Causal Inference in Experiments and Observational Studies

Han Zhang

Outline

Readings

Today's topics are drawn from:

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapters 2 - 3)

Readings

Today's topics are drawn from:

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapters 2 - 3)
  - MHE later in this class

Readings

Today's topics are drawn from:

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapters 2 - 3)
  - MHE later in this class
- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019. (Chapters 6 - 7)

Readings

Today's topics are drawn from:

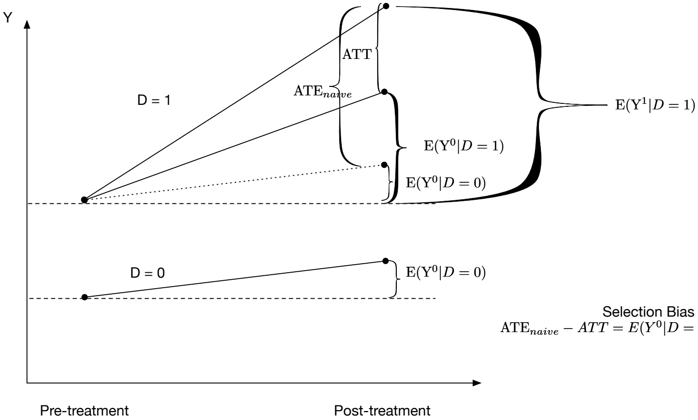
- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapters 2 - 3)
  - MHE later in this class
- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019. (Chapters 6 - 7)
- Proofs:

Readings

Today's topics are drawn from:

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapters 2 - 3)
  - MHE later in this class
- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019. (Chapters 6 - 7)
- Proofs:
  - Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015 (Chapter 6 - 7).

Observational Studies



Selection Bias =
$\text{ATE}_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in "treatment" and "control" outcomes.

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in "treatment" and "control" outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in "treatment" and "control" outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in “treatment” and “control” outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

- Selection bias becomes 0 and ATT is identified, if Y^0 is independent of D

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in "treatment" and "control" outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

- Selection bias becomes 0 and ATT is identified, if Y^0 is independent of D
  - in other words, if the treated units were not treated, their counterfactual outcome would be the same as that of the untreated users

Observational Studies

- $ATE_{naive} = E(Y^1|D = 1) - E(Y^0|D = 0)$, which is the mean differences in "treatment" and "control" outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D = 1) - E(Y^0|D = 0)$$

- Selection bias becomes 0 and ATT is identified, if Y^0 is independent of D
  - in other words, if the treated units were not treated, their counterfactual outcome would be the same as that of the untreated users
  - e.g., college-educated would have the same earning as non college-educated, if college-educated did not go to college

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in "treatment" and "control" outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

- Selection bias becomes 0 and ATT is identified, if Y^0 is independent of D
    - in other words, if the treated units were not treated, their counterfactual outcome would be the same as that of the untreated users
    - e.g., college-educated would have the same earning as non college-educated, if college-educated did not go to college
- If we further assume Y^1 is independent of D

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in "treatment" and "control" outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

- Selection bias becomes 0 and ATT is identified, if Y^0 is independent of D
  - in other words, if the treated units were not treated, their counterfactual outcome would be the same as that of the untreated users
  - e.g., college-educated would have the same earning as non college-educated, if college-educated did not go to college
- If we further assume Y^1 is independent of D
  - $ATT = ATE$

Design-based causal inference

- If you can, think and perform real randomized experiment

Design-based causal inference

- If you can, think and perform real randomized experiment
- If you cannot, try to approximate an experiment by adding assumptions

Design-based causal inference

- If you can, think and perform real randomized experiment
- If you cannot, try to approximate an experiment by adding assumptions
  - A better study design uses assumption that makes your study more like an experiment

Design-based causal inference

- If you can, think and perform real randomized experiment
- If you cannot, try to approximate an experiment by adding assumptions
  - A better study design uses assumption that makes your study more like an experiment
- examples: natural experiment, matching, DID, modern IV, RD

Design-based causal inference

- If you can, think and perform real randomized experiment
- If you cannot, try to approximate an experiment by adding assumptions
  - A better study design uses assumption that makes your study more like an experiment
- examples: natural experiment, matching, DID, modern IV, RD
- Rule of thumb: the gold-standard is always randomized controlled experiment

Model-based causal inference

- start from a regression and gradually add assumptions to the regression model (e.g., assumptions about endogenous or exogenous regressor)

Model-based causal inference

- start from a regression and gradually add assumptions to the regression model (e.g., assumptions about endogenous or exogenous regressor)
- example: traditional IV, fixed effects

Model-based causal inference

- start from a regression and gradually add assumptions to the regression model (e.g., assumptions about endogenous or exogenous regressor)
- example: traditional IV, fixed effects
- Historically people are more familiar with model-based causal inference

Model-based causal inference

- start from a regression and gradually add assumptions to the regression model (e.g., assumptions about endogenous or exogenous regressor)
- example: traditional IV, fixed effects

- Historically people are more familiar with model-based causal inference
- The trend is leaning toward design-based causal inference

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random
- Does political leaders (Presidents, Prime Ministers, etc) matter for regime type?

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random
- Does political leaders (Presidents, Prime Ministers, etc) matter for regime type?
- The Neyman-Rubin model suggests that we need to manipulate the treatment variable, and compare counterfactual outcomes

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random
- Does political leaders (Presidents, Prime Ministers, etc) matter for regime type?
- The Neyman-Rubin model suggests that we need to manipulate the treatment variable, and compare counterfactual outcomes
- A better way to ask the question: do leadership changes lead to changes in regime type?

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random
- Does political leaders (Presidents, Prime Ministers, etc) matter for regime type?
- The Neyman-Rubin model suggests that we need to manipulate the treatment variable, and compare counterfactual outcomes
- A better way to ask the question: do leadership changes lead to changes in regime type?
    - Problem: we do not know all factors that determine leadership changes

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random
- Does political leaders (Presidents, Prime Ministers, etc) matter for regime type?
- The Neyman-Rubin model suggests that we need to manipulate the treatment variable, and compare counterfactual outcomes
- A better way to ask the question: do leadership changes lead to changes in regime type?
    - Problem: we do not know all factors that determine leadership changes
        - A non-exhaustive lists include economic growth itself, leadership personality, geospatial conditions, etc.

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:
  - failed assassinations are a control group for successful assassinations

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:
  - failed assassinations are a control group for successful assassinations
  - e.g., compare the assassination of JFK to the assassination attempt on Ronald Reagan. Bullet killed JFK but missed Reagans heart by inches.

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:
  - failed assassinations are a control group for successful assassinations
  - e.g., compare the assassination of JFK to the assassination attempt on Ronald Reagan. Bullet killed JFK but missed Reagans heart by inches.
- Outcome is the change of regime type after the leadership transition

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:
  - failed assassinations are a control group for successful assassinations
  - e.g., compare the assassination of JFK to the assassination attempt on Ronald Reagan. Bullet killed JFK but missed Reagans heart by inches.
- Outcome is the change of regime type after the leadership transition
  - binary (democratic vs autocratic)

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:
  - failed assassinations are a control group for successful assassinations
  - e.g., compare the assassination of JFK to the assassination attempt on Ronald Reagan. Bullet killed JFK but missed Reagans heart by inches.
- Outcome is the change of regime type after the leadership transition
  - binary (democratic vs autocratic)
  - POLITY Score (democracy scores)

Estimating causal effects in natural experiment

- Because of the as-if-random assumption, observed and unobserved conditions, which may be related to treatment assignment, are randomized by treatment and control groups

$$Y = \alpha + \rho D + \epsilon$$

Estimating causal effects in natural experiment

- Because of the as-if-random assumption, observed and unobserved conditions, which may be related to treatment assignment, are randomized by treatment and control groups
- We can use Neyman estimator (compare changes in regime type in successful assassinations vs unsuccessful assassinations)

$$Y = \alpha + \rho D + \epsilon$$

Estimating causal effects in natural experiment

- Because of the as-if-random assumption, observed and unobserved conditions, which may be related to treatment assignment, are randomized by treatment and control groups
- We can use Neyman estimator (compare changes in regime type in successful assassinations vs unsuccessful assassinations)
- We can also use regression estimator: $\hat{\rho}$ estimates ATE

$$Y = \alpha + \rho D + \epsilon$$

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)
- Because we assume the natural experiments are as-if-random, it is important to check pre-treatment balance to see whether treatment/control group are actually balanced

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)
- Because we assume the natural experiments are as-if-random, it is important to check pre-treatment balance to see whether treatment/control group are actually balanced
- If the assumptions is true, we would expect that other covariates are indeed balanced across treatment and control groups

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)
- Because we assume the natural experiments are as-if-random, it is important to check pre-treatment balance to see whether treatment/control group are actually balanced
- If the assumptions is true, we would expect that other covariates are indeed balanced across treatment and control groups
    - That is, the other variables should be independent of D, if the assumption is correct

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)
- Because we assume the natural experiments are as-if-random, it is important to check pre-treatment balance to see whether treatment/control group are actually balanced
- If the assumptions is true, we would expect that other covariates are indeed balanced across treatment and control groups
    - That is, the other variables should be independent of D, if the assumption is correct
- But the reverse is not true: covariates are balanced $\not\Rightarrow$ exogenous shocks are truly random

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)
- Because we assume the natural experiments are as-if-random, it is important to check pre-treatment balance to see whether treatment/control group are actually balanced
- If the assumptions is true, we would expect that other covariates are indeed balanced across treatment and control groups
    - That is, the other variables should be independent of D, if the assumption is correct
- But the reverse is not true: covariates are balanced $\nRightarrow$ exogenous shocks are truly random
- There is no substitute for a good research design (here, exogenous shocks)

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)
- Say we think that within assassination attempts, those using guns are more likely to succeed than those using bombs

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)
- Say we think that within assassination attempts, those using guns are more likely to succeed than those using bombs
  - Assassinators also know this! So their decisions are conditioned on weapons

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)
- Say we think that within assassination attempts, those using guns are more likely to succeed than those using bombs
  - Assassinators also know this! So their decisions are conditioned on weapons
- We should add weapon types as additional controls

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)
- Say we think that within assassination attempts, those using guns are more likely to succeed than those using bombs
    - Assassinators also know this! So their decisions are conditioned on weapons
- We should add weapon types as additional controls

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

- In analyzing natural experiments, it is recommended to use all other observed control variables you think are relevant to your outcome Y (no post-treatment controls, of course)

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)
- Say we think that within assassination attempts, those using guns are more likely to succeed than those using bombs
  - Assassinators also know this! So their decisions are conditioned on weapons
- We should add weapon types as additional controls

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

- In analyzing natural experiments, it is recommended to use all other observed control variables you think are relevant to your outcome Y (no post-treatment controls, of course)
  - Another differences from the randomized experiments

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways
    - Any assassinatons (including those still in secret planning?)

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways
  - Any assassinatons (including those still in secret planning?)
  - Any failed assassinatons planning that made to newspapers

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways
    - Any assassinatons (including those still in secret planning?)
    - Any failed assassinatons planning that made to newspapers
    - Assassinations whose weapons were already discharged (serious attempts)

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways
    - Any assassinatons (including those still in secret planning?)
    - Any failed assassinatons planning that made to newspapers
    - Assassinations whose weapons were already discharged (serious attempts)
- Be clear what you are comparing with

Many different ways to define control groups

- In natural experiments, there are often multiple ways to choose control groups
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways
  - Any assassinatons (including those still in secret planning?)
  - Any failed assassinatons planning that made to newspapers
  - Assassinations whose weapons were already discharged (serious attempts)
- Be clear what you are comparing with
- In randomized controlled experiments, treatment and control groups were chosen with clear standard so there is no such problem

Selection on Observables/ Ignorability

- To identify causal effects without random assignment, we have to add strong assumptions (analogous to MAR)

Definition (Selection on observable, or ignorability, or exogeneity)

Selection on Observables/ Ignorability

- To identify causal effects without random assignment, we have to add strong assumptions (analogous to MAR)

Definition (Selection on observable, or ignorability, or exogeneity)

- $Y_i^0, Y_i^1 \perp\!\!\!\perp D_i | X_i$ (Potential outcome is independent of treatment assignment, condition on observed X_i)

Selection on Observables/ Ignorability

- To identify causal effects without random assignment, we have to add strong assumptions (analogous to MAR)

Definition (Selection on observable, or ignorability, or exogeneity)

- $Y_i^0, Y_i^1 \perp\!\!\!\perp D_i | X_i$ (Potential outcome is independent of treatment assignment, condition on observed X_i)
- $P(D = 1) > 0$ (non-zero treatment probability)

Selection on Observables/ Ignorability

- To identify causal effects without random assignment, we have to add strong assumptions (analogous to MAR)

Definition (Selection on observable, or ignorability, or exogeneity)

- $Y_i^0, Y_i^1 \perp\!\!\!\perp D_i | X_i$ (Potential outcome is independent of treatment assignment, condition on observed X_i)
- $P(D = 1) > 0$ (non-zero treatment probability)

- Note that randomized experiments automatically satisfy this assumption

Ignorability

- There may always be unknown/unmeasured factors that contribute to treatment assignments

Ignorability

- There may always be unknown/unmeasured factors that contribute to treatment assignments
- That is, we are selecting on unobservables

Ignorability

- There may always be unknown/unmeasured factors that contribute to treatment assignments
- That is, we are selecting on unobservables
- or we have non-ignorability

Ignorability

- There may always be unknown/unmeasured factors that contribute to treatment assignments
- That is, we are selecting on unobservables
- or we have non-ignorability
- or we have omitted variable bias

Ignorability

- There may always be unknown/unmeasured factors that contribute to treatment assignments
- That is, we are selecting on unobservables
- or we have non-ignorability
- or we have omitted variable bias
- or, there is selection bias due to unobservables

Regression estimator

- If ignorability assumption is true, and you assume the effect of treatment is constant on Y, we can use regression to estimate causal effects:

$$Y = \alpha + \rho D + \gamma X + \eta$$

Regression estimator

- If ignorability assumption is true, and you assume the effect of treatment is constant on Y, we can use regression to estimate causal effects:

$$Y = \alpha + \rho D + \gamma X + \eta$$

- Estimate of ATE is the regression coefficient ρ

Regression estimator

- If ignorability assumption is true, and you assume the effect of treatment is constant on Y, we can use regression to estimate causal effects:

$$Y = \alpha + \rho D + \gamma X + \eta$$

- Estimate of ATE is the regression coefficient ρ
- Here we want as many X as possible

Regression estimator

- If ignorability assumption is true, and you assume the effect of treatment is constant on Y, we can use regression to estimate causal effects:

$$Y = \alpha + \rho D + \gamma X + \eta$$

- Estimate of ATE is the regression coefficient ρ
- Here we want as many X as possible
  - We are making assumptions that potential outcome is independent of D, conditional on X

Regression estimator

- If ignorability assumption is true, and you assume the effect of treatment is constant on Y, we can use regression to estimate causal effects:

$$Y = \alpha + \rho D + \gamma X + \eta$$

- Estimate of ATE is the regression coefficient ρ
- Here we want as many X as possible
  - We are making assumptions that potential outcome is independent of D, conditional on X
  - More X increases the possibility that you do not missed anything important confounders

Regression as Imputation

- If ignorability is true, the regression estimator of ATE is implicitly making counterfactual imputation using linear regression.

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	2	1	1	7
2	5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	3	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	10	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	2	1	5	2

Regression as Imputation

- If ignorability is true, the regression estimator of ATE is implicitly making counterfactual imputation using linear regression.
- Hence the similar form of ignorability and MAR assumption

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	2	1	1	7
2	5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	3	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	10	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	2	1	5	2

Regression as Imputation

- If ignorability is true, the regression estimator of ATE is implicitly making counterfactual imputation using linear regression.
- Hence the similar form of ignorability and MAR assumption

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

- Run a regression as $Y = \beta_0 + \beta_1 D_i + \beta_2 X_{[1]i} + \beta_3 X_{[2]i}$, and impute counterfactual outcome using the linear regression:

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	2	1	1	7
2	5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	3	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	10	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	2	1	5	2

Ignorability: example

- Ignorability assumption basically says that we can add any confounder that may impact treatment assignment and the outcome variable

Ignorability: example

- Ignorability assumption basically says that we can add any confounder that may impact treatment assignment and the outcome variable
  - You can add them -> you can observe them

Ignorability: example

- Ignorability assumption basically says that we can add any confounder that may impact treatment assignment and the outcome variable
  - You can add them -> you can observe them
- This is usually a very strong assumption

Ignorability: example

- Ignorability assumption basically says that we can add any confounder that may impact treatment assignment and the outcome variable
    - You can add them -> you can observe them
- This is usually a very strong assumption
- But in some occasional cases, this assumption is likely to be true

Ignorability: example

- Ignorability assumption basically says that we can add any confounder that may impact treatment assignment and the outcome variable
  - You can add them -> you can observe them
- This is usually a very strong assumption
- But in some occasional cases, this assumption is likely to be true
- Jens Hainmueller and Dominik Hangartner, *Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination*, American Political Science Review **107** (2013), no. 01, 159–187

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
    - if you ask people whether they will tend to discriminate against certain types of immigrants

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
  - if you ask people whether they will tend to discriminate against certain types of immigrants
  - even if a people discriminate, he/she probably will not tell you

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
  - if you ask people whether they will tend to discriminate against certain types of immigrants
  - even if a people discriminate, he/she probably will not tell you
  - this latent factor is unobserved

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
  - if you ask people whether they will tend to discriminate against certain types of immigrants
  - even if a people discriminate, he/she probably will not tell you
  - this latent factor is unobserved
  - Other unobserved confounders you can think of?

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
  - if you ask people whether they will tend to discriminate against certain types of immigrants
  - even if a people discriminate, he/she probably will not tell you
  - this latent factor is unobserved
  - Other unobserved confounders you can think of?
    - networks or living environments

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
  - if you ask people whether they will tend to discriminate against certain types of immigrants
  - even if a people discriminate, he/she probably will not tell you
  - this latent factor is unobserved
  - Other unobserved confounders you can think of?
    - networks or living environments
    - if a voter has many immigrant friends, he/she probably is less likely to discriminate

Ignorability: example

- Research question: what types of immigrants are more likely to be discriminated?
- Empirical problem:
    - if you ask people whether they will tend to discriminate against certain types of immigrants
    - even if a people discriminate, he/she probably will not tell you
    - this latent factor is unobserved
    - Other unobserved confounders you can think of?
        - networks or living environments
        - if a voter has many immigrant friends, he/she probably is less likely to discriminate
        - but you failed to measure this social network information in your survey

Ignorability: example

- Some Municipalities in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not

Ignorability: example

- Some Municipalities in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not
- Unless one knows the immigrant in person, he is most likely to make a decision based on "voting leaflets summarizing the applicant characteristics were sent to all citizens usually about two to six weeks before"

Ignorability: example

- Some Municipalities in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not
- Unless one knows the immigrant in person, he is most likely to make a decision based on "voting leaflets summarizing the applicant characteristics were sent to all citizens usually about two to six weeks before"
  - Votes are secret; people can reveal their real preferences

Ignorability: example

- Some Municipalities in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not
- Unless one knows the immigrant in person, he is most likely to make a decision based on "voting leaflets summarizing the applicant characteristics were sent to all citizens usually about two to six weeks before"
  - Votes are secret; people can reveal their real preferences
  - Votes are real; people are more likely to express their behavioral preferences

Ignorability: example

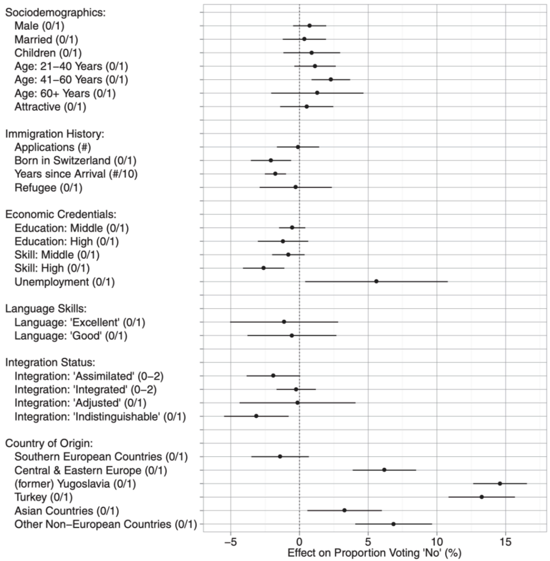
- Some Municipalities in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not
- Unless one knows the immigrant in person, he is most likely to make a decision based on "voting leaflets summarizing the applicant characteristics were sent to all citizens usually about two to six weeks before"
    - Votes are secret; people can reveal their real preferences
    - Votes are real; people are more likely to express their behavioral preferences
    - Most people don't know the applicant; they only judge from the leaflet that gives a description of the applicant

Ignorability: example

- Some Municipalities in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not
- Unless one knows the immigrant in person, he is most likely to make a decision based on "voting leaflets summarizing the applicant characteristics were sent to all citizens usually about two to six weeks before"
  - Votes are secret; people can reveal their real preferences
  - Votes are real; people are more likely to express their behavioral preferences
  - Most people don't know the applicant; they only judge from the leaflet that gives a description of the applicant
- These features ensures minimum measurement error, and also ignorability

Ignorability: example

FIGURE 2. Effect of Applicant Characteristics on Opposition to Naturalization Requests



Ignorability

- What if some people know the applicants in person?

Ignorability

- What if some people know the applicants in person?
- That is, they have private information other than that listed in the leaflets

Matching estimator using original data

- With ignorability, we can also use matching estimator

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units

Matching estimator using original data

- With ignorability, we can also use matching estimator
    - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
    - Find the j in the control group, whose X_j is the closest X_i
        - j is called the matched unit of i
    - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units
- We can do the similar things for control units:

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units
- We can do the similar things for control units:
  - For unit i in the control ($D_i = 1$), we want to impute its Y_i^1

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units
- We can do the similar things for control units:
  - For unit i in the control ($D_i = 1$), we want to impute its Y_i^1
    - Find the j in the control group, whose X_j is the closest X_i

Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units
- We can do the similar things for control units:
  - For unit i in the control ($D_i = 1$), we want to impute its Y_i^1
    - Find the j in the control group, whose X_j is the closest X_i
    - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i

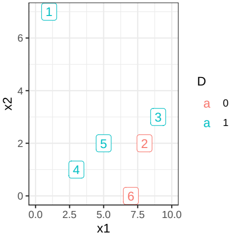
Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units
- We can do the similar things for control units:
  - For unit i in the control ($D_i = 1$), we want to impute its Y_i^1
    - Find the j in the control group, whose X_j is the closest X_i
    - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- Then we can estimate ATE as the the difference between the mean of Y_i^1 and Y_i^0 for all units

Matching estimator using original data

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

(2)

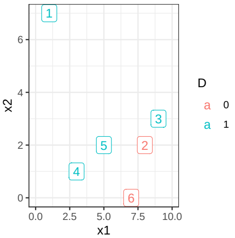


- Unit 3 is treated; it is closest to unit 2; unit 2 is the matched unit of unit 3

Matching estimator using original data

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

(2)



- Unit 3 is treated; it is closest to unit 2; unit 2 is the matched unit of unit 3
- $Y_3^0 \leftarrow Y_2^0 = 5$

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.

$$P(D = 1|X) \tag{3}$$

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again

$$P(D = 1|X) \tag{3}$$

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

- Rosenbaum and Rubin proves (propensity score theorem)

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

- Rosenbaum and Rubin proves (propensity score theorem)
  - If you have the correct propensity score

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

- Rosenbaum and Rubin proves (propensity score theorem)
  - If you have the correct propensity score
  - Then conditioning on X is equivalent to conditioning on $P(D = 1|X)$

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
    - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

- Rosenbaum and Rubin proves (propensity score theorem)
    - If you have the correct propensity score
    - Then conditioning on X is equivalent to conditioning on $P(D = 1|X)$
- Treatment propensity score provides a single-number summary of treatment probability

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

- Rosenbaum and Rubin proves (propensity score theorem)
  - If you have the correct propensity score
  - Then conditioning on X is equivalent to conditioning on $P(D = 1|X)$
- Treatment propensity score provides a single-number summary of treatment probability
- We should match treatment and control users with similar treatment propensity scores

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{3}$$

- Rosenbaum and Rubin proves (propensity score theorem)
  - If you have the correct propensity score
  - Then conditioning on X is equivalent to conditioning on $P(D = 1|X)$
- Treatment propensity score provides a single-number summary of treatment probability
- We should match treatment and control users with similar treatment propensity scores
  - This is usually called propensity score matching

Matching estimator using Propensity Score

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$	$p(D_i = 1 \vert X_i)$
1	?	2	1	1	7	0.33
2	5	?	0	8	2	0.14
3	?	3	1	10	3	0.73
4	?	10	1	3	1	0.35
5	?	2	1	5	2	0.78
6	0	?	0	7	0	0.70

(4)

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$	$p(D_i = 1 \vert X_i)$
1	5	2	1	1	7	0.33
2	5	2	0	8	2	0.14
3	0	3	1	10	3	0.73
4	5	10	1	3	1	0.35
5	0	2	1	5	2	0.78
6	0	3	0	7	0	0.70

(5)

- Estimated ATE is $7/6$

Regression vs Matching

- Regression estimates of ATE in general will be different from matching estimates of ATE (MHE 3.3)

$$A\hat{T}E_{ols} = \hat{\rho} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot A\hat{T}E_x$$

$$\omega(x) = P(X = x) \cdot P(D = 1|X = x) \cdot (1 - P(D = 1|X = x))$$

$$A\hat{T}E_{matching} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot A\hat{T}E_x$$

$$\omega(x) = P(X = x)$$

Regression vs Matching

- Regression estimates of ATE in general will be different from matching estimates of ATE (MHE 3.3)

$$\hat{ATE}_{ols} = \hat{\rho} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x) \cdot P(D = 1|X = x) \cdot (1 - P(D = 1|X = x))$$

$$\hat{ATE}_{matching} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x)$$

- Regression estimator is generally inconsistent, while matching estimator approximates Neyman estimator and is consistent

Regression vs Matching

- Regression estimates of ATE in general will be different from matching estimates of ATE (MHE 3.3)

$$\hat{ATE}_{ols} = \hat{\rho} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x) \cdot P(D = 1|X = x) \cdot (1 - P(D = 1|X = x))$$

$$\hat{ATE}_{matching} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x)$$

- Regression estimator is generally inconsistent, while matching estimator approximates Neyman estimator and is consistent
- Regression give more weights to data whose propensity score is close to 0.5

Regression vs Matching

- Regression estimates of ATE in general will be different from matching estimates of ATE (MHE 3.3)

$$\hat{ATE}_{ols} = \hat{\rho} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x) \cdot P(D = 1|X = x) \cdot (1 - P(D = 1|X = x))$$

$$\hat{ATE}_{matching} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x)$$

- Regression estimator is generally inconsistent, while matching estimator approximates Neyman estimator and is consistent
- Regression give more weights to data whose propensity score is close to 0.5
  - These are observations whose treatment status cannot be predicted well by X, thus could have omitted variable bias

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
  - Pros: consistent

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
  - Pros: consistent
    - also reduce model dependency; make your model more robust under misspecification

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
  - Pros: consistent
    - also reduce model dependency; make your model more robust under misspecification
  - Cons:

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
  - Pros: consistent
    - also reduce model dependency; make your model more robust under misspecification
  - Cons:
    - Do you have the correct propensity score? (if you use propensity score matching)

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
  - Pros: consistent
    - also reduce model dependency; make your model more robust under misspecification
  - Cons:
    - Do you have the correct propensity score? (if you use propensity score matching)
    - confidence intervals are hard to calculate analytically

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
    - Pros: consistent
        - also reduce model dependency; make your model more robust under misspecification
    - Cons:
        - Do you have the correct propensity score? (if you use propensity score matching)
        - confidence intervals are hard to calculate analytically
        - The first theoretical work is by Abadie and Imbens, 2006, Econometrica.

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
    - Pros: consistent
        - also reduce model dependency; make your model more robust under misspecification
    - Cons:
        - Do you have the correct propensity score? (if you use propensity score matching)
        - confidence intervals are hard to calculate analytically
        - The first theoretical work is by Abadie and Imbens, 2006, Econometrica.
- Regression

Regression vs Matching

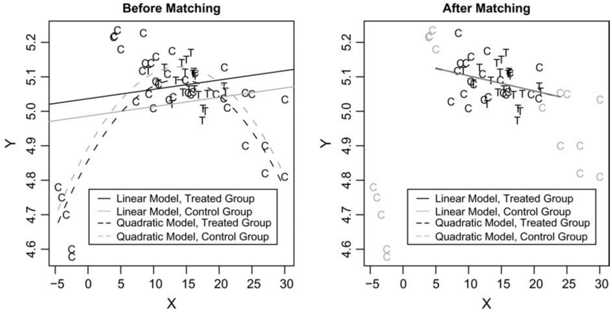
- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
    - Pros: consistent
        - also reduce model dependency; make your model more robust under misspecification
    - Cons:
        - Do you have the correct propensity score? (if you use propensity score matching)
        - confidence intervals are hard to calculate analytically
        - The first theoretical work is by Abadie and Imbens, 2006, Econometrica.
- Regression
    - Pros: easier to work with (especially the standard error)

Regression vs Matching

- Regression and matching estimates of causal effect share the same assumption: ignorability
- Usually matching is considered a superior methods than simple regression, because matching explicitly approaches experimental ideal
- Matching
    - Pros: consistent
        - also reduce model dependency; make your model more robust under misspecification
    - Cons:
        - Do you have the correct propensity score? (if you use propensity score matching)
        - confidence intervals are hard to calculate analytically
        - The first theoretical work is by Abadie and Imbens, 2006, Econometrica.
- Regression
    - Pros: easier to work with (especially the standard error)
    - Cons: inconsistent; more weights to data whose treatment status we cannot predict

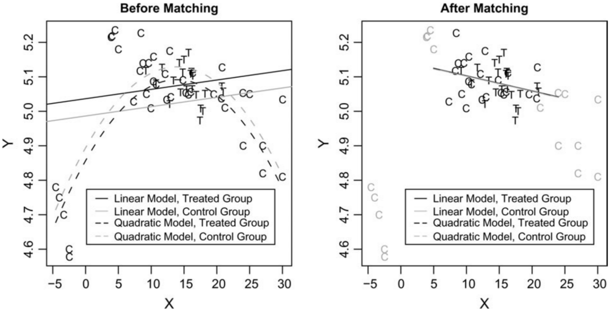
Regression vs matching

- Why approaching experimental ideal is important?



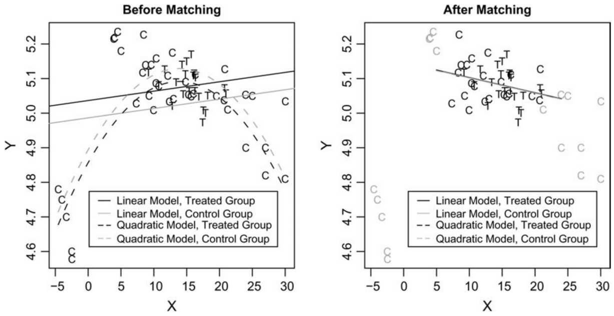
Regression vs matching

- Why approaching experimental ideal is important?
- It makes your model more robust to misspecification



Regression vs matching

- Why approaching experimental ideal is important?
- It makes your model more robust to misspecification
- Ho, King, Imai and Stuart, “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference”, *Political Analysis*, 2007



Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
    - Outcome: earning after 1/2 years

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
  - Outcome: earning after 1/2 years
  - the (randomly selected) treatment and control groups consist of 297 and 425 such workers, respectively.

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
    - Outcome: earning after 1/2 years
    - the (randomly selected) treatment and control groups consist of 297 and 425 such workers, respectively.
- CPS: survey data that is much larger in size, but with similar variables

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
  - Outcome: earning after 1/2 years
  - the (randomly selected) treatment and control groups consist of 297 and 425 such workers, respectively.
- CPS: survey data that is much larger in size, but with similar variables
  - clearly no one in CPS received the work training so they served as a non-experimental control group

Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
    - Outcome: earning after 1/2 years
    - the (randomly selected) treatment and control groups consist of 297 and 425 such workers, respectively.
- CPS: survey data that is much larger in size, but with similar variables
    - clearly no one in CPS received the work training so they served as a non-experimental control group
    - That is, construct a data as (NSW(treated), CPS); CPS replaced NSW control units

Regression vs Matching

- What the authors did:

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)

Regression vs Matching

- What the authors did:
  - Experiments: raw ATE vs. regression estimated ATE based on [NSW(treated), NSW(control)]

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)

Regression vs Matching

- What the authors did:
  - Experiments: raw ATE vs. regression estimated ATE based on [NSW(treated), NSW(control)]
  - CPS: ATE_{naive} vs. regression adjusted ATE_{naive} based on [NSW(treated), CPS)]

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)

Regression vs Matching

- What the authors did:
  - Experiments: raw ATE vs. regression estimated ATE based on [NSW(treated), NSW(control)]
  - CPS: ATE_{naive} vs. regression adjusted ATE_{naive} based on [NSW(treated), CPS)]
  - Matching: match each unit in NSW treatment group with a control in CPS, then calculate ATE_{naive} and regress adjusted ATE_{naive} based on [NSW(treated), CPS(matched)]

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good
- In particular, never condition on post-treatment X (Judea Pearl)

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good
- In particular, never condition on post-treatment X (Judea Pearl)
    - This is like cheating: if X is determined by your D, X would not be randomized even under random assignment of D, which recreates selection biases

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good
- In particular, never condition on post-treatment X (Judea Pearl)
    - This is like cheating: if X is determined by your D, X would not be randomized even under random assignment of D, which recreates selection biases
- Example:

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good
- In particular, never condition on post-treatment X (Judea Pearl)
  - This is like cheating: if X is determined by your D, X would not be randomized even under random assignment of D, which recreates selection biases
- Example:
  - Effect of college education on future earning

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good
- In particular, never condition on post-treatment X (Judea Pearl)
    - This is like cheating: if X is determined by your D, X would not be randomized even under random assignment of D, which recreates selection biases
- Example:
    - Effect of college education on future earning
    - And it will be dangerous to add *occupation* as a control variable, since occupation may be the result of treatment

Matching + Regression

- Matching and Regression are not mutually exclusive

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units
  - Basic idea: assume you have 100 treated units and 400 control units

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units
  - Basic idea: assume you have 100 treated units and 400 control units
  - Instead of running a regression with all of them

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units
    - Basic idea: assume you have 100 treated units and 400 control units
    - Instead of running a regression with all of them
    - Find the 100 control units that are closet to 100 treated units

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units
    - Basic idea: assume you have 100 treated units and 400 control units
    - Instead of running a regression with all of them
    - Find the 100 control units that are closet to 100 treated units
    - And then run regressions based on 200 units

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units
  - Basic idea: assume you have 100 treated units and 400 control units
  - Instead of running a regression with all of them
  - Find the 100 control units that are closet to 100 treated units
  - And then run regressions based on 200 units
  - This gives you the ATT estimates

Ignorability vs. Non-ignorability

- Randomized experiments:

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:

Ignorability vs. Non-ignorability

- Randomized experiments:
    - Automatically satisfies ignorability by randomization
- Ignorability:
    - Very strong assumption; unrealistic in most settings

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:
  - Very strong assumption; unrealistic in most settings
  - In this case, people just control a bunch of things but do not claim that they find any causal effect

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:
  - Very strong assumption; unrealistic in most settings
  - In this case, people just control a bunch of things but do not claim that they find any causal effect
  - Or avoiding using causal language; just say X predicts Y of X is associated of Y

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:
  - Very strong assumption; unrealistic in most settings
  - In this case, people just control a bunch of things but do not claim that they find any causal effect
  - Or avoiding using causal language; just say X predicts Y of X is associated of Y
- The third approach:

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:
  - Very strong assumption; unrealistic in most settings
  - In this case, people just control a bunch of things but do not claim that they find any causal effect
  - Or avoiding using causal language; just say X predicts Y of X is associated of Y
- The third approach:
  - Do not assume ignorability (or assume non-ignorability)

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:
  - Very strong assumption; unrealistic in most settings
  - In this case, people just control a bunch of things but do not claim that they find any causal effect
  - Or avoiding using causal language; just say X predicts Y of X is associated of Y
- The third approach:
  - Do not assume ignorability (or assume non-ignorability)
  - Essentially admit that we cannot control everything; there are some unobserved variables we cannot control for

Econometric tools in working with non-ignorability

- Fixed effect and diff-in-diff

Econometric tools in working with non-ignorability

- Fixed effect and diff-in-diff
- IV

Econometric tools in working with non-ignorability

- Fixed effect and diff-in-diff
- IV
- RD