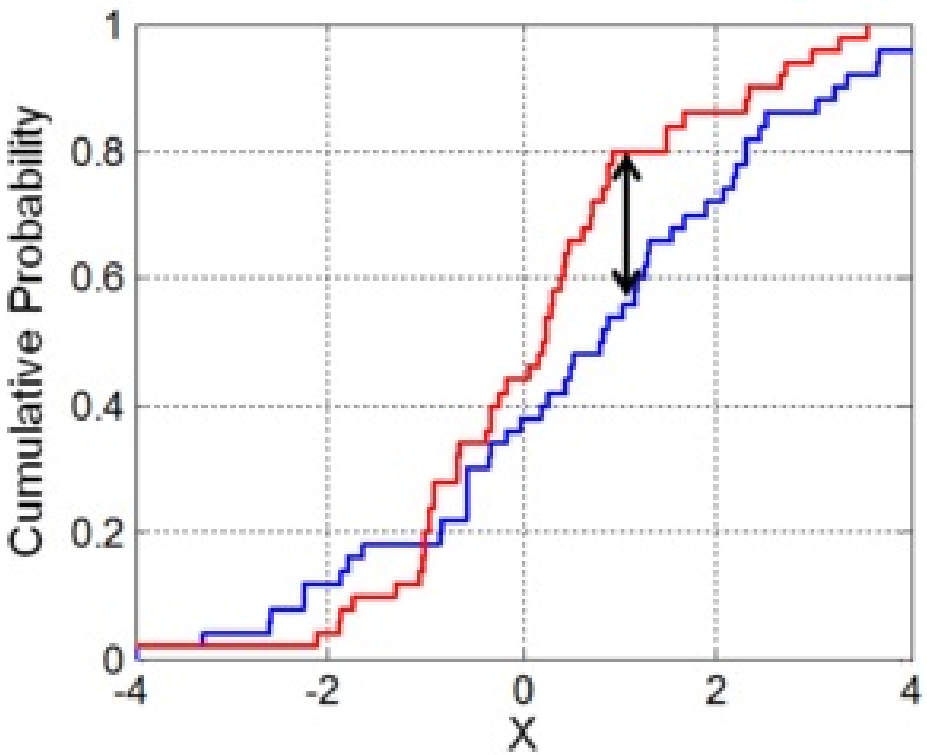
- The KS test is used to test whether two random variables are sampled from the same distribution
- The test is nonparametric, meaning that it works (asymptotically) without assumptions about the form of the underlying distribution

Kolmogorov-Smirnov (KS) Test

- The KS test is used to test whether two random variables are sampled from the same distribution
- The test is nonparametric, meaning that it works (asymptotically) without assumptions about the form of the underlying distribution
- Consider N observations of two random variables, X_0 and X_1.

Kolmogorov-Smirnov (KS) Test

- The KS test is used to test whether two random variables are sampled from the same distribution
- The test is nonparametric, meaning that it works (asymptotically) without assumptions about the form of the underlying distribution
- Consider N observations of two random variables, X_0 and X_1.



Checking balance: Theory

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|$$

Checking balance: Theory

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|$$

where $\widehat{F}_0(x)$, $\widehat{F}_1(x)$ is the empirical CDF of X_0, X_1.

Checking balance: Theory

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|$$

where $\widehat{F}_0(x)$, $\widehat{F}_1(x)$ is the empirical CDF of X_0, X_1.

The KS null hypothesis: $F_1(x) = F_0(x)$ (no difference in true distributions)

Checking balance: Theory

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|$$

where $\widehat{F}_0(x)$, $\widehat{F}_1(x)$ is the empirical CDF of X_0, X_1.

The KS null hypothesis: $F_1(x) = F_0(x)$ (no difference in true distributions)

Under the null, D has the Kolmogorov distribution as $n \to \infty$.

Checking balance: Theory

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|$$

where $\widehat{F}_0(x)$, $\widehat{F}_1(x)$ is the empirical CDF of X_0, X_1.

The KS null hypothesis: $F_1(x) = F_0(x)$ (no difference in true distributions)

Under the null, D has the Kolmogorov distribution as $n \to \infty$.

Reject the null at level α if

$$\frac{D}{\sqrt{(n_1 + n_0)/n_1 n_0}} > c_\alpha$$

level (α)	.1	.05	.01
critical value (c_α)	1.22	1.36	1.63

Balance Checks

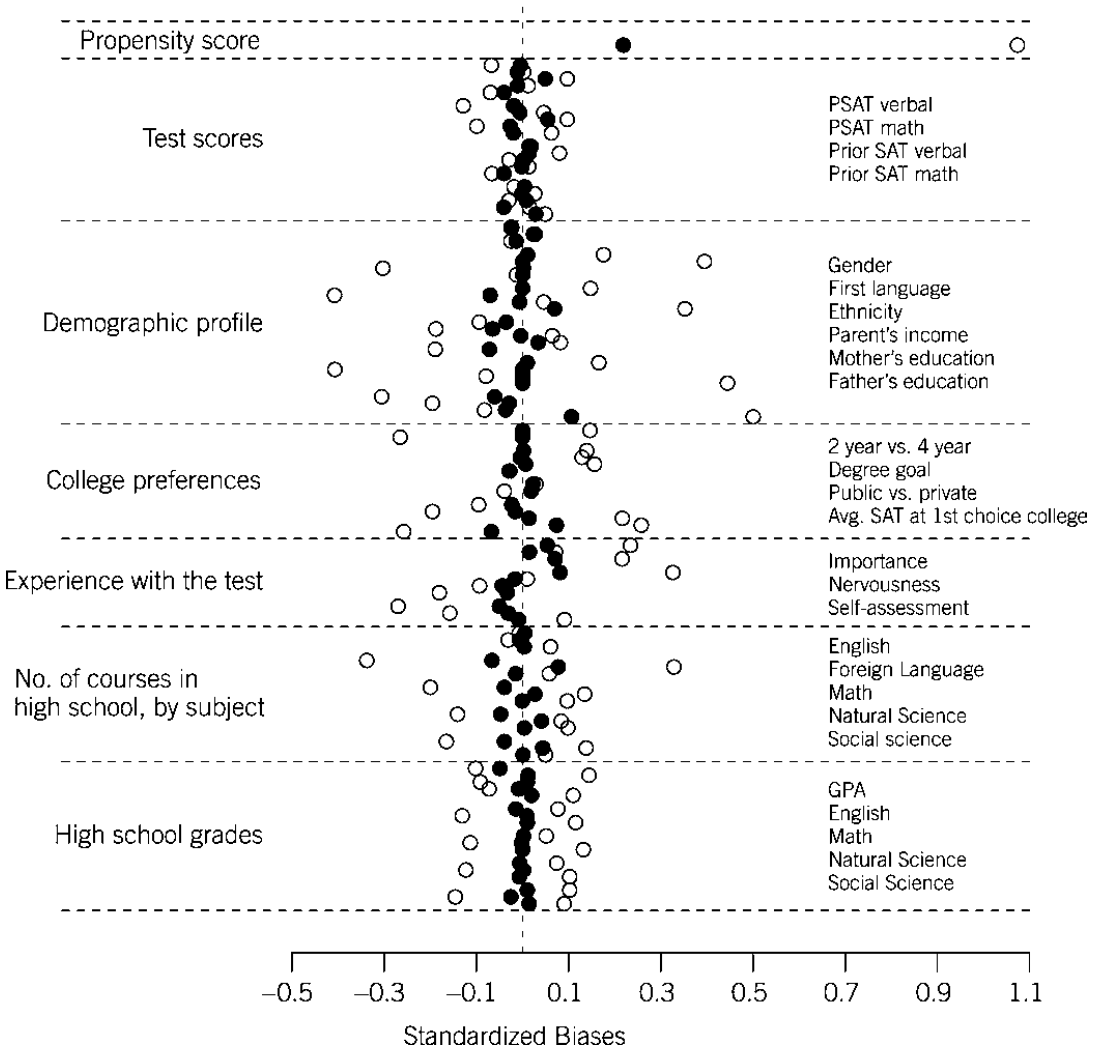


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Balance Checks

TABLE 2. Balance Summary Statistics and Tests: Russian and Chechen Sweeps

Pretreatment Covariates	Mean Treated	Mean Control	Mean Difference	Std. Bias	Rank Sum Test	K-S Test
Demographics						
Population	8.657	8.606	0.049	0.033	0.708	0.454
Tariqa	0.076	0.048	0.028	0.104	0.331	—
Poverty	1.917	1.931	−0.016	−0.024	0.792	1.000
Spatial						
Elevation	5.078	5.233	−0.155	−0.135	0.140	0.228
Isolation	1.007	1.070	−0.063	−0.096	0.343	0.851
Groznyy	0.131	0.138	−0.007	−0.018	0.864	—
War Dynamics						
TAC	0.241	0.282	−0.041	−0.095	0.424	—
Garrison	0.379	0.414	−0.035	−0.072	0.549	—
Rebel	0.510	0.441	0.070	0.139	0.240	—
Selection						
Presweep violence	3.083	3.117	−0.034	0.009	0.454	0.292
Large-scale theft	0.034	0.055	−0.021	−0.115	0.395	—
Killing	0.117	0.090	0.027	0.084	0.443	—
Violence Inflicted						
Total abuse	0.970	0.833	0.137	0.124	0.131	0.454
Prior sweeps	1.729	1.812	−0.090	−0.089	0.394	0.367
Other						
Month	7.428	6.986	0.442	0.130	0.260	0.292
Year	2004.159	2004.110	0.049	0.043	0.889	1.000

Note: 145 matched pairs. Matching with replacement.

Variance of the Matching Estimator

Option 1. Ignore the matching uncertainty and estimate the SEs from whatever model you run on the matched sample

- treats the matched sample as fixed
- thus ignores uncertainty in matching

Variance of the Matching Estimator

Option 1. Ignore the matching uncertainty and estimate the SEs from whatever model you run on the matched sample

- treats the matched sample as fixed
- thus ignores uncertainty in matching

Option 2. Bootstrap

- NO. Not consistent
- "extreme non-smoothness" (Abadie & Imbens 2006)

Variance of the Matching Estimator

Option 1. Ignore the matching uncertainty and estimate the SEs from whatever model you run on the matched sample

- treats the matched sample as fixed
- thus ignores uncertainty in matching

Option 2. Bootstrap

- NO. Not consistent
- "extreme non-smoothness" (Abadie & Imbens 2006)

Option 3. Abadie-Imbens asymptotic SEs

- uses matched pairs to estimate local variance in Y_0 and takes weighted sum that accounts for matching
- provided by `Match()` package.
- generally use these
- but still an open area of research

Useful Matching Functions

The workhorse model is the `Match()` function in the `Matching` package:

```
Match(Y = NULL, Tr, X, Z = X, V = rep(1, length(Y)),
    estimand = "ATT", M = 1, BiasAdjust = FALSE, exact = NULL,
    caliper = NULL,  replace = TRUE, ties = TRUE,
    CommonSupport = FALSE, Weight = 1, Weight.matrix = NULL,
    weights = NULL, Var.calc = 0,  sample = FALSE, restrict = NULL
    match.out = NULL, distance.tolerance = 1e-05,
    tolerance = sqrt(.Machine$double.eps), version = "standard")
```

Default distance metric (`Weight=1`) is normalized Euclidean distance

- `MatchBalance(formu)` for balance checking
- `GenMatch()` for genetic matching

Example: Blattman and Annan (2009)

The Consequences of Child Soldiering. The Review of Economics and Statistics.

Example: Blattman and Annan (2009)

The Consequences of Child Soldiering. The Review of Economics and Statistics.

Example: Blattman and Annan (2009)

The Consequences of Child Soldiering. The Review of Economics and Statistics.

Question: what is the impact of abduction by the rebel group *Lord's Resistance Army* on education?

Example: Blattman and Annan (2009)

The Consequences of Child Soldiering. The Review of Economics and Statistics.

Question: what is the impact of abduction by the rebel group *Lord's Resistance Army* on education?

ID Strategy: SOO. According to the logic by which abduction occurred, we will assume that abduction is indiscriminate, conditional on age and location.

> *Abduction was large-scale and seemingly indiscriminate; 60,000 to 80,000 youth are estimated to have been abducted and more than a quarter of males currently aged 14 to 30 in our study region were abducted for at least two weeks...*
>
> *Youth were typically taken by roving groups of 10 to 20 rebels during night raids on rural homes. Adolescent males appear to have been the most pliable, reliable and effective forced recruits, and so were disproportionately targeted by the LRA. Youth under age 11 and over 24 tended to be avoided and had a high probability of immediate release.*

Example: Blattman and Annan (2009)

Data: panel survey of male youth in war-afflicted regions of Uganda.

- `abd`: abducted by the LRA (the treatment)
- `c_ach - c_pal`: Location indicators
- `age`: age in years
- `fthr_ed`,`mthr_ed` : father's/mother's education (years)
- `orphan96`: indicator if parent's died before 1997
- `hh_fthr_frm`: indicator if father is a farmer
- `hh_size96`: household size in 1996
- `educ`: years of education
- `distress`: index of emotional distress (0-15)
- `logwage`: log of average daily wage earned in last 4 weeks

Example: Blattman and Annan (2009)

Data: panel survey of male youth in war-afflicted regions of Uganda.

- `abd`: abducted by the LRA (the treatment)
- `c_ach - c_pal`: Location indicators
- `age`: age in years
- `fthr_ed`,`mthr_ed` : father's/mother's education (years)
- `orphan96`: indicator if parent's died before 1997
- `hh_fthr_frm`: indicator if father is a farmer
- `hh_size96`: household size in 1996
- `educ`: years of education
- `distress`: index of emotional distress (0-15)
- `logwage`: log of average daily wage earned in last 4 weeks

Example: Blattman and Annan (2009)

Data: panel survey of male youth in war-afflicted regions of Uganda.

- `abd`: abducted by the LRA (the treatment)
- `c_ach - c_pal`: Location indicators
- `age`: age in years
- `fthr_ed`,`mthr_ed` : father's/mother's education (years)
- `orphan96`: indicator if parent's died before 1997
- `hh_fthr_frm`: indicator if father is a farmer
- `hh_size96`: household size in 1996
- `educ`: years of education
- `distress`: index of emotional distress (0-15)
- `logwage`: log of average daily wage earned in last 4 weeks

NB: `educ`, `distress`, and `logwage` are measured after abduction.

Example: Blattman and Annan (2009)

Data: panel survey of male youth in war-afflicted regions of Uganda.

- `abd`: abducted by the LRA (the treatment)
- `c_ach - c_pal`: Location indicators
- `age`: age in years
- `fthr_ed`,`mthr_ed` : father's/mother's education (years)
- `orphan96`: indicator if parent's died before 1997
- `hh_fthr_frm`: indicator if father is a farmer
- `hh_size96`: household size in 1996
- `educ`: years of education
- `distress`: index of emotional distress (0-15)
- `logwage`: log of average daily wage earned in last 4 weeks

NB: `educ`, `distress`, and `logwage` are measured after abduction.

Now, to `R`.

Roadmap

1. Theory: Potential outcomes, identification, key quantities
2. Randomization
   - difference in means, variance estimation
   - covariate adjustment
   - blocking
   - cluster randomization
3. Selection on Observables
   - sub-classification
   - matching (on X)
   - weighting (on X)
   - matching and weighting on $Pr(D = 1)$ (propensity scores)
   - regression
4. Instrumental Variables
5. Regression Discontinuity
6. Difference in Differences and Synthetic Control
7. Sensitivity and Bounds

We have to keep saying this...

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

We have to keep saying this...

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

We have to keep saying this...

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

- We took the literal approach with sub-classiciation and matching

We have to keep saying this...

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

- We took the literal approach with sub-classiciation and matching
- Now, let's think about re-weighting in ways that make the overall distribution of treated and control more similar

We have to keep saying this...

Assumption: Conditional Ignorability

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

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variables)

How to actually estimate something that conditions on X?

- We took the literal approach with sub-classiciation and matching
- Now, let's think about re-weighting in ways that make the overall distribution of treated and control more similar
- This will allow us to estimate things as if we have done so conditionally on X then averaged over some density in X

Why Weighting?

A few ways to come up with the weighting idea...

Why Weighting?

A few ways to come up with the weighting idea...

First, we have been getting conditional treatment effect estimates, $\hat{\mathbb{E}}[Y_{1i} - Y_{0i}|X]$, then putting these together by some distribution of X.

Why Weighting?

A few ways to come up with the weighting idea...

First, we have been getting conditional treatment effect estimates, $\hat{\mathbb{E}}[Y_{1i} - Y_{0i}|X]$, then putting these together by some distribution of X.

What if we moved the treated and control to have the same distribution in X first, then take a simple average? More formally,

Why Weighting?

A few ways to come up with the weighting idea...

First, we have been getting conditional treatment effect estimates, $\hat{\mathbb{E}}[Y_{1i} - Y_{0i}|X]$, then putting these together by some distribution of X.

What if we moved the treated and control to have the same distribution in X first, then take a simple average? More formally,

- We have conditional treatment effects in each stratum of X:

$$\mathbb{E}[\tau|X] = \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$$

Why Weighting?

A few ways to come up with the weighting idea...

First, we have been getting conditional treatment effect estimates, $\hat{\mathbb{E}}[Y_{1i} - Y_{0i}|X]$, then putting these together by some distribution of X.

What if we moved the treated and control to have the same distribution in X first, then take a simple average? More formally,

- We have conditional treatment effects in each stratum of X:

$$\mathbb{E}[\tau|X] = \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$$

- Then we integrate them using some $p(X^\star)$:

$$\begin{aligned}\mathbb{E}_X\mathbb{E}[\tau|X] &= \int \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]p(X^\star)dx \\ &= \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X^\star)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X^\star)dx \\ &= \text{mean for treated minus controls when they have same density in } X\end{aligned}$$

Why Weighting?

A few ways to come up with the weighting idea...

First, we have been getting conditional treatment effect estimates, $\hat{\mathbb{E}}[Y_{1i} - Y_{0i}|X]$, then putting these together by some distribution of X.

What if we moved the treated and control to have the same distribution in X first, then take a simple average? More formally,

- We have conditional treatment effects in each stratum of X:

$$\mathbb{E}[\tau|X] = \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$$

- Then we integrate them using some $p(X^\star)$:

$$\begin{aligned}\mathbb{E}_X\mathbb{E}[\tau|X] &= \int \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]p(X^\star)dx \\ &= \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X^\star)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X^\star)dx \\ &= \text{mean for treated minus controls when they have same density in } X\end{aligned}$$

Suggests "moving" treated and control to same density in X, then diff. in means

What Weights?

Without weighting, we are estimating:

$$\mathbb{E}[\hat{\tau}] = \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X_i|D = 1)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X_i|D_i = 0)dx$$

What Weights?

Without weighting, we are estimating:

$$\mathbb{E}[\hat{\tau}] = \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X_i|D = 1)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X_i|D_i = 0)dx$$

Consider weighting by $\frac{p(D_i)}{p(D_i|X_i)}$:

What Weights?

Without weighting, we are estimating:

$$\mathbb{E}[\hat{\tau}] = \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X_i|D = 1)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X_i|D_i = 0)dx$$

Consider weighting by $\frac{p(D_i)}{p(D_i|X_i)}$:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_{IPW}] &= \int \mathbb{E}[Y_i|D_i = 1, X_i]\frac{p(X|D = 1)p(D_i)}{p(D_i|X_i)}dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]\frac{p(X|D_i = 0)p(D_i)}{p(D_i|X_i)}d \\ &= \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X_i)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X_i)dx\end{aligned}$$

What Weights?

Without weighting, we are estimating:

$$\mathbb{E}[\hat{\tau}] = \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X_i|D = 1)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X_i|D_i = 0)dx$$

Consider weighting by $\frac{p(D_i)}{p(D_i|X_i)}$:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_{IPW}] &= \int \mathbb{E}[Y_i|D_i = 1, X_i]\frac{p(X|D = 1)p(D_i)}{p(D_i|X_i)}dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]\frac{p(X|D_i = 0)p(D_i)}{p(D_i|X_i)}d \\ &= \int \mathbb{E}[Y_i|D_i = 1, X_i]p(X_i)dx - \int \mathbb{E}[Y_i|D_i = 0, X_i]p(X_i)dx\end{aligned}$$

How convenient! These are the (stabilized) Inverse Propensity Score (IPW) weights.

- these are great in theory: not matching discrepancy to worry about, $dim(X)$ doesn't matter
- but require a model to get $p(D_i|X_i)$, so vulnerable to misspecification
- This is one view of propensity scores, $p(D_i = 1|X_i)$. More on that soon.

Weighting without a Model

Other approaches to weighting don't require a model of the treatment assignment

Weighting without a Model

Other approaches to weighting don't require a model of the treatment assignment

Rather, the logic comes back to balance:

- recall, when we matched we wanted to make sure we got balance
- equivalently, we know we want to move treated and control units to have similar distributions in X as a way of conditioning

Weighting without a Model

Other approaches to weighting don't require a model of the treatment assignment

Rather, the logic comes back to balance:

- recall, when we matched we wanted to make sure we got balance
- equivalently, we know we want to move treated and control units to have similar distributions in X as a way of conditioning

So let's choose weights that "get us balance":

- typically we settle for equal means in X
- though we can try to get equal means on X^2, $X_1 X_2$, etc.
- it turns out that these get you unbiased ATT estimates when $Y_i 0$ is linear in the things you get mean balance on

Weighting without a Model

Other approaches to weighting don't require a model of the treatment assignment

Rather, the logic comes back to balance:

- recall, when we matched we wanted to make sure we got balance
- equivalently, we know we want to move treated and control units to have similar distributions in X as a way of conditioning

So let's choose weights that "get us balance":

- typically we settle for equal means in X
- though we can try to get equal means on X^2, $X_1 X_2$, etc.
- it turns out that these get you unbiased ATT estimates when $Y_i 0$ is linear in the things you get mean balance on

Entropy balancing (Hainmueller, 2011) is one option (`ebal` in `R`), as well as generalized method of moments (GMM), empirical likelihood (EL)

Weighting for (mean) balance

Entropy balancing (and similar) try to find weights on control units s.t. the mean of X (and possible higher-order transforms) for controls matches that of treated.

Weighting for (mean) balance

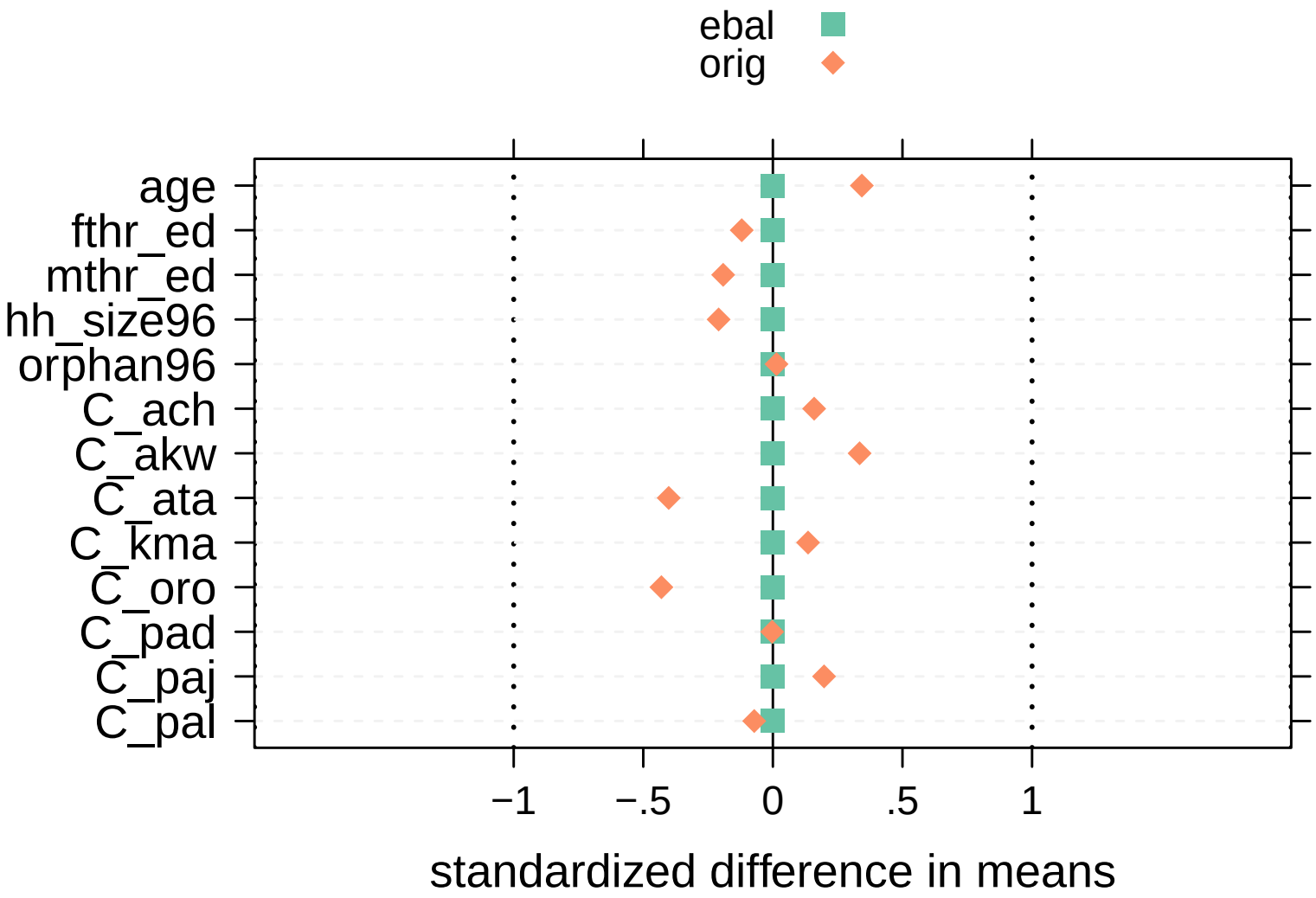
Entropy balancing (and similar) try to find weights on control units s.t. the mean of X (and possible higher-order transforms) for controls matches that of treated.

Resulting weights can be use to get weighted difference in means, or applied to any model (e.g. regression)

- Could also move treated to controls, or move both to common means
- Fast, often finds perfect mean balance, no iteration
- Can include higher order terms if you like
- Main downside: only getting balance on means or other included functions
- Standard errors are still an open question, but bootstrap has not been ruled out.

Now let's try it in `R`...

Balance on Blattman data with Ebal



Back to Propensity Scores

We earlier thought about reweighting the data to get both treated and control distributions onto common $p(X)$.

Back to Propensity Scores

We earlier thought about reweighting the data to get both treated and control distributions onto common $p(X)$.

One useful set of weights was the inverse propensity score weights.

Back to Propensity Scores

We earlier thought about reweighting the data to get both treated and control distributions onto common $p(X)$.

One useful set of weights was the inverse propensity score weights.

More commonly, we arrive at this idea through the propensity score:

Propensity Score

$$\text{Propensity Score} = \pi_i = p(D_i = 1|X_i)$$

Propensity Scores: Standard Motivation

- It can be hard to get close matches on X_i if X_i is multi-dimensional
- What if you could just match on a one-dimensional summary?

Conditioning on the Propensity Score

Under SOO and common support,

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$$

Propensity Scores: Standard Motivation

- It can be hard to get close matches on X_i if X_i is multi-dimensional
- What if you could just match on a one-dimensional summary?

Conditioning on the Propensity Score

Under SOO and common support,

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$$

The intuition:

- consider units with equal probabilities of getting the treatment as far as X can predict ($\pi(X_i)$ constant)
- recall SOO: conditionally on X, treatment is random
- now, conditionally on $\pi(X_i)$, $Pr(D = 1)$ does not depend on X.
- thus conditioning on $\pi(X)$ gets us random assignment (if SOO is true and $\pi(X)$ correct)
- See next slide for proof along these lines

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1 | Y_0, Y_1, \pi(X)) = \Pr(D = 1 | \pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\Pr(D = 1 | Y_1, Y_0, \pi(X)) \quad = \quad \mathbb{E}[D | Y_1, Y_0, \pi(X)]$$

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1 | Y_0, Y_1, \pi(X)) = \Pr(D = 1 | \pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned} \Pr(D = 1 | Y_1, Y_0, \pi(X)) &= \mathbb{E}[D | Y_1, Y_0, \pi(X)] \\ &= \mathbb{E}\left[\mathbb{E}[D | Y_1, Y_0, X] | Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \end{aligned}$$

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= \mathbb{E}[D|Y_1, Y_0, \pi(X)] \\
&= \mathbb{E}\left[\mathbb{E}[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= \mathbb{E}\left[\mathbb{E}[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)}
\end{aligned}$$

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1 | Y_0, Y_1, \pi(X)) = \Pr(D = 1 | \pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$
\begin{aligned}
\Pr(D = 1 | Y_1, Y_0, \pi(X)) &= \mathbb{E}[D | Y_1, Y_0, \pi(X)] \\
&= \mathbb{E}\left[\mathbb{E}[D | Y_1, Y_0, X] | Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= \mathbb{E}\left[\mathbb{E}[D | X] | Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= \mathbb{E}\left[\pi(X) | Y_1, Y_0, \pi(X)\right]
\end{aligned}
$$

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= \mathbb{E}[D|Y_1, Y_0, \pi(X)] \\
&= \mathbb{E}\left[\mathbb{E}[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= \mathbb{E}\left[\mathbb{E}[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= \mathbb{E}\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\Pr(D = 1|\pi(X)) = \mathbb{E}[D|\pi(X)] = \mathbb{E}[\mathbb{E}[D|X]|\pi(x)]$$

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= \mathbb{E}[D|Y_1, Y_0, \pi(X)] \\
&= \mathbb{E}\left[\mathbb{E}[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= \mathbb{E}\left[\mathbb{E}[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= \mathbb{E}\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= \mathbb{E}[D|\pi(X)] = \mathbb{E}[\mathbb{E}[D|X]|\pi(x)] \\
&= \mathbb{E}[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Using Propensity Scores

Once estimated, you can match or weight on $\hat{\pi}_i$.

- Matching:
  - may need a *caliper* to trim cases without common support
  - or, some trim cases with extreme pscores by hand to improve common support
  - but beware, this changes your estimand
- Weighting
  - Simple IPW weights: $\frac{1}{p(D_i|X_i)} = \frac{1}{D_i\pi_i+(1-D_i)(1-\pi_i)}$
  - Stabilized IPW weights: $\frac{p(D_i)}{p(D_i|X_i)} = \frac{(D_i)Pr(D_i=1)+(1-D_i)(1-Pr(D_i0))}{(D_i)\pi_i+(1-D_i)(1-\pi_i)}$
  - But if using weights, check how extreme they get and how few units are doing most of the work

Using Propensity Scores

Pros:

- Just one thing to match on
- Ignores imbalance on X's that do not predict D...very helpful for large P
- Will give you balance ("balancing property"), but *only in expectation* and if correctly estimated

Using Propensity Scores

Pros:

- Just one thing to match on
- Ignores imbalance on X's that do not predict D...very helpful for large P
- Will give you balance ("balancing property"), but *only in expectation* and if correctly estimated

Cons:

- Sometimes balancing property seen as the *goal* of pscores – by that metric it often under-performs.
- requires correct estimation of the propensity score!
- ...requires iterating between pscore estimation and balance checking
- extreme weights can cause large bias (so check)

Using Propensity Scores

Pros:

- Just one thing to match on
- Ignores imbalance on X's that do not predict D...very helpful for large P
- Will give you balance ("balancing property"), but *only in expectation* and if correctly estimated

Cons:

- Sometimes balancing property seen as the *goal* of pscores – by that metric it often under-performs.
- requires correct estimation of the propensity score!
- ...requires iterating between pscore estimation and balance checking
- extreme weights can cause large bias (so check)

Both at once:

Covariate balancing propensity scores (Imai, Ratkovic 2014)

Using Propensity Scores: Blattman & Annan

```
pi.out = glm(abd~age+fthr_ed+mthr_ed+hh_size96+orphan96+C_ach+C_akw+C_ata+
C_kma+C_oro+C_pad+C_paj+C_pal,data=dat,family="binomial"(link=logit))

matchout.pi=Match(Y=dat$educ,Tr=dat$abd,X=pi.out$fit,M=1,estimand="ATT")
summary(matchout.pi)

#Check balance
mb.out.pi =MatchBalance(match.out=matchout.pi,abd~age+fthr_ed+mthr_ed+
  hh_size96+orphan96+C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal,data=

btest_after_pi=baltest.collect(mb.out.pi,var.names=varnames,after=TRUE)

balancecompare.pi=cbind(round(btest[,c("mean.Tr","mean.Co","T pval","KS pv
                    round(btest_after_pi[,c("mean.Tr","mean.Co","T pval",
```

Using Propensity Scores: Blattman & Annan

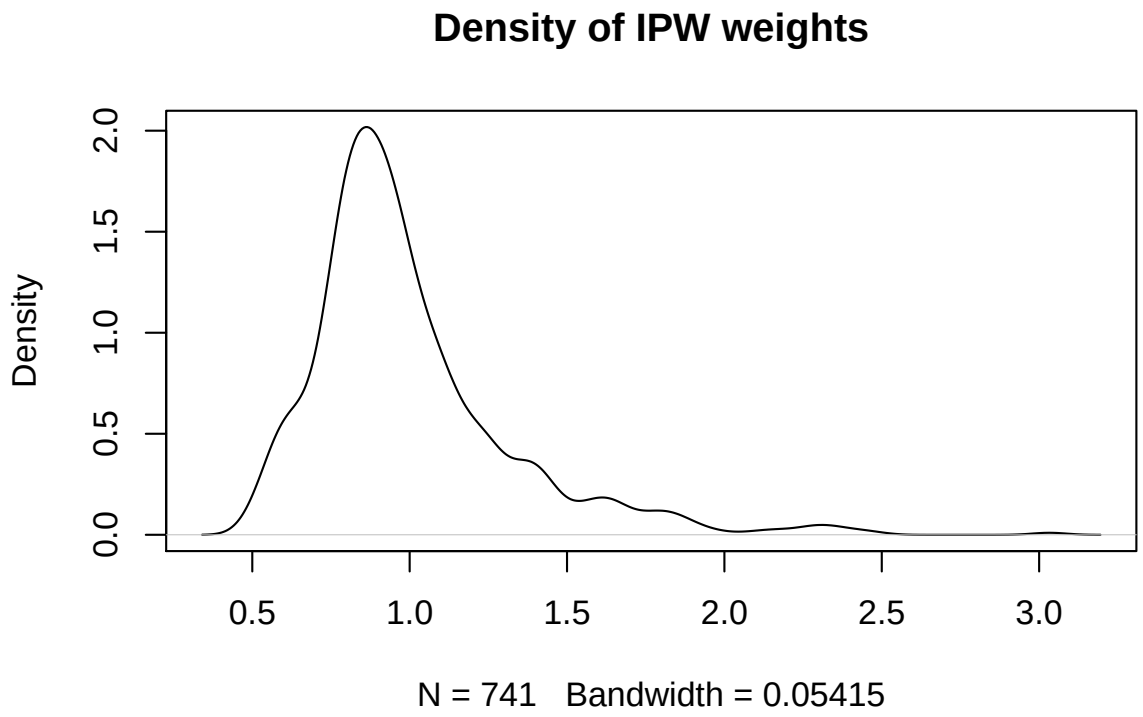
balancecompare.pi

	mean.Tr	mean.Co	T pval	KS pval	mean.Tr	mean.Co	T pval	KS pval
age	21.37	20.15	0.00	0.00	21.37	21.04	0.27	0.17
fthr_ed	5.76	6.07	0.27	0.87	5.76	5.60	0.48	0.09
mthr_ed	2.09	2.49	0.09	0.35	2.09	2.40	0.11	0.20
hh_size96	8.09	8.70	0.06	0.04	8.09	8.02	0.79	0.03
orphan96	0.08	0.08	0.90	NA	0.08	0.11	0.07	NA
C_ach	0.15	0.11	0.13	NA	0.15	0.17	0.60	NA
C_akw	0.16	0.08	0.00	NA	0.16	0.17	0.62	NA
C_ata	0.10	0.20	0.00	NA	0.10	0.09	0.55	NA
C_kma	0.15	0.12	0.19	NA	0.15	0.16	0.87	NA
C_oro	0.05	0.14	0.00	NA	0.05	0.05	0.95	NA
C_pad	0.12	0.12	0.98	NA	0.12	0.09	0.10	NA
C_paj	0.15	0.10	0.06	NA	0.15	0.15	0.91	NA
C_pal	0.11	0.13	0.51	NA	0.11	0.13	0.45	NA

Using Propensity Scores: Blattman & Annan

Weighting on the propensity scores (stabilized IPW):

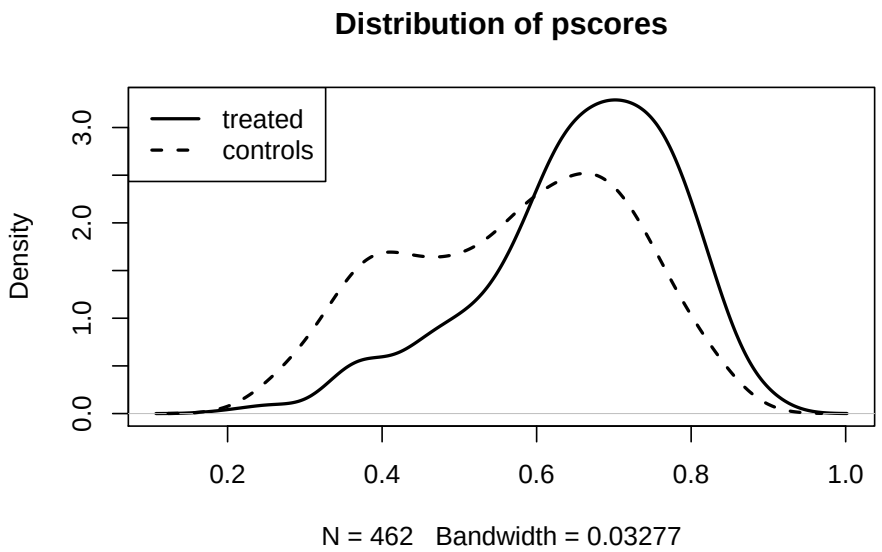
```
ps=pi.out$fit
D=dat$abd
PrD=mean(D)
IPW= (D*PrD+(1-D)*(1-PrD))/(D*ps+(1-D)*(1-ps))

#Good to check how crazy the weights are:
plot(density(IPW))
```

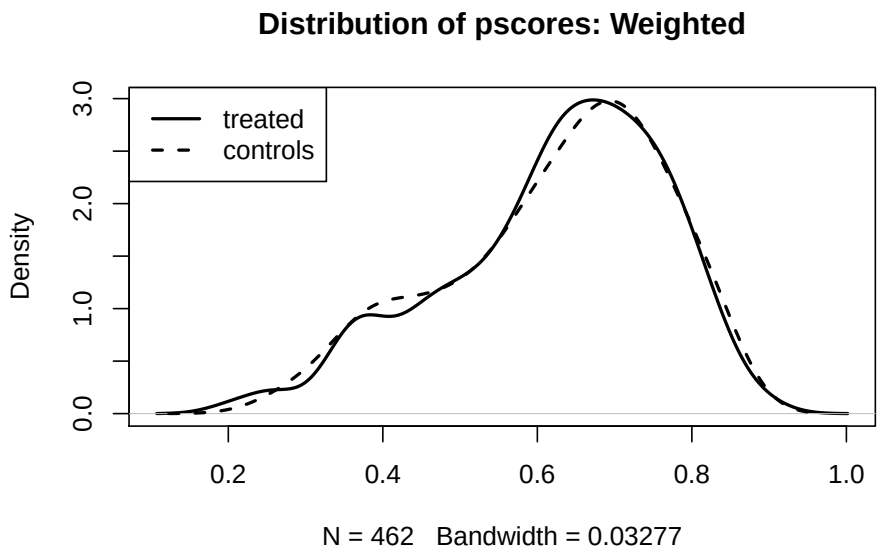


How does $p(\pi_i)$ look before/after weighting?

How does $p(\pi_i)$ look before/after weighting?

```
plot(density(pi.out$fit[D==1], weight=IPW[D==1]/sum(IPW[D==1])), lwd=2, ma
lines(density(pi.out$fit[D==0], weight=IPW[D==0]/sum(IPW[D==0])),lwd=2, lt
legend("topleft", legend=c("treated","controls"), lty=c(1,2), lwd=2)
```





Check Omnibus Balance

Check Omnibus Balance

```
omnibus.bal = lm(abd~age+fthr_ed+mthr_ed+hh_size96+orphan96+
                   C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal, data=dat)
summary(omnibus.bal)
#R2=0.08, F-statistic: 5.585 on 12 and 728 DF,  p-value: 3.303e-09
```

Check Omnibus Balance

```
omnibus.bal = lm(abd~age+fthr_ed+mthr_ed+hh_size96+orphan96+
                   C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal, data=dat)
summary(omnibus.bal)
#R2=0.08, F-statistic: 5.585 on 12 and 728 DF,  p-value: 3.303e-09

omnibus.bal.ipw = lm(abd~age+fthr_ed+mthr_ed+hh_size96+orphan96+
                   C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal,
                   weight=IPW, data=dat)
summary(omnibus.bal.ipw)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.138e-01  1.058e-01   5.804 9.69e-09 ***
age          5.474e-04  3.691e-03   0.148    0.882
fthr_ed     -4.053e-04  5.352e-03  -0.076    0.940
mthr_ed     -2.116e-04  6.628e-03  -0.032    0.975
hh_size96    4.875e-04  4.689e-03   0.104    0.917
orphan96     3.979e-03  6.873e-02   0.058    0.954
...
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.4888 on 728 degrees of freedom
Multiple R-squared:  0.0002747,Adjusted R-squared:  -0.0162
F-statistic: 0.01667 on 12 and 728 DF,  p-value: 1
```

Check Omnibus Balance

```
omnibus.bal = lm(abd~age+fthr_ed+mthr_ed+hh_size96+orphan96+
                   C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal, data=dat)
summary(omnibus.bal)
#R2=0.08, F-statistic: 5.585 on 12 and 728 DF,  p-value: 3.303e-09

omnibus.bal.ipw = lm(abd~age+fthr_ed+mthr_ed+hh_size96+orphan96+
                   C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal,
                   weight=IPW, data=dat)
summary(omnibus.bal.ipw)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.138e-01  1.058e-01   5.804 9.69e-09 ***
age          5.474e-04  3.691e-03   0.148    0.882
fthr_ed     -4.053e-04  5.352e-03  -0.076    0.940
mthr_ed     -2.116e-04  6.628e-03  -0.032    0.975
hh_size96    4.875e-04  4.689e-03   0.104    0.917
orphan96     3.979e-03  6.873e-02   0.058    0.954
...
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.4888 on 728 degrees of freedom
Multiple R-squared:  0.0002747,Adjusted R-squared:  -0.0162
F-statistic: 0.01667 on 12 and 728 DF,  p-value: 1
```

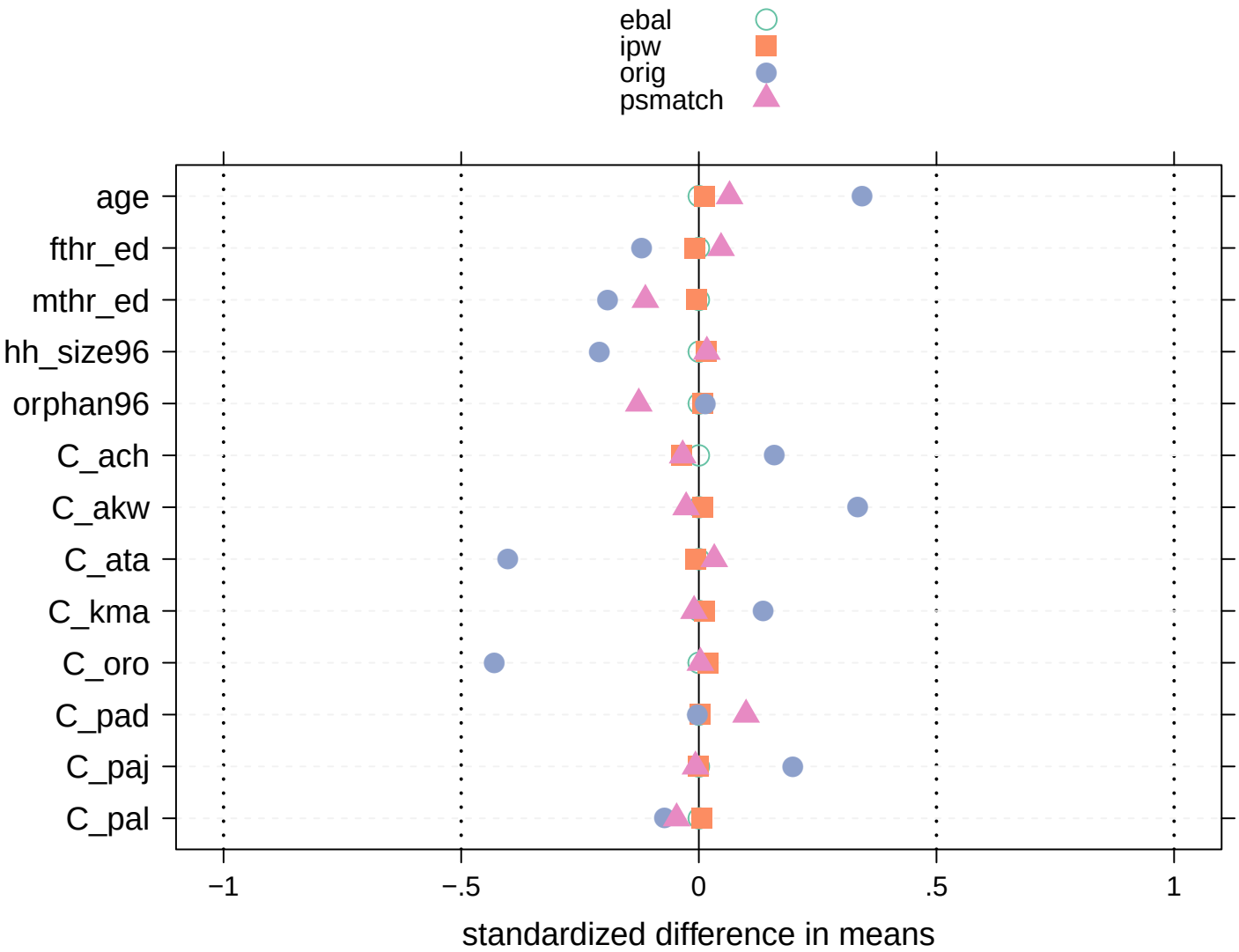
Yay!

Estimating the ATE with IPW weights

```
lm.out.naive = lm(educ~abd, data=dat)
lm.out.ipw = lm(educ~abd, weight=IPW, data=dat)
lm.out.ipw2=lm(educ~abd+age+fthr_ed+mthr_ed+hh_size96+orphan96+
    C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C_pal,weight=IPW,data=dat)
```

	Dependent variable:		
	educ		
	(1)	(2)	(3)
abd	−0.595***	−0.726***	−0.735***
	(0.218)	(0.220)	(0.210)
Constant	7.416***	7.499***	6.113***
	(0.172)	(0.174)	(0.610)
Observations	741	741	741
R^2	0.010	0.014	0.128
Adjusted R^2	0.009	0.013	0.112
Residual Std. Error (df = 739)	2.876	2.904	2.754
Note:			*p<0.1; **p<0.05; ***p<0.01

Balance Plot after IPW



Model-based Estimation of Causal Effects

Regression has been popular to “control for” variables. Is this causal inference?

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Which under CI we can identify:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Which under CI we can identify:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

And us to compute things like:

$$\tau_{ATE} = \mathbb{E}[\hat{\tau}(x)] \quad \text{and} \quad \tau_{ATT} = \mathbb{E}[\hat{\tau}(x) \mid D_i = 1]$$

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Which under CI we can identify:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

And us to compute things like:

$$\tau_{ATE} = \mathbb{E}[\hat{\tau}(x)] \quad \text{and} \quad \tau_{ATT} = \mathbb{E}[\hat{\tau}(x) \mid D_i = 1]$$

This suggests a model-based approach for estimating causal effects:

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Which under CI we can identify:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

And us to compute things like:

$$\tau_{ATE} = \mathbb{E}[\hat{\tau}(x)] \quad \text{and} \quad \tau_{ATT} = \mathbb{E}[\hat{\tau}(x) \mid D_i = 1]$$

This suggests a model-based approach for estimating causal effects:

- Suppose $\mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_i \mid D_i = 0, X_i] = \beta_0 + \gamma^\top X_i$

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Which under CI we can identify:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

And us to compute things like:

$$\tau_{ATE} = \mathbb{E}[\hat{\tau}(x)] \quad \text{and} \quad \tau_{ATT} = \mathbb{E}[\hat{\tau}(x) \mid D_i = 1]$$

This suggests a model-based approach for estimating causal effects:

- Suppose $\mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_i \mid D_i = 0, X_i] = \beta_0 + \gamma^\top X_i$
- ...and $\mathbb{E}[Y_{1i}|D_i = 1, X_i] = \mathbb{E}[Y_i \mid D_i = 1, X_i] = \beta_0 + \beta_1 + \gamma^\top X_i$

Model-based Estimation of Causal Effects

Regression has been popular to "control for" variables. Is this causal inference?

Take the quantity:

$$\hat{\tau}(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x]$$

Which under CI we can identify:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

And us to compute things like:

$$\tau_{ATE} = \mathbb{E}[\hat{\tau}(x)] \quad \text{and} \quad \tau_{ATT} = \mathbb{E}[\hat{\tau}(x) \mid D_i = 1]$$

This suggests a model-based approach for estimating causal effects:

- Suppose $\mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_i \mid D_i = 0, X_i] = \beta_0 + \gamma^\top X_i$
- ...and $\mathbb{E}[Y_{1i}|D_i = 1, X_i] = \mathbb{E}[Y_i \mid D_i = 1, X_i] = \beta_0 + \beta_1 + \gamma^\top X_i$
- Which gives $\hat{\tau}(X) = \beta_1$. This implies the single model:

$$\mathbb{E}[Y_i \mid D_i, X_i] = \beta_0 + \beta_1 D_i + \gamma^\top X_i,$$

OLS as an Estimator of Causal Effects

When is $\hat{\beta}_{OLS}$ a good estimator of τ_{ATE}?

OLS as an Estimator of Causal Effects

When is $\hat{\beta}_{OLS}$ a good estimator of τ_{ATE}?

We slipped in two assumptions:

OLS as an Estimator of Causal Effects

When is $\hat{\beta}_{OLS}$ a good estimator of τ_{ATE}?

We slipped in two assumptions:

(1) Constant treatment effect. We assumed

- $\hat{\tau}(X_i) = \mathbb{E}[Y_{1i} - Y_{0i}|X_i]$
- ...which implies $\tau_i = \tau$ for all i

OLS as an Estimator of Causal Effects

When is $\hat{\beta}_{OLS}$ a good estimator of τ_{ATE}?

We slipped in two assumptions:

(1) Constant treatment effect. We assumed

- $\hat{\tau}(X_i) = \mathbb{E}[Y_{1i} - Y_{0i}|X_i]$
- ...which implies $\tau_i = \tau$ for all i

(2) Linearity: Between our model and the CI assumption, we asserted

$$\mathbb{E}[Y_{id}|X_i] = \beta_0 + \beta_1 d_i + \gamma^\top X_i \quad \text{for} \quad d = 0, 1$$

Equivalently,

$$Y_{id} = \beta_0 + \beta_1 d_i + \gamma^\top X_i + \varepsilon \quad \text{for} \quad d = 0, 1$$

OLS as an Estimator of Causal Effects

When is $\hat{\beta}_{OLS}$ a good estimator of τ_{ATE}?

We slipped in two assumptions:

(1) Constant treatment effect. We assumed

- $\hat{\tau}(X_i) = \mathbb{E}[Y_{1i} - Y_{0i}|X_i]$
- ...which implies $\tau_i = \tau$ for all i

(2) Linearity: Between our model and the CI assumption, we asserted

$$\mathbb{E}[Y_{id}|X_i] = \beta_0 + \beta_1 d_i + \gamma^\top X_i \quad \text{for} \quad d = 0, 1$$

Equivalently,

$$Y_{id} = \beta_0 + \beta_1 d_i + \gamma^\top X_i + \varepsilon \quad \text{for} \quad d = 0, 1$$

Noting that (2) implies (1) (such that $\beta_1 = \tau$), there are 3 possible scenarios:

1. Both (1) and (2) are true.
2. Only (1) is true.
3. Neither (1) nor (2) is true.

Case 1: Constant Effect w/ Linear Potential Outcomes

Result: If treatment effect is constant across units and potential outcomes are linear in X_i, then the OLS estimate of β_1 in the following regression model

$$Y_i = \beta_0 + \beta_1 D_i + \gamma^\top X_i + \varepsilon_i$$

is an unbiased and consistent estimator of τ_{ATE}.

Case 1: Constant Effect w/ Linear Potential Outcomes

Result: If treatment effect is constant across units and potential outcomes are linear in X_i, then the OLS estimate of β_1 in the following regression model

$$Y_i = \beta_0 + \beta_1 D_i + \gamma^\top X_i + \varepsilon_i$$

is an unbiased and consistent estimator of τ_{ATE}.

Proof (just how we got here):

$$\begin{aligned}
\mathbb{E}[\beta_1] &= \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i] \quad \text{(correct specification)} \\
&= \mathbb{E}[Y_{1i}|X_i] - \mathbb{E}[Y_{0i}|X_i] \quad \text{(SOO)} \\
&= \tau(X) \\
\mathbb{E}[\beta_1] &= \tau \quad \text{(constant effect assumption)}
\end{aligned}$$

Case 1: Constant Effect w/ Linear Potential Outcomes

Result: If treatment effect is constant across units and potential outcomes are linear in X_i, then the OLS estimate of β_1 in the following regression model

$$Y_i = \beta_0 + \beta_1 D_i + \gamma^\top X_i + \varepsilon_i$$

is an unbiased and consistent estimator of τ_{ATE}.

Proof (just how we got here):

$$\begin{aligned}
\mathbb{E}[\beta_1] &= \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i] \quad \text{(correct specification)} \\
&= \mathbb{E}[Y_{1i}|X_i] - \mathbb{E}[Y_{0i}|X_i] \quad \text{(SOO)} \\
&= \tau(X) \\
\mathbb{E}[\beta_1] &= \tau \quad \text{(constant effect assumption)}
\end{aligned}$$

Note that if CI and linearity hold, ε cannot be related to D: traditional CIA assumption

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

What if $\mathbb{E}[Y_i(d)|X_i]$ is an unknown, nonlinear function of X and D but we use OLS?

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

What if $\mathbb{E}[Y_i(d)|X_i]$ is an unknown, nonlinear function of X and D but we use OLS?

Recall that OLS is still the best linear predictor in terms of MSE:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i)^2\right]$$

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

What if $\mathbb{E}[Y_i(d)|X_i]$ is an unknown, nonlinear function of X and D but we use OLS?

Recall that OLS is still the best linear predictor in terms of MSE:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i)^2\right]$$

Implies $\hat{\beta}_{OLS}$ is best linear approximation to the population regression function:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[\left(\mathbb{E}[Y_i \mid D_i, X_i] - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i\right)^2\right]$$

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

What if $\mathbb{E}[Y_i(d)|X_i]$ is an unknown, nonlinear function of X and D but we use OLS?

Recall that OLS is still the best linear predictor in terms of MSE:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i)^2\right]$$

Implies $\hat{\beta}_{OLS}$ is best linear approximation to the population regression function:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[\left(\mathbb{E}[Y_i \mid D_i, X_i] - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i\right)^2\right]$$

- $\therefore$ we say "$\hat{\beta}_{OLS}$ is best linear approximation to the true treatment effect", whatever the true functional form is.

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

What if $\mathbb{E}[Y_i(d)|X_i]$ is an unknown, nonlinear function of X and D but we use OLS?

Recall that OLS is still the best linear predictor in terms of MSE:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i)^2\right]$$

Implies $\hat{\beta}_{OLS}$ is best linear approximation to the population regression function:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[\left(\mathbb{E}[Y_i \mid D_i, X_i] - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i\right)^2\right]$$

- $\therefore$ we say "$\hat{\beta}_{OLS}$ is best linear approximation to the true treatment effect", whatever the true functional form is.
- This approximation may or may not be good! Danger of mis-specification bias.

Case 2: Constant Effect w/ Unknown Functional Form

Linearity is a very strong assumption in most practical situations.

What if $\mathbb{E}[Y_i(d)|X_i]$ is an unknown, nonlinear function of X and D but we use OLS?

Recall that OLS is still the best linear predictor in terms of MSE:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i)^2\right]$$

Implies $\hat{\beta}_{OLS}$ is best linear approximation to the population regression function:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\text{argmin}}\ \mathbb{E}\left[\left(\mathbb{E}[Y_i \mid D_i, X_i] - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\gamma}^\top X_i\right)^2\right]$$

- $\therefore$ we say "$\hat{\beta}_{OLS}$ is best linear approximation to the true treatment effect", whatever the true functional form is.
- This approximation may or may not be good! Danger of mis-specification bias.
- More flexible models (nonlinear, semi-/non-parametric, etc.) are a good idea

Case 3: Heterogeneous Treatment Effects

We typically cannot assume constant treatment effects. Is this a problem?

Case 3: Heterogeneous Treatment Effects

We typically cannot assume constant treatment effects. Is this a problem?

Recall the subclassification estimator for the ATE:

$$\hat{\tau}_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

Case 3: Heterogeneous Treatment Effects

We typically cannot assume constant treatment effects. Is this a problem?

Recall the subclassification estimator for the ATE:

$$\hat{\tau}_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

Similarly, the subclassification estimator for the ATT:

$$\hat{\tau}_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

Case 3: Heterogeneous Treatment Effects

We typically cannot assume constant treatment effects. Is this a problem?

Recall the subclassification estimator for the ATE:

$$\hat{\tau}_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

Similarly, the subclassification estimator for the ATT:

$$\hat{\tau}_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

Result: The OLS estimator can be written as a subclassification estimator, where the weights come from conditional variances of D_i in each subgroup:

Case 3: Heterogeneous Treatment Effects

We typically cannot assume constant treatment effects. Is this a problem?

Recall the subclassification estimator for the ATE:

$$\hat{\tau}_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

Similarly, the subclassification estimator for the ATT:

$$\hat{\tau}_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

Result: The OLS estimator can be written as a subclassification estimator, where the weights come from conditional variances of D_i in each subgroup:

$$\hat{\beta}_{OLS} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \frac{\mathrm{Var}(D_i \mid X_i = x) \Pr(X_i = x)}{\sum_{x'} \mathrm{Var}(D_i \mid X_i = x') \Pr(X_i = x')}$$

- Intuition: OLS is minimizing MSE and learns more from strata of X where π_X closer to 0.50.
- think about what this means for your estimate, especially with lack of overlap
- (See Angrist and Pishke for derivation)

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\mathrm{Var}(D_i \mid X_i = x)\Pr(X_i = x)}{\sum_{x'} \mathrm{Var}(D_i \mid X_i = x')\Pr(X_i = x')}$	"$\tau_{\text{CVW-ATE}}$"

- With heterogeneous treatment effects, OLS provides an unbiased estimator for a *conditional-variance-weighted average treatment effect*.

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\text{Var}(D_i \mid X_i = x)\Pr(X_i = x)}{\sum_{x'} \text{Var}(D_i \mid X_i = x')\Pr(X_i = x')}$	"$\tau_{\text{CVW-ATE}}$"

- With heterogeneous treatment effects, OLS provides an unbiased estimator for a *conditional-variance-weighted average treatment effect*.
- Estimand has a causal meaning, but difficult to interpret. Not the ATE or ATT.

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\mathrm{Var}(D_i \mid X_i = x)\Pr(X_i = x)}{\sum_{x'} \mathrm{Var}(D_i \mid X_i = x')\Pr(X_i = x')}$	"$\tau_{\text{CVW-ATE}}$"

- With heterogeneous treatment effects, OLS provides an unbiased estimator for a *conditional-variance-weighted average treatment effect*.
- Estimand has a causal meaning, but difficult to interpret. Not the ATE or ATT.
- Note: $\mathrm{Var}(D_i \mid X_i = x) = \pi(x)(1 - \pi(x))$. Thus,
  - what happens when π_i is 0 or 1?

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\mathrm{Var}(D_i \mid X_i = x)\Pr(X_i = x)}{\sum_{x'} \mathrm{Var}(D_i \mid X_i = x')\Pr(X_i = x')}$	"$\tau_{\text{CVW-ATE}}$"

- With heterogeneous treatment effects, OLS provides an unbiased estimator for a *conditional-variance-weighted average treatment effect*.
- Estimand has a causal meaning, but difficult to interpret. Not the ATE or ATT.
- Note: $\mathrm{Var}(D_i \mid X_i = x) = \pi(x)(1 - \pi(x))$. Thus,
  - what happens when π_i is 0 or 1?
  - largest weights for strata π_i close to 0.5,

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\mathrm{Var}(D_i \mid X_i = x)\Pr(X_i = x)}{\sum_{x'} \mathrm{Var}(D_i \mid X_i = x')\Pr(X_i = x')}$	"$\tau_{\text{CVW-ATE}}$"

- With heterogeneous treatment effects, OLS provides an unbiased estimator for a *conditional-variance-weighted average treatment effect*.
- Estimand has a causal meaning, but difficult to interpret. Not the ATE or ATT.
- Note: $\mathrm{Var}(D_i \mid X_i = x) = \pi(x)(1 - \pi(x))$. Thus,
  - what happens when π_i is 0 or 1?
  - largest weights for strata π_i close to 0.5,
- This result assumes discrete Xs, but intuition holds for continuous Xs.

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\text{Var}(D_i \mid X_i = x)\Pr(X_i = x)}{\sum_{x'} \text{Var}(D_i \mid X_i = x')\Pr(X_i = x')}$	"$\tau_{\text{CVW-ATE}}$"

- With heterogeneous treatment effects, OLS provides an unbiased estimator for a *conditional-variance-weighted average treatment effect*.
- Estimand has a causal meaning, but difficult to interpret. Not the ATE or ATT.
- Note: $\text{Var}(D_i \mid X_i = x) = \pi(x)(1 - \pi(x))$. Thus,
  - what happens when π_i is 0 or 1?
  - largest weights for strata π_i close to 0.5,
- This result assumes discrete Xs, but intuition holds for continuous Xs.
- Another option: use regression to impute missing potential outcomes. E.g. Estimate $\mathbb{E}[Y_{0i}|X_i]$ and use as counterfactuals for observed Y_{1i}. Can weight each pairwise difference as you like.

Combining Regression, Matching and Weighting

The different logics of matching, weighting, and regression can be combined to improve finite-sample performance or robustness:

Combining Regression, Matching and Weighting

The different logics of matching, weighting, and regression can be combined to improve finite-sample performance or robustness:

- Bias-corrected matching (Abadie and Imbens 2005):
  - estimate bias due to unequal Y_{0i} of matched units due to imperfect match on X's.
  - subtract it off from the matching estimate for correction

Combining Regression, Matching and Weighting

The different logics of matching, weighting, and regression can be combined to improve finite-sample performance or robustness:

- Bias-corrected matching (Abadie and Imbens 2005):
  - estimate bias due to unequal Y_{0i} of matched units due to imperfect match on X's.
  - subtract it off from the matching estimate for correction
- "Doubly-robust" estimation
  - Use a weighted average of regression and IPW estimators (Robins and Rotnitzky 2001):
  - Use IPW weighting in a regression that re-includes covariates

Combining Regression, Matching and Weighting

The different logics of matching, weighting, and regression can be combined to improve finite-sample performance or robustness:

- Bias-corrected matching (Abadie and Imbens 2005):
  - estimate bias due to unequal Y_{0i} of matched units due to imperfect match on X's.
  - subtract it off from the matching estimate for correction
- "Doubly-robust" estimation
  - Use a weighted average of regression and IPW estimators (Robins and Rotnitzky 2001):
  - Use IPW weighting in a regression that re-includes covariates
- Matching as nonparametric data preprocessing (Ho et al. 2007):
  - model-based estimation of causal effect is most likely to go wrong when it involves extrapolation due to poor overlap in covariates
  - use matching to make treatment and control groups similar
  - then run regression models to estimate causal effects

Summary: Estimation under Conditional Ignorability

- Matching, weighting and regression are main methods to estimate average causal effects when one can assume conditional ignorability
- These are estimation strategies; the validity of the identification strategy (SOO) remains a first-order concern
- No single method is dominant
- Key considerations
  - does it get you good balance?
  - is there risk of extrapolation due to non-overlap?
  - is there risk you are not doing the conditioning you mean to do? Ask yourself:
  - "what assumptions of the estimation procedure might be invalid?" (e.g. close matches, common support/overlap linearity, constant treatment effects...)