Statistical Methods of Causal Inference
Lecture 2: RCTs and Matching Analysis

Dr. Sanchayan Banerjee
(largely adapted from Dr. David Hendry)

Institute for Environmental Studies
Vrije Universiteit Amsterdam

July 18, 2023

S.Banerjee@vu.nl

Outline

Outline

Course Roadmap:

1. 17 July: Potential Outcomes Framework
2. 18 July: RCTs and Matching $\rightarrow$ We are here!
3. 19 July: Panel Data Models
4. 20 July: Difference in Differences
5. 21 July: Instrumental Variable Regression
6. 24 July: Regression Discontinuity Design
7. 25 July: Power Analysis and Advanced Experimental Design
8. 26 July: Practical Issues in Experiments
9. 27 July: Exam Review (afternoon)
10. 28 July: Final Matters (exam in am; Causal inference socials in pm)

Mantra: People who look comparable are not really comparable. They differ in ways we have not observed.

Outline

Observational Studies

- Experiments form gold standard for causal inference, because randomization balances confounding characteristics between treated and non-treated

Observational Studies

- Experiments form gold standard for causal inference, because randomization balances confounding characteristics between treated and non-treated
- Cannot always randomize (e.g., effect of smoking)

Observational Studies

- Experiments form gold standard for causal inference, because randomization balances confounding characteristics between treated and non-treated
- Cannot always randomize (e.g., effect of smoking)
- Main problem in observational studies is selection bias

Observational Studies

- Experiments form gold standard for causal inference, because randomization balances confounding characteristics between treated and non-treated
- Cannot always randomize (e.g., effect of smoking)
- Main problem in observational studies is selection bias
- Goal is to design observational study to approximate an experiment

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

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

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

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

Subclassification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use subclassification:

- for each country, divide each group into different age subgroups
- calculate death rates within age subgroups
- average within age subgroup death rates using fixed weights (e.g., number of cigarette smokers)

Subclassification: Example

	Death Rates for 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

Question

What is the average death rate for Pipe Smokers?

Subclassification: Example

	Death Rates for 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

Question

What is the average death rate for Pipe Smokers?

Answer

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

Subclassification: Example

	Death Rates for 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

Question

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

Subclassification: Example

	Death Rates for 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

Question

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

Answer

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

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

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

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

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

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

Selection On Observables: The Familiar Case

Subclassification, matching, and propensity-score weighting, when used to make causal inferences, are all examples of causal identification via "selection on observables"

Selection On Observables: The Familiar Case

Subclassification, matching, and propensity-score weighting, when used to make causal inferences, are all examples of causal identification via "selection on observables"

- Intuitively: If we are able to account for all of the X variables that affect both selection into treatment, D, AND variation in the outcome, Y, then we can claim independence between D and potential outcomes, which we need to talk about causality

Selection On Observables: The Familiar Case

Subclassification, matching, and propensity-score weighting, when used to make causal inferences, are all examples of causal identification via "selection on observables"

- Intuitively: If we are able to account for all of the X variables that affect both selection into treatment, D, AND variation in the outcome, Y, then we can claim independence between D and potential outcomes, which we need to talk about causality
- The big problem: We have to be able to observe and measure all of the X variables that affect selection into treatment and variation in outcomes

Selection On Observables: The Familiar Case

Subclassification, matching, and propensity-score weighting, when used to make causal inferences, are all examples of causal identification via "selection on observables"

- Intuitively: If we are able to account for all of the X variables that affect both selection into treatment, D, AND variation in the outcome, Y, then we can claim independence between D and potential outcomes, which we need to talk about causality
- The big problem: We have to be able to observe and measure all of the X variables that affect selection into treatment and variation in outcomes

Multivariate linear regression, which you learned prior to this class, if used to make causal inferences, is another example

Selection On Observables: The Familiar Case

Subclassification, matching, and propensity-score weighting, when used to make causal inferences, are all examples of causal identification via "selection on observables"

- Intuitively: If we are able to account for all of the X variables that affect both selection into treatment, D, AND variation in the outcome, Y, then we can claim independence between D and potential outcomes, which we need to talk about causality
- The big problem: We have to be able to observe and measure all of the X variables that affect selection into treatment and variation in outcomes

Multivariate linear regression, which you learned prior to this class, if used to make causal inferences, is another example

- Omitted variable bias

Selection on Observables: The Familiar Case

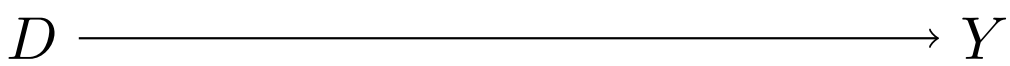
Multivariate linear regression estimates the impact of a treatment, D on an outcome, Y, overcoming omitted variable bias, as follows

$$Y_i = \alpha + \delta D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$

Selection on Observables: The Familiar Case

Multivariate linear regression estimates the impact of a treatment, D on an outcome, Y, overcoming omitted variable bias, as follows

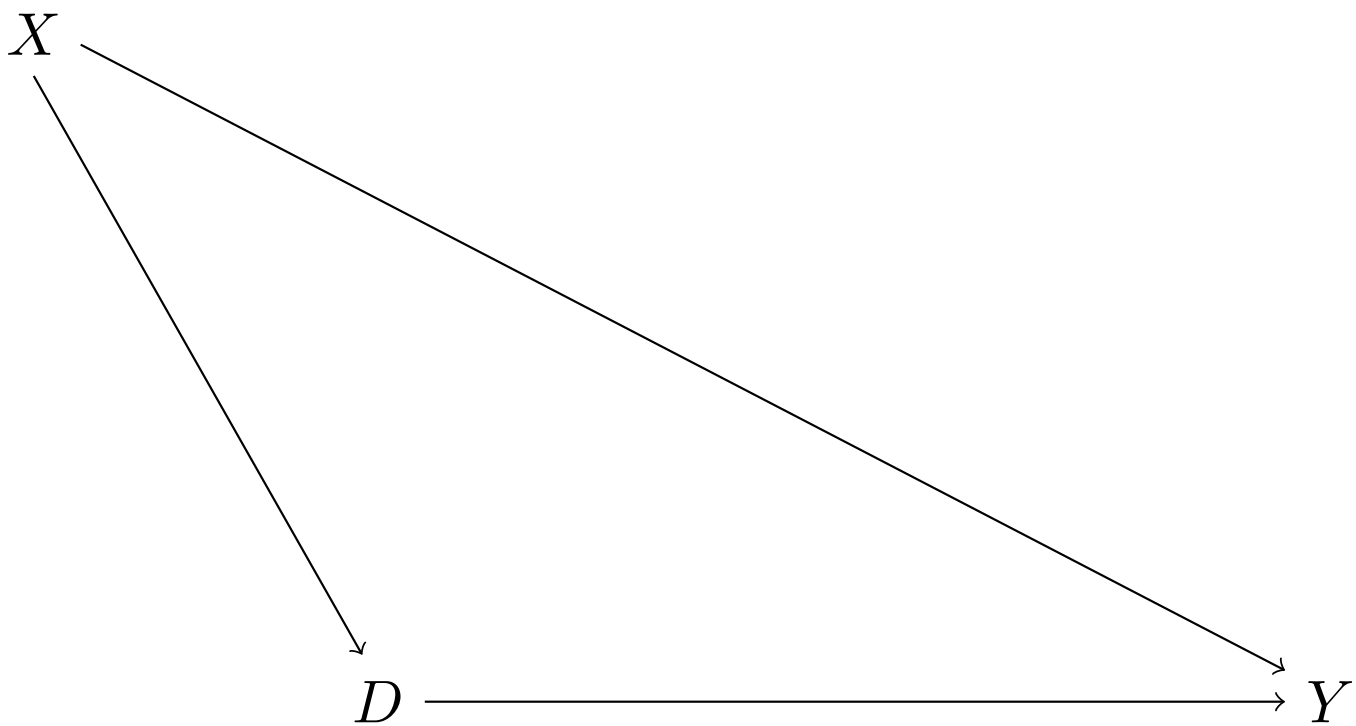
$$Y_i = \alpha + \delta D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$



Selection on Observables: The Familiar Case

Multivariate linear regression estimates the impact of a treatment, D on an outcome, Y, overcoming omitted variable bias, as follows

$$Y_i = \alpha + \delta D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$



Covariates, Outcomes, and Post-Treatment Bias

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D if for each individual i, $X_{0i} = X_{1i}$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

Covariates, Outcomes, and Post-Treatment Bias

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D if for each individual i, $X_{0i} = X_{1i}$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Does not imply that X and D are independent

Covariates, Outcomes, and Post-Treatment Bias

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D if for each individual i, $X_{0i} = X_{1i}$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Does not imply that X and D are independent
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary

Covariates, Outcomes, and Post-Treatment Bias

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D if for each individual i, $X_{0i} = X_{1i}$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Does not imply that X and D are independent
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary
- Baseline measures are often not only the most predictive variables, but also most important to remove confounding

Covariates, Outcomes, and Post-Treatment Bias

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D if for each individual i, $X_{0i} = X_{1i}$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Does not imply that X and D are independent
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary
- Baseline measures are often not only the most predictive variables, but also most important to remove confounding

Definition (Outcomes)

Those variables, Y, that are (possibly) not predetermined are called outcomes (for some individual i, $Y_{0i} \neq Y_{1i}$)

Covariates, Outcomes, and Post-Treatment Bias

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D if for each individual i, $X_{0i} = X_{1i}$, i.e., the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Does not imply that X and D are independent
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary
- Baseline measures are often not only the most predictive variables, but also most important to remove confounding

Definition (Outcomes)

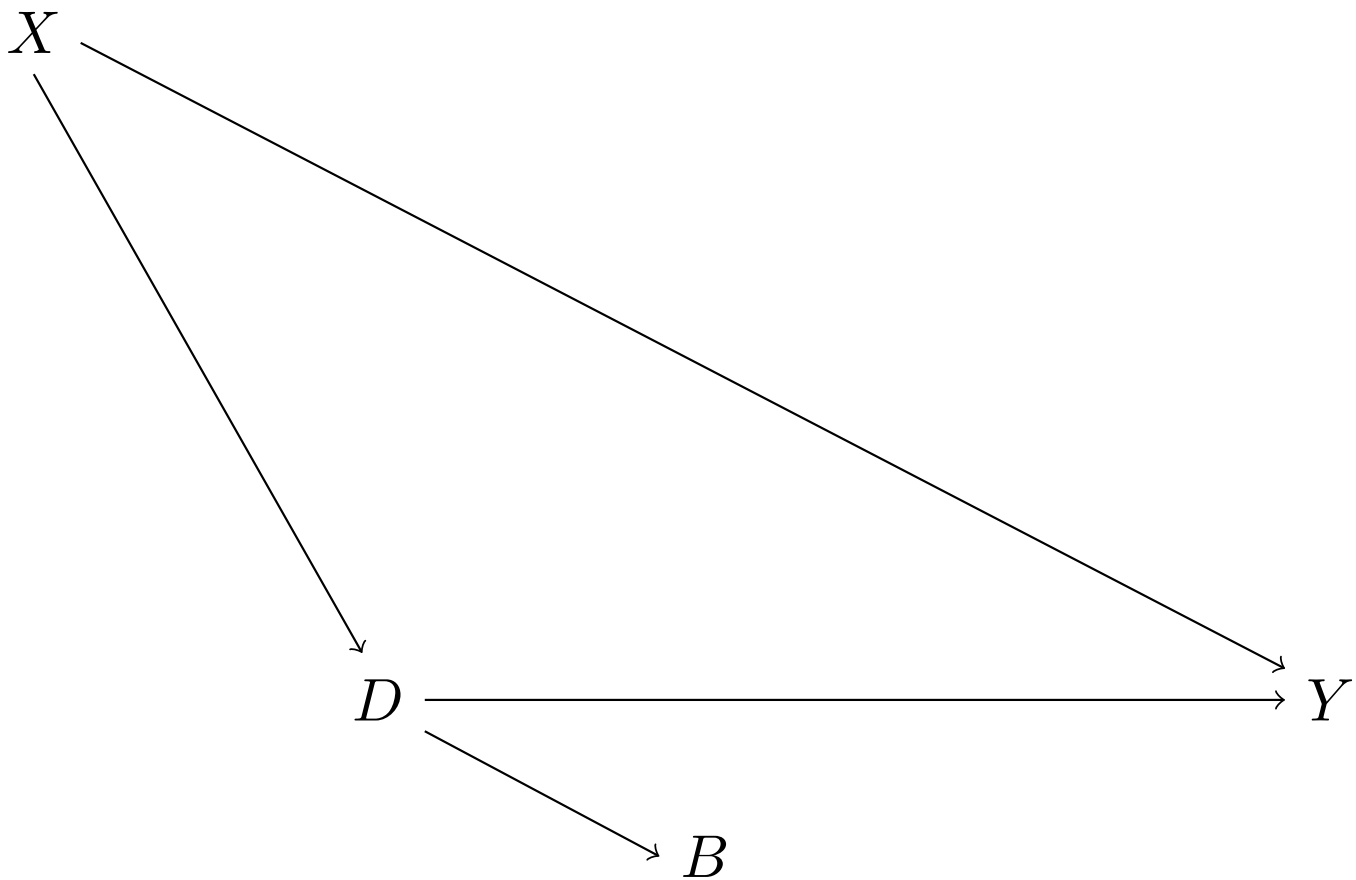
Those variables, Y, that are (possibly) not predetermined are called outcomes (for some individual i, $Y_{0i} \neq Y_{1i}$)

In general, one should not condition on outcomes, because this may induce post-treatment bias

Selection on Observables: The Familiar Case

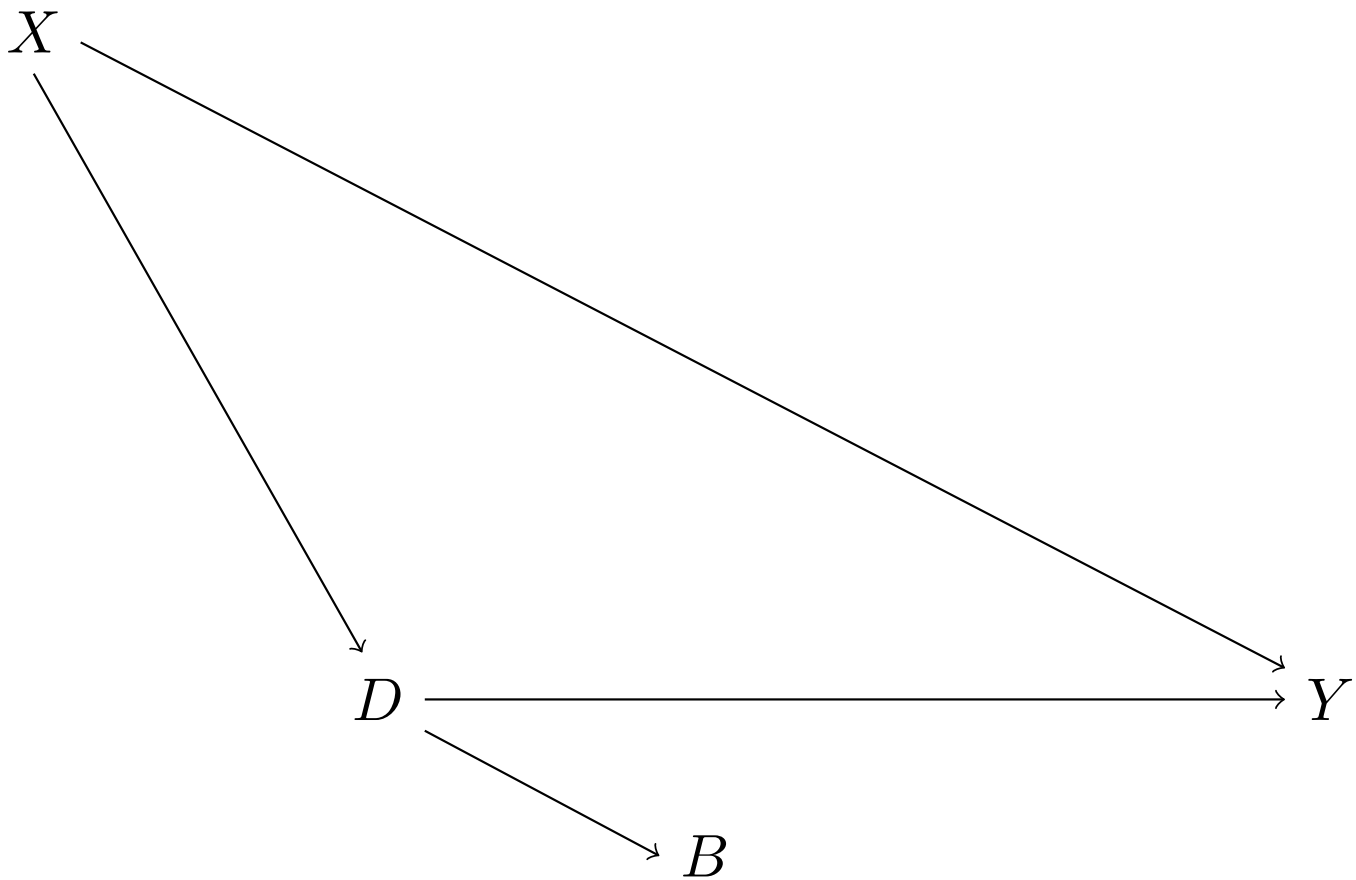
In multivariate linear regression, post-treatment bias would be induced by controlling for variables like B

$$Y_i = \alpha + \delta D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$



Selection on Observables: The Familiar Case

$$Y_i = \alpha + \delta D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$



If we control for variables like B, some of the real explained variance from D to Y will be accounted for by the coefficient estimate for B, which is not what we want

Regression vs. Matching and Weighting

A note of caution:

- For those interested in causal inference, matching and weighting are often promoted as causal inference methods, while multivariate regression is criticized as a valid way to achieve causal identification

Regression vs. Matching and Weighting

A note of caution:

- For those interested in causal inference, matching and weighting are often promoted as causal inference methods, while multivariate regression is criticized as a valid way to achieve causal identification
- But be very careful about this!

Regression vs. Matching and Weighting

A note of caution:

- For those interested in causal inference, matching and weighting are often promoted as causal inference methods, while multivariate regression is criticized as a valid way to achieve causal identification
- But be very careful about this!

- Both regression and matching/weighting achieve causal identification via the selection-on-observables assumption
  - Implies no unmeasured confounders
  - Equally (im)plausible irrespective of the estimation method

The Logic of Matching Methods

So if multivariate regression and matching/weighting are relying on the same underlying assumption to reach a causal interpretation of their estimates, then why do we not just use regression?

The Logic of Matching Methods

So if multivariate regression and matching/weighting are relying on the same underlying assumption to reach a causal interpretation of their estimates, then why do we not just use regression?

- Matching and weighting have the additional benefit of reducing so-called "model dependence"
- That is, we rid ourselves of the functional-form assumptions necessary for multivariate regression to work

The Logic of Matching Methods

Instead of relying on a functional form assumption, we...

1. Construct a sample of treated units and their counterfactuals (i.e., their "best" matches from the pool of control units

The Logic of Matching Methods

Instead of relying on a functional form assumption, we...

1. Construct a sample of treated units and their counterfactuals (i.e., their "best" matches from the pool of control units
    - Matching/weighting relies on the common support assumption to ensure comparability between treatment and control cases
    - Regression is not concerned with comparability, and simply uses the functional form of the model in regions where there are no comparable cases

The Logic of Matching Methods

Instead of relying on a functional form assumption, we...

1. Construct a sample of treated units and their counterfactuals (i.e., their "best" matches from the pool of control units
   - Matching/weighting relies on the common support assumption to ensure comparability between treatment and control cases
   - Regression is not concerned with comparability, and simply uses the functional form of the model in regions where there are no comparable cases
2. Calculate a simple difference of means (with some corrections to the point estimates and standard errors) to estimate a causal effect

The Logic of Matching Methods

Achieve an observational study that looks as close as possible to an experiment by focusing on observed treated units and constructing a set of control units that are "as similar as possible" on all relevant units

The Logic of Matching Methods

Achieve an observational study that looks as close as possible to an experiment by focusing on observed treated units and constructing a set of control units that are "as similar as possible" on all relevant units

- Suggests we will often be interested in the *average treatment effect on the treated* rather than the *average treatment effect*
    - Take the treated units as given and worry about constructing their counterfactuals

The Logic of Matching Methods

Achieve an observational study that looks as close as possible to an experiment by focusing on observed treated units and constructing a set of control units that are "as similar as possible" on all relevant units

- Suggests we will often be interested in the *average treatment effect on the treated* rather than the *average treatment effect*
  - Take the treated units as given and worry about constructing their counterfactuals
- Often, but not always, all of the hard work of achieving a plausible case for causal inference will happen through sample construction

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
   - *Treatment is random, i.e., independent of potential outcomes,* ***conditional*** *on* X
   - *Often called the Conditional Independence Assumption*
   - *Implies that we need to observe all factors that are correlated with both the outcome,* Y*, and treatment assignment,* D *(i.e., confounders)*
     - *Equivalent to "no omitted variables" assumption in regression*

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
   - *Treatment is random, i.e., independent of potential outcomes,* ***conditional*** *on* X
   - *Often called the Conditional Independence Assumption*
   - *Implies that we need to observe all factors that are correlated with both the outcome,* Y*, and treatment assignment,* D *(i.e., confounders)*
     - *Equivalent to "no omitted variables" assumption in regression*
2. $0 < \Pr(D = 1|X = x) < 1 \ \forall \ X$ *(common support)*

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
   - *Treatment is random, i.e., independent of potential outcomes,* ***conditional*** *on* X
   - *Often called the Conditional Independence Assumption*
   - *Implies that we need to observe all factors that are correlated with both the outcome,* Y*, and treatment assignment,* D *(i.e., confounders)*
     - *Equivalent to "no omitted variables" assumption in regression*
2. $0 < \Pr(D = 1|X = x) < 1 \ \forall \ X$ *(common support)*
   - *For any given realized value of* $X = x$*, there is at least one unit in each of the experimental groups (in this case,* $D = 1$ *and* $D = 0$*)*
   - *If we do not have this, we are unable to estimate counterfactual expected outcomes conditional on* $X = x$
   - *Note: In least squares estimation, we ignore common support and simply use linearity to project conditional expectations onto regions of the distribution of* X *where there is no data*

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \; \forall \; X$ *(common support)*

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \; \forall \; X$ *(common support)*

Identification Result

Given selection on observables we have

$$\begin{aligned} E[Y_1 - Y_0|X] &= E[Y_1 - Y_0|X, D = 1] \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

Not Examinable

Therefore, under the common support condition:

$$\begin{aligned} \alpha_{ATE} &= E[Y_1 - Y_0] = \int E[Y_1 - Y_0|X] \, dP(X) \\ &= \int \big(E[Y|X, D = 1] - E[Y|X, D = 0]\big) \, dP(X) \end{aligned}$$

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \ \forall \ X$ *(common support)*

Not Examinable

Identification Result

$$\begin{aligned} \alpha_{ATE} &= E[Y_1 - Y_0] = \int E[Y_1 - Y_0|X] \, dP(X) \\ &= \int \left(E[Y|X, D = 1] - E[Y|X, D = 0] \right) dP(X) \end{aligned}$$

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \ \forall \ X$ *(common support)*

Not Examinable

Identification Result

$$\begin{aligned} \alpha_{ATE} &= E[Y_1 - Y_0] = \int E[Y_1 - Y_0|X]\, dP(X) \\ &= \int \left(E[Y|X, D = 1] - E[Y|X, D = 0]\right) dP(X) \end{aligned}$$

- *Here, we want to be able to use the unconditional expectation of the mean difference, but we have to consider the conditional expectation*
- *Integrating over all possible values of X gives us the marginal distribution of the expectation of the mean difference of Y*
  - *Like a continuous analog to the partition theorem:*
    For any $\{B_n : n = 1, 2, \ldots\}$ that represents a finite or infinitely countable partition of the sample space and for which each B_n is measurable, then for any event A in the same probability space:

$$\Pr(A) = \sum_n \Pr(A \cap B_n) = \sum_n \Pr(A|B_n) \Pr(B_n)$$

Identification under Selection on Observables

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \; \forall \; X$ *(common support)*

Not Examinable

Identification Result

Similarly,

$$\begin{aligned} \alpha_{ATT} &= E[Y_1 - Y_0|D = 1] \\ &= \int \left(E[Y|X, D = 1] - E[Y|X, D = 0]\right) dP(X|D = 1) \end{aligned}$$

To identify α_{ATT} the selection on observables and common support conditions can be relaxed to:

- $Y_0 \perp\!\!\!\perp D|X$
- $\Pr(D = 1|X = x) < 1$ *(with* $\Pr(D = 1) > 0$*)*

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates,* X*, such that after controlling for* X*, potential outcomes are independent of treatment status*

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates,* X*, such that after controlling for* X*, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*
2. *Common Support*

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*
2. *Common Support*
   - *ATE Version:* $0 < \Pr(D = 1|X = x) < 1$

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*
2. *Common Support*
   - *ATE Version:* $0 < \Pr(D = 1|X = x) < 1$
     - *For each value of X, there is a positive probability of observing both treated and untreated units*

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*
2. *Common Support*
   - *ATE Version:* $0 < \Pr(D = 1|X = x) < 1$
     - *For each value of X, there is a positive probability of observing both treated and untreated units*
   - *ATT Version:* $\Pr(D = 1|X = x) < 1$, *with* $\Pr(D = 1) > 0$

Identification under Selection on Observables

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $(Y_1, Y_0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y_0 \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*
2. *Common Support*
   - *ATE Version:* $0 < \Pr(D = 1|X = x) < 1$
     - *For each value of X, there is a positive probability of observing both treated and untreated units*
   - *ATT Version:* $\Pr(D = 1|X = x) < 1$, *with* $\Pr(D = 1) > 0$
     - *For each value of X observed for a treated unit, there should exist an untreated unit with the same value of X*

Outline

Matching

When X is continuous we can estimate α_{ATT} by "imputing" the missing potential outcome of each treated unit using the observed outcome from the "closest" control unit:

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

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.

Matching

When X is continuous we can estimate α_{ATT} by "imputing" the missing potential outcome of each treated unit using the observed outcome from the "closest" control unit:

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

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 for M closest matches:

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

Works well when we can find good matches for each treated unit.

Matching: Example with 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

Question

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

Matching: Example with 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

Question

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

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

Matching: Distance Metric

Note: Matrix multiplication and inversion are not examinable

Matching: Distance Metric

"Closeness" is often defined by a distance metric that projects the distance between the multivariate covariate vectors $X_i = (X_{i1}, X_{i2}, ..., X_{ik})'$ and $X_j = (X_{j1}, X_{j2}, ..., X_{jk})'$ onto a univariate scale

Matching: Distance Metric

"Closeness" is often defined by a distance metric that projects the distance between the multivariate covariate vectors $X_i = (X_{i1}, X_{i2}, ..., X_{ik})'$ and $X_j = (X_{j1}, X_{j2}, ..., X_{jk})'$ onto a univariate scale

A commonly used distance is the Mahalanobis distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)} \ ,$$

where S is the sample variance-covariance-matrix.

Matching: Distance Metric

Through multiplication by S^{-1}, the Mahalanobis distance takes scale and correlation of covariates into account by:

- standardization via main-diagonal
- down-weighting via off-diagonal

Matching: Distance Metric

Through multiplication by S^{-1}, the Mahalanobis distance takes scale and correlation of covariates into account by:

- standardization via main-diagonal
- down-weighting via off-diagonal

Works great for multivariate normal-distributed data, not so great with outliers or binary data.

Matching: Distance Metric

Through multiplication by S^{-1}, the Mahalanobis distance takes scale and correlation of covariates into account by:

- standardization via main-diagonal
- down-weighting via off-diagonal

Works great for multivariate normal-distributed data, not so great with outliers or binary data.

A popular alternative is the normalized Euclidean distance

$$D_A(X_i, X_j) = \sqrt{(X_i - X_j)' A (X_i - X_j)} \ ,$$

where A is the $k \times k$ diagonal matrix such that $diag(A) = diag(S^{-1})$.

Matching on Covariates: Example

Covariate vector treated unit i:

$$X_i = (38, 1, 0) \ .$$

Covariate vector control unit j:

$$X_j = (48, 0, 0) \ .$$

Matching on Covariates: Example

Covariate vector treated unit i:

$$X_i = (38, 1, 0) \ .$$

Covariate vector control unit j:

$$X_j = (48, 0, 0) \ .$$

The sample variance-covariance matrix of X:

$$S = \begin{pmatrix} 54.04 & 0.39 & -0.94 \\ 0.39 & 0.13 & 0.02 \\ -0.94 & 0.02 & 0.25 \end{pmatrix} \ ,$$

Matching on Covariates: Example

Covariate vector treated unit i:

$$X_i = (38, 1, 0) \ .$$

Covariate vector control unit j:

$$X_j = (48, 0, 0) \ .$$

The sample variance-covariance matrix of X:

$$S = \begin{pmatrix} 54.04 & 0.39 & -0.94 \\ 0.39 & 0.13 & 0.02 \\ -0.94 & 0.02 & 0.25 \end{pmatrix} \ ,$$

and its inverse

$$S^{-1} = \begin{pmatrix} 0.021 & -0.077 & 0.084 \\ -0.077 & 8.127 & -1.009 \\ 0.084 & -1.009 & 4.407 \end{pmatrix} \ .$$

Matching on Covariates: Example

Distance Matrix:

$$\begin{bmatrix} D(X_1,X_1) & D(X_1,X_2) & D(X_1,X_3) & \cdots & D(X_1,X_n) \\ D(X_2,X_1) & D(X_2,X_2) & D(X_2,X_3) & \cdots & D(X_2,X_n) \\ D(X_3,X_1) & D(X_3,X_2) & D(X_3,X_3) & \cdots & D(X_3,X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D(X_n,X_1) & D(X_n,X_2) & D(X_n,X_3) & \cdots & D(X_n,X_n) \end{bmatrix}$$

$$= \begin{bmatrix} 0 & D(X_1,X_2) & D(X_1,X_3) & \cdots & D(X_1,X_n) \\ D(X_2,X_1) & 0 & D(X_2,X_3) & \cdots & D(X_2,X_n) \\ D(X_3,X_1) & D(X_3,X_2) & 0 & \cdots & D(X_3,X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D(X_n,X_1) & D(X_n,X_2) & D(X_n,X_3) & \cdots & 0 \end{bmatrix}$$

Matching on Covariates: Example

Under matching with replacement, pick a treated unit i and, for each control unit j, calculate the distance:

Matching on Covariates: Example

Under matching with replacement, pick a treated unit i and, for each control unit j, calculate the distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

$$= \sqrt{\begin{pmatrix} 38-48 & 1-0 & 0-0 \end{pmatrix} \begin{pmatrix} 0.021 & -0.077 & 0.084 \\ -0.077 & 8.127 & -1.009 \\ 0.084 & -1.009 & 4.407 \end{pmatrix} \begin{pmatrix} 38-48 \\ 1-0 \\ 0-0 \end{pmatrix}}$$

$$= \sqrt{11.59} = 3.4$$

Matching on Covariates: Example

Under matching with replacement, pick a treated unit i and, for each control unit j, calculate the distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

$$= \sqrt{\begin{pmatrix} 38-48 & 1-0 & 0-0 \end{pmatrix} \begin{pmatrix} 0.021 & -0.077 & 0.084 \\ -0.077 & 8.127 & -1.009 \\ 0.084 & -1.009 & 4.407 \end{pmatrix} \begin{pmatrix} 38-48 \\ 1-0 \\ 0-0 \end{pmatrix}}$$

$$= \sqrt{11.59} = 3.4$$

or

$$D_A(X_i, X_j) = \sqrt{(X_i - X_j)' A (X_i - X_j)}$$

$$= \sqrt{\begin{pmatrix} -10 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.021 & 0 & 0 \\ 0 & 8.127 & 0 \\ 0 & 0 & 4.407 \end{pmatrix} \begin{pmatrix} -10 \\ 1 \\ 0 \end{pmatrix}}$$

$$= \sqrt{10.1} = 3.2$$

Matching on Covariates: Example

Under matching with replacement, pick a treated unit i and, for each control unit j, calculate the distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

$$= \sqrt{\begin{pmatrix} 38-48 & 1-0 & 0-0 \end{pmatrix} \begin{pmatrix} 0.021 & -0.077 & 0.084 \\ -0.077 & 8.127 & -1.009 \\ 0.084 & -1.009 & 4.407 \end{pmatrix} \begin{pmatrix} 38-48 \\ 1-0 \\ 0-0 \end{pmatrix}}$$

$$= \sqrt{11.59} = 3.4$$

or

$$D_A(X_i, X_j) = \sqrt{(X_i - X_j)' A (X_i - X_j)}$$

$$= \sqrt{\begin{pmatrix} -10 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.021 & 0 & 0 \\ 0 & 8.127 & 0 \\ 0 & 0 & 4.407 \end{pmatrix} \begin{pmatrix} -10 \\ 1 \\ 0 \end{pmatrix}}$$

$$= \sqrt{10.1} = 3.2$$

and match with the one control unit $j(i)$ that minimizes $D(X_i, X_j)$ or $D_A(X_i, X_j)$, respectively.

Matching on Covariates: Example

Distance Matrix:

$$= \begin{bmatrix} 0 & D(X_1, X_2) & D(X_1, X_3) & \cdots & D(X_1, X_n) \\ D(X_2, X_1) & 0 & D(X_2, X_3) & \cdots & D(X_2, X_n) \\ D(X_3, X_1) & D(X_3, X_2) & 0 & \cdots & D(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D(X_n, X_1) & D(X_n, X_2) & D(X_n, X_3) & \cdots & 0 \end{bmatrix}$$

Matching on Covariates: Example

Distance Matrix:

$$= \begin{bmatrix} 0 & D(X_1, X_2) & D(X_1, X_3) & \cdots & D(X_1, X_n) \\ D(X_2, X_1) & 0 & D(X_2, X_3) & \cdots & D(X_2, X_n) \\ D(X_3, X_1) & D(X_3, X_2) & 0 & \cdots & D(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D(X_n, X_1) & D(X_n, X_2) & D(X_n, X_3) & \cdots & 0 \end{bmatrix}$$

- For 1:1 matching with replacement
  - Look across each row and select the column for the lowest value as the matching case

Matching on Covariates: Example

Distance Matrix:

$$= \begin{bmatrix} 0 & D(X_1, X_2) & D(X_1, X_3) & \cdots & D(X_1, X_n) \\ D(X_2, X_1) & 0 & D(X_2, X_3) & \cdots & D(X_2, X_n) \\ D(X_3, X_1) & D(X_3, X_2) & 0 & \cdots & D(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D(X_n, X_1) & D(X_n, X_2) & D(X_n, X_3) & \cdots & 0 \end{bmatrix}$$

- For 1:1 matching with replacement
  - Look across each row and select the column for the lowest value as the matching case
- For 1:2 matching with replacement
  - Look across each row and select the columns with the two lowest values and take their average as the matching case

Matching on Covariates: Example

Distance Matrix:

$$= \begin{bmatrix} 0 & D(X_1, X_2) & D(X_1, X_3) & \cdots & D(X_1, X_n) \\ D(X_2, X_1) & 0 & D(X_2, X_3) & \cdots & D(X_2, X_n) \\ D(X_3, X_1) & D(X_3, X_2) & 0 & \cdots & D(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D(X_n, X_1) & D(X_n, X_2) & D(X_n, X_3) & \cdots & 0 \end{bmatrix}$$

- For 1:1 matching with replacement
  - Look across each row and select the column for the lowest value as the matching case
- For 1:2 matching with replacement
  - Look across each row and select the columns with the two lowest values and take their average as the matching case
- For 1:1 matching without replacement
  - Look across the first row and select the column for the lowest value as the matching case
  - Remove that column from consideration
  - Repeat above two steps for all remaining rows

Matching: Bias Correction

Matching estimators may behave badly if X contains multiple continuous variables.

Need to adjust matching estimators in the following way:

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

where $\mu_0(x) = E[Y|X = x, D = 0]$ and $\widehat{\mu}_0$ is an estimate of μ_0.

These "bias-corrected" matching estimators behave well even if μ_0 is estimated using a simple linear regression (i.e., $\mu_0(x) = \beta_0 + \beta_1 x$) (Abadie and Imbens, 2006).

Matching: Bias Derivation

It can be shown that if the dimension of X_i is large enough

$$\sqrt{N_1}(X_{j(i)} - X_i) \not\to 0.$$

This implies:

$$\sqrt{N_1}\Big(\mu_0(X_{j(i)}) - \mu_0(X_i)\Big) \not\to 0,$$

$$\frac{1}{\sqrt{N_1}} \sum_{D_i=1} \left(\mu_0(X_i) - \mu_0(X_{j(i)})\right) \not\to 0,$$

and therefore

$$\sqrt{N_1}\Big(\widehat{\alpha}_{ATT} - \alpha_{ATT}\Big) \not\to 0.$$

Matching with Bias Correction

Each treated observation contributes

$$\mu_0(X_i) - \mu_0(X_{j(i)})$$

to the bias.

Bias-corrected matching:

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

The large sample distribution of this estimator (for the case of matching with replacement) is derived in Abadie and Imbens (2006). For matching without replacement, things are more complicated, but see the Martingale representation in Abadie and Imbens (2009).

Bias-Variance Tradeoff in Matching

Once we settle on a particular distance metric that we want to use, we might use:

Bias-Variance Tradeoff in Matching

Once we settle on a particular distance metric that we want to use, we might use:

1. Matching with replacement or matching without replacement, and

Bias-Variance Tradeoff in Matching

Once we settle on a particular distance metric that we want to use, we might use:

1. Matching with replacement or matching without replacement, and
2. 1:1 or 1:M nearest-neighbor matching, with $M > 1$

Bias-Variance Tradeoff in Matching

Once we settle on a particular distance metric that we want to use, we might use:

1. Matching with replacement or matching without replacement, and
2. 1:1 or 1:M nearest-neighbor matching, with $M > 1$
   - 1:1 matching means that for every treated unit, we construct a control by identifying the one closest match among the control cases
   - 1:M matching means that for every treated unit, we construct a control by identifying the M closest matches among the control cases, averaging over their X and Y values, thereby constructing a composite control

Bias-Variance Tradeoff in Matching

The decisions of whether to use matching with vs. without replacement and 1:1 vs. 1:M matching both involve a decision to use more or less cases

Bias-Variance Tradeoff in Matching

The decisions of whether to use matching with vs. without replacement and 1:1 vs. 1:M matching both involve a decision to use more or less cases

- In 1:1 matching, we use 1 control unit per treated unit, while in 1:M matching, we use M control units per treated unit

Bias-Variance Tradeoff in Matching

The decisions of whether to use matching with vs. without replacement and 1:1 vs. 1:M matching both involve a decision to use more or less cases

- In 1:1 matching, we use 1 control unit per treated unit, while in 1:M matching, we use M control units per treated unit
- In matching with replacement, we use the same case(s) potentially more than once, and hence end up using fewer cases than in matching without replacement

Bias-Variance Tradeoff in Matching

Bias-Variance Tradeoff in Matching

Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

Bias-Variance Tradeoff in Matching

Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

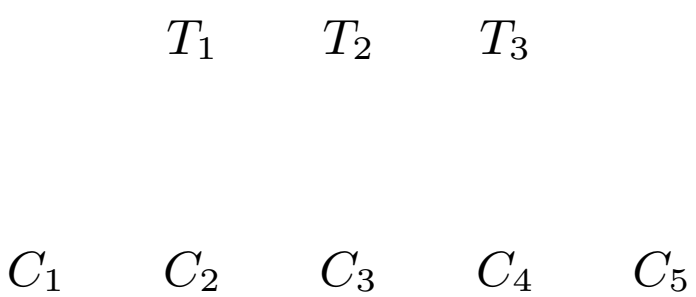
- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

Bias-Variance Tradeoff in Matching

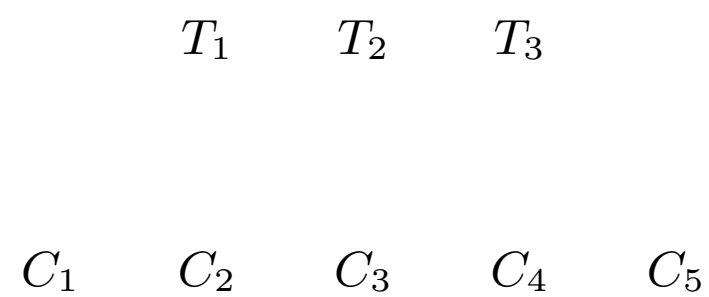
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:1 Matching without Replacement

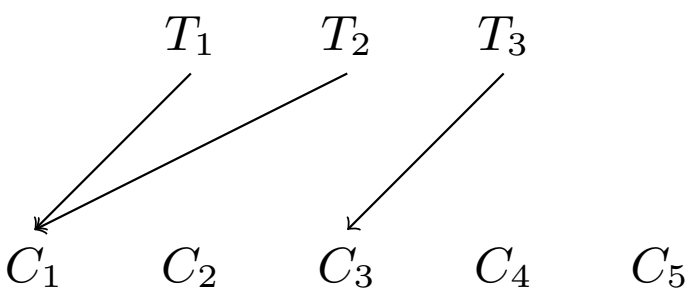


Bias-Variance Tradeoff in Matching

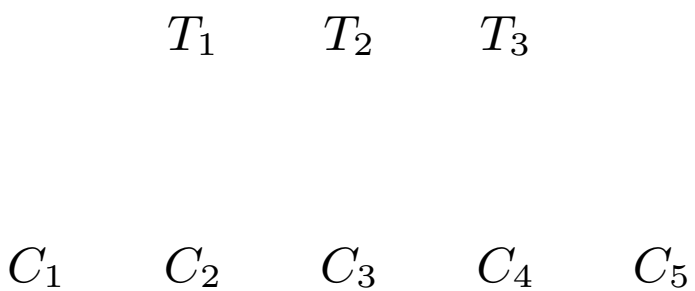
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:1 Matching without Replacement

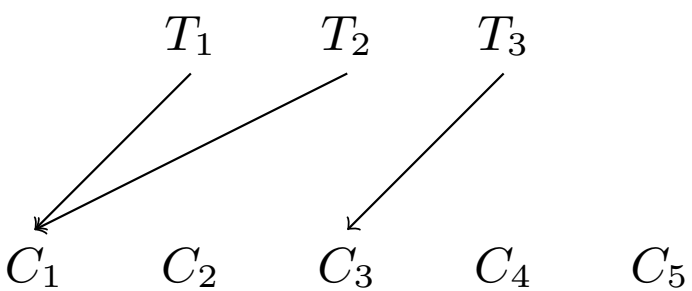


Bias-Variance Tradeoff in Matching

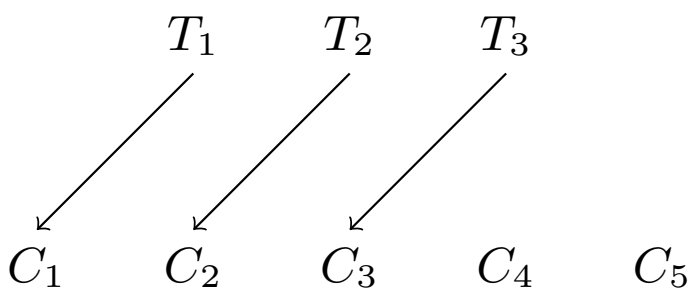
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:1 Matching without Replacement

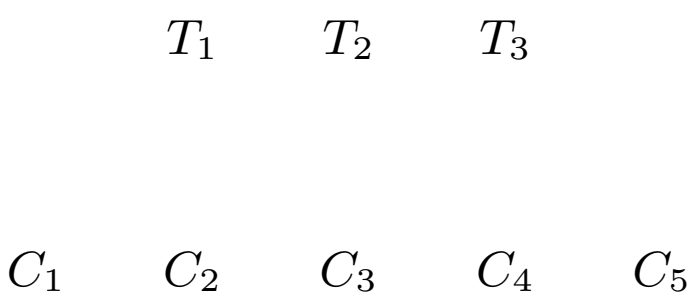


Bias-Variance Tradeoff in Matching

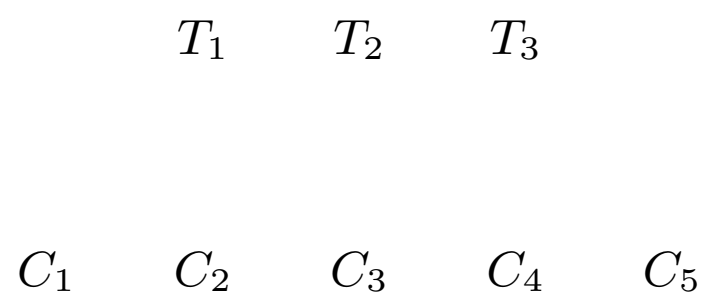
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:2 Matching with Replacement

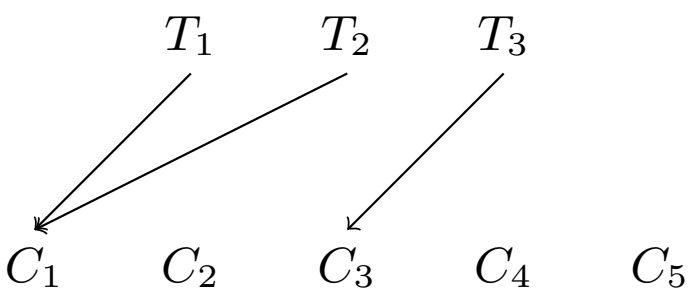


Bias-Variance Tradeoff in Matching

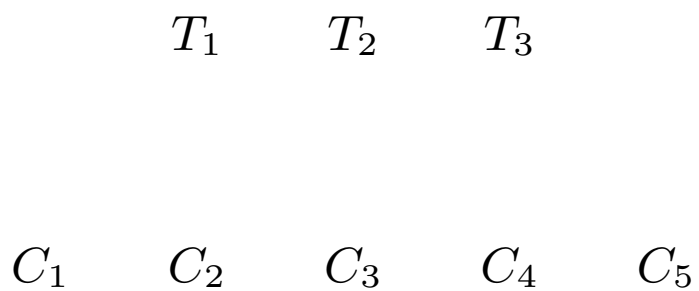
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:2 Matching with Replacement

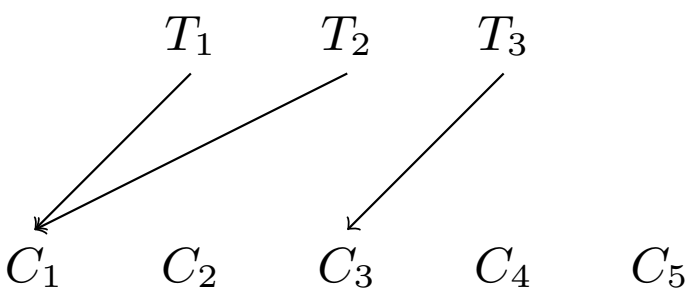


Bias-Variance Tradeoff in Matching

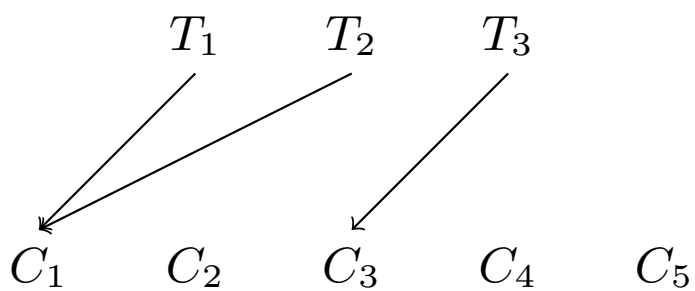
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:2 Matching with Replacement

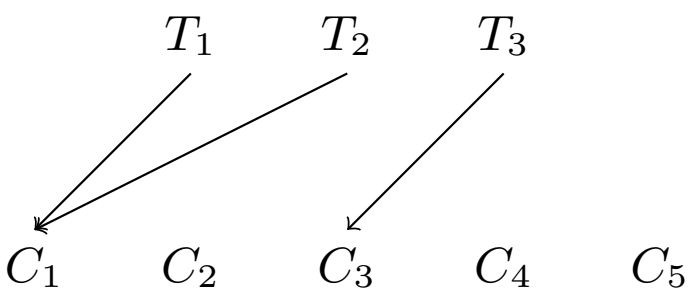


Bias-Variance Tradeoff in Matching

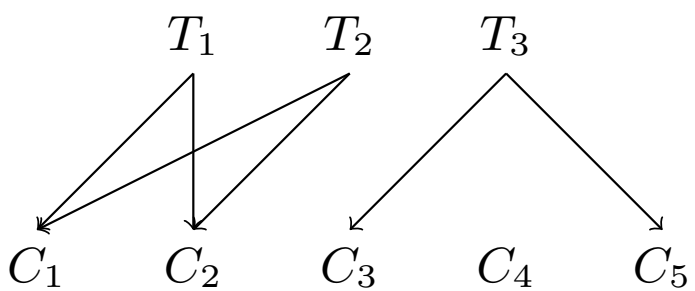
Example Scenario: 3 Treatment cases (T_1, T_2, T_3) and 5 control cases (C_1, C_2, C_3, C_4, C_5)

- T_1's closest match is C_1 and second closest match is C_2
- T_2's closest match is C_1 and second closest match is C_2
- T_3's closest match is C_3 and second closest match is C_5

1:1 Matching with Replacement



1:2 Matching with Replacement



Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Likewise, methods that use more control cases introduce more bias, but lower variance

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Likewise, methods that use more control cases introduce more bias, but lower variance

General Rules:

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Likewise, methods that use more control cases introduce more bias, but lower variance

General Rules:

- Using 1:1 matching achieves bias that is as small as or smaller than any 1:M matching procedure on the same data

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Likewise, methods that use more control cases introduce more bias, but lower variance

General Rules:

- Using 1:1 matching achieves bias that is as small as or smaller than any 1:M matching procedure on the same data
- Using matching with replacement achieves bias that is as small as or smaller than any procedure using matching without replacement on the same data

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Likewise, methods that use more control cases introduce more bias, but lower variance

General Rules:

- Using 1:1 matching achieves bias that is as small as or smaller than any 1:M matching procedure on the same data
- Using matching with replacement achieves bias that is as small as or smaller than any procedure using matching without replacement on the same data
- Using 1:M matching achieves variance that is as small as or smaller than any 1:1 matching procedure on the same data

Bias-Variance Tradeoff in Matching

In the previous example situations, the method that uses fewer control cases

- Achieves superior matches, and hence less bias
  - Matching to the one closest case instead of averaging over the closest, second closest, and so on
  - Matching to the one closest case, even if that case has to be used more than once
- But also higher variance, and hence less precise estimates, because we are using less information

Likewise, methods that use more control cases introduce more bias, but lower variance

General Rules:

- Using 1:1 matching achieves bias that is as small as or smaller than any 1:M matching procedure on the same data
- Using matching with replacement achieves bias that is as small as or smaller than any procedure using matching without replacement on the same data
- Using 1:M matching achieves variance that is as small as or smaller than any 1:1 matching procedure on the same data
- Using matching without replacement achieves variance that is as small as or smaller than any procedure using matching with replacement on the same data

Identification with the Propensity Score

Definition

Propensity score is defined as the selection probability conditional on the confounding variables: $\pi(X) = \Pr(D = 1|X)$

Identification with the Propensity Score

Definition

Propensity score is defined as the selection probability conditional on the confounding variables: $\pi(X) = \Pr(D = 1|X)$

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \forall X$ *(common support)*

Identification with the Propensity Score

Definition

Propensity score is defined as the selection probability conditional on the confounding variables: $\pi(X) = \Pr(D = 1|X)$

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \forall X$ *(common support)*

Identification Result

If $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *and* $0 < \Pr(D = 1|X = x) < 1 \forall X$, *then* $(Y_1, Y_0) \perp\!\!\!\perp D|\pi(X)$

Identification with the Propensity Score

Definition

Propensity score is defined as the selection probability conditional on the confounding variables: $\pi(X) = \Pr(D = 1|X)$

Identification Assumptions

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X = x) < 1 \forall X$ *(common support)*

Identification Result

If $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *and* $0 < \Pr(D = 1|X = x) < 1 \forall X$, *then* $(Y_1, Y_0) \perp\!\!\!\perp D|\pi(X)$

- *I.e., conditioning on the propensity score is enough to have independence between the treatment indicator and potential outcomes if and only if selection on observables and common support hold.*
- *Implies substantial dimension reduction*

Identification with the Propensity Score

Not Examinable:

Proof.

Properties of mathematical expectation:

- Law of Iterated Expectations: If $E[A]$ exists and A and B are defined on the same probability space, then $E[A] = E[E[A|B]]$
- If A and B are independent, $E[A|B] = E[A]$

Note that

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

Since

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

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

Matching on the Propensity Score

Corollary

If $(Y_1, Y_0) \perp\!\!\!\perp D|X$, *then*

$$E[Y|D=1, \pi(X)=\bar{\pi}] - E[Y|D=0, \pi(X)=\bar{\pi}] = E[Y_1 - Y_0|\pi(X)=\bar{\pi}]$$

Suggests a two step procedure to estimate causal effects under selection on covariates:

1. Estimate the propensity score $\pi(X) = \Pr(D=1|X)$
   - In practice, this is almost always done using logistic regression
2. Match units on their propensity scores rather than the multidimensional distance between their X variables

Matching on the Propensity Score: Example

Distance Matrix:

$$\begin{bmatrix} \pi(X)_1 - \pi(X)_1 & \pi(X)_1 - \pi(X)_2 & \pi(X)_1 - \pi(X)_3 & \cdots & \pi(X)_1 - \pi(X)_n \\ \pi(X)_2 - \pi(X)_1 & \pi(X)_2 - \pi(X)_2 & \pi(X)_2 - \pi(X)_3 & \cdots & \pi(X)_2 - \pi(X)_n \\ \pi(X)_3 - \pi(X)_1 & \pi(X)_3 - \pi(X)_2 & \pi(X)_3 - \pi(X)_3 & \cdots & \pi(X)_3 - \pi(X)_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi(X)_n - \pi(X)_1 & \pi(X)_n - \pi(X)_2 & \pi(X)_n - \pi(X)_3 & \cdots & \pi(X)_n - \pi(X)_n \end{bmatrix}$$

$$= \begin{bmatrix} 0 & \pi(X)_1 - \pi(X)_2 & \pi(X)_1 - \pi(X)_3 & \cdots & \pi(X)_1 - \pi(X)_n \\ \pi(X)_2 - \pi(X)_1 & 0 & \pi(X)_2 - \pi(X)_3 & \cdots & \pi(X)_2 - \pi(X)_n \\ \pi(X)_3 - \pi(X)_1 & \pi(X)_3 - \pi(X)_2 & 0 & \cdots & \pi(X)_3 - \pi(X)_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi(X)_n - \pi(X)_1 & \pi(X)_n - \pi(X)_2 & \pi(X)_n - \pi(X)_3 & \cdots & 0 \end{bmatrix}$$

Weighting on the Propensity Score

Proposition

Provided that the relevant moments exists, if $Y_1, Y_0 \perp\!\!\!\perp D|X$*, then*

$$\begin{aligned}
\alpha_{ATE} &= E[Y_1 - Y_0] = E\left[Y \cdot \frac{D - \pi(X)}{\pi(X) \cdot (1 - \pi(X))}\right] \\
\alpha_{ATT} &= E[Y_1 - Y_0|D = 1] = \frac{1}{\Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - \pi(X)}{1 - \pi(X)}\right]
\end{aligned}$$

Not Examinable:

Proof.

Property of mathematical expectation:

- If A is a random variable, $E[A]$ exists, and $f(\cdot)$ is a real-valued function, then $E[f(A)B|A] = f(A)E[B|A]$

$$\begin{aligned}
E\left[Y \cdot \frac{D - \pi(X)}{\pi(X)(1 - \pi(X))}\middle| X\right] &= \\
E[Y/\pi(X)|X, D = 1]\pi(X) + E[-Y/(1 - \pi(X))|X, D = 0](1 - \pi(X)) &= \\
E[Y|X, D = 1] - E[Y|X, D = 0]
\end{aligned}$$

Integrate over $\Pr(X)$ to get α_{ATE} and over $\Pr(X|D = 1)$ to get α_{ATT} □

Weighting on the Propensity Score

Proposition

$$\begin{aligned}
\alpha_{ATE} &= E[Y_1 - Y_0] = E\left[Y \cdot \frac{D - \pi(X)}{\pi(X) \cdot (1 - \pi(X))}\right] \\
\alpha_{ATT} &= E[Y_1 - Y_0 | D = 1] = \frac{1}{\Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - \pi(X)}{1 - \pi(X)}\right]
\end{aligned}$$

Suggests a two step procedure:

1. Estimate the propensity score ($\widehat{\pi}(X)$)
2. Use sample averages and the estimated propensity score to produce analog estimators of α_{ATE} and α_{ATT}:

$$\begin{aligned}
\widehat{\alpha}_{ATE} &= \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i) \cdot (1 - \widehat{\pi}(X_i))}, \\
\widehat{\alpha}_{ATT} &= \frac{1}{N_1} \sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{1 - \widehat{\pi}(X)}
\end{aligned}$$

Variance Estimation: Matching

- Best with replacement to eliminate biases.
- Let $K_i = \#$ times that observation i is used as a match,

$$\widehat{\alpha}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)}) = \frac{1}{N_1} \sum_{i=1} (D_i - (1 - D_i)K_i)Y_i.$$

- "Usual" standard errors come from

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_{ATT}) = \frac{1}{N_1^2} \sum_{D_i=1} (Y_i - Y_{j(i)} - \widehat{\alpha}_{ATT})^2.$$

- Correct standard errors come from

$$\begin{aligned} \widehat{\mathrm{Var}}(\widehat{\alpha}_{ATT}) &= \frac{1}{N_1^2} \sum_{D_i=1} (Y_i - Y_{j(i)} - \widehat{\alpha}_{ATT})^2 \\ &+ \frac{1}{N_1^2} \sum_{D_i=0} K_i(K_i - 1)\widehat{\mathrm{Var}}(Y_i|X_i, D_i). \end{aligned}$$

$\longrightarrow$ I.e., the usual standard errors are too small.

Variance Estimation: Weighting on Propensity Score

Recall the propensity score weighting estimator:

$$\widehat{\alpha}_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} Y_i \frac{D_i - \widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)}$$

Need to adjust standard errors for first-step estimation of $\pi(X)$

Two options:

Variance Estimation: Weighting on Propensity Score

Recall the propensity score weighting estimator:

$$\widehat{\alpha}_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} Y_i \frac{D_i - \widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)}$$

Need to adjust standard errors for first-step estimation of $\pi(X)$

Two options:

1. Analytical solution
   - Parametric first-step: Newey & McFadden (1994)
   - Non-parametric first-step: Newey (1994)
2. Bootstrap

Pushing Our Identification Assumptions

Usually, thinking that we have truly achieved the selection-on-observables assumption is unrealistic

- We are often just trying to convince ourselves that we have done "good enough"
- And, in any case, selection on observables is untestable

Pushing Our Identification Assumptions

Usually, thinking that we have truly achieved the selection-on-observables assumption is unrealistic

- We are often just trying to convince ourselves that we have done "good enough"
- And, in any case, selection on observables is untestable

But is there anything that we can do?

Sensitivity Analysis

- Sensitivity analysis is a way to subject our identification assumptions to a stress test

Sensitivity Analysis

- Sensitivity analysis is a way to subject our identification assumptions to a stress test
- It is not a formal test of whether an assumption is true or false

Sensitivity Analysis

- Sensitivity analysis is a way to subject our identification assumptions to a stress test
- It is not a formal test of whether an assumption is true or false
- Rather, it asks the question: How badly could we violate an assumption and still be confident in our results?

Sensitivity Analysis

- Sensitivity analysis is a way to subject our identification assumptions to a stress test
- It is not a formal test of whether an assumption is true or false
- Rather, it asks the question: How badly could we violate an assumption and still be confident in our results?

- In matching, most sensitivity analyses ask: How badly can we violate the selection-on-observables assumption and still be confident in our estimate of the treatment effect?

Sensitivity Analysis Using Rosenbaum Bounds

The logic of the most common form of sensitivity analysis in matching is as follows:

- Suppose we have units in our data that have an exact match on their covariates, but yet have different probabilities of being assigned to treatment
  - Formally, $X_j = X_k$ but $\pi(X_j) \neq \pi(X_k)$ for units j and k
  - Implies that there are one or more covariates needed to account for selection on observables that we have not accounted for
  - Known as "hidden bias"

Sensitivity Analysis Using Rosenbaum Bounds

The logic of the most common form of sensitivity analysis in matching is as follows:

- Suppose we have units in our data that have an exact match on their covariates, but yet have different probabilities of being assigned to treatment
  - Formally, $X_j = X_k$ but $\pi(X_j) \neq \pi(X_k)$ for units j and k
  - Implies that there are one or more covariates needed to account for selection on observables that we have not accounted for
  - Known as "hidden bias"
- Think of a factor Γ that represents how much two units with the same X values could differ in their odds of receiving treatment
  - $\Gamma = 1$ means no hidden bias
  - $\Gamma = 2$ means that two units with the same X could differ in their odds of receiving treatment by a factor of 2
  - ... and so on

Sensitivity Analysis Using Rosenbaum Bounds

Rosenbaum shows that for units j and k with an exact match on X_j and X_k,

$$\frac{1}{\Gamma} \leq \frac{\pi(X_j)/(1-\pi(X_j))}{\pi(X_k)/(1-\pi(X_k))} \leq \Gamma$$

Sensitivity Analysis Using Rosenbaum Bounds

Rosenbaum shows that for units j and k with an exact match on X_j and X_k,

$$\frac{1}{\Gamma} \leq \frac{\pi(X_j)/(1-\pi(X_j))}{\pi(X_k)/(1-\pi(X_k))} \leq \Gamma$$

In a sensitivity test using this framework, we would try out different values of Γ to show how inferences might change in the presence of hidden bias

Practical Steps in Matching

1. Like all data analysis exercises, explore your raw data before making a lot of assumptions, transformations, etc.

Practical Steps in Matching

1. Like all data analysis exercises, explore your raw data before making a lot of assumptions, transformations, etc.
    - In matching exercises, prior to constructing a matched sample, this can involve...
        - Simple difference-of-means tests on the outcome variable
        - Regression-adjusted difference-of-means tests on the outcome variable using pre-treatment covariates

Practical Steps in Matching

2. Construct the matched sample
   - Whether using distance-matching or propensity scores, try multiple specifications, using different choices for covariates, numbers of control units for each treatment unit, matching with and without replacement, etc.
   - Check for balance on pre-treatment covariates (e.g., t-tests for balance on pre-treatment covariates, graphical overlays of covariate densities between treatment groups)
   - For propensity-score matching, explore the distribution of the propensity score across treatment groups

Practical Steps in Matching

2. Construct the matched sample
    - Whether using distance-matching or propensity scores, try multiple specifications, using different choices for covariates, numbers of control units for each treatment unit, matching with and without replacement, etc.
    - Check for balance on pre-treatment covariates (e.g., t-tests for balance on pre-treatment covariates, graphical overlays of covariate densities between treatment groups)
    - For propensity-score matching, explore the distribution of the propensity score across treatment groups
    - Choose your method based on balance, not treatment effects!
        - For best practices, do not even look at treatment effects before the matching method is chosen

Practical Steps in Matching

3. Estimate treatment effects
    - Simple difference-of-means tests between experimental groups
    - Regression-adjusted difference of means tests using pre-treatment variables as controls
    - Same types of graphical techniques used for difference-of-means and regressions
    - Comparisons to unmatched analyses
    - If possible, comparisons to an experimental benchmark

Practical Steps in Matching

4. Evaluate the selection-on-observables assumption with sensitivity analysis
    - Rosenbaum bounds are probably the most widely used test
    - Usually will take the form of asking: How bad will our inferences be if we left out an important covariate?

Summary

The goal of matching is to generate greater balance in the distribution of pretreatment potential confounders between treatment and control groups in a non-experimental setting

- A search for a subset or transformed set of data that resembles an experiment but is contained within an observational study
- Requires some version of the selection on observables and common support assumptions

Selection on observables (along with other assumptions) gives us causal identification if we can claim that we can account for all of the variables that affect both treatment assignment and outcomes

- Very demanding assumption
- No omitted variables assumption of linear regression is one form of the assumption
- Other methods like subclassification and matching require it, but do not rely on the parametric assumptions of linear regression

Common support

- Not required in linear regression
- Required in matching because it does not rely on the functional form assumptions of linear regression
- Depending on our chosen estimand, requires us to have comparable cases with respect to X in the treatment and control groups