FLS 6441 - Methods III: Explanation and Causation

Week 12 - Review & Frontiers

Jonathan Phillips

June 2019

Section 1

Review

Classification of Research Designs

- Correlation is not causation
  - And regresssion is just fancy correlation
- So how do we provide evidence of causation?

Classification of Research Designs

		Independence of Treatment Assignment	Researcher Controls Treatment Assignment?
Controlled Experiments	Field Experiments	✓	✓
	Survey and Lab Experiments	✓	✓
Natural Experiments	Natural Experiments	✓	
	Instrumental Variables	✓	
	Discontinuities	✓	
Observational Studies	Difference-in-Differences		
	Controlling for Confounding		
	Matching		
	Comparative Cases and Process Tracing		

Definitions

1. Potential Outcomes

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance
7. Hawthorne Effects

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance
7. Hawthorne Effects
8. Time-invariant confounder

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance
7. Hawthorne Effects
8. Time-invariant confounder
9. Exclusion Restriction

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance
7. Hawthorne Effects
8. Time-invariant confounder
9. Exclusion Restriction
10. Back-door path

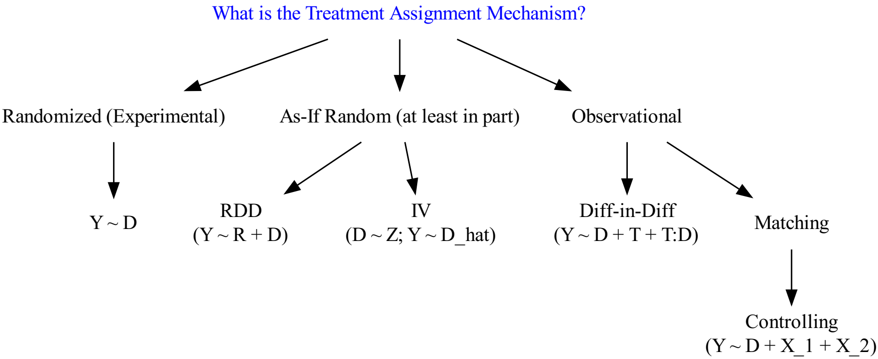
Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance
7. Hawthorne Effects
8. Time-invariant confounder
9. Exclusion Restriction
10. Back-door path
11. SUTVA

Definitions

1. Potential Outcomes
2. Treatment Assignment Mechanism
3. Independence of Potential Outcomes from Treatment
4. Average Treatment Effect
5. Local Average Treatment Effect
6. Non-compliance
7. Hawthorne Effects
8. Time-invariant confounder
9. Exclusion Restriction
10. Back-door path
11. SUTVA
12. Overlap in sample characteristics

Choosing a Method



Choosing a Method

- How do we decide which causal inference strategy to use?

Choosing a Method

- How do we decide which causal inference strategy to use?
  1. What is the treatment assignment mechanism?

Choosing a Method

- How do we decide which causal inference strategy to use?
  1. What is the treatment assignment mechanism?
     - Randomized: field experiment
     - As-if random: natural experiment
     - Messy: Observational study
  2. Where is the as-if variation in treatment?

Choosing a Method

- How do we decide which causal inference strategy to use?
  1. What is the treatment assignment mechanism?
     - Randomized: field experiment
     - As-if random: natural experiment
     - Messy: Observational study
  2. Where is the as-if variation in treatment?
     - Across time: Diff-in-diff
     - Across threshold: RDD
     - Before treatment: IV
  3. How many units can we get accurate measures for?

Choosing a Method

- How do we decide which causal inference strategy to use?
  1. What is the treatment assignment mechanism?
     - Randomized: field experiment
     - As-if random: natural experiment
     - Messy: Observational study
  2. Where is the as-if variation in treatment?
     - Across time: Diff-in-diff
     - Across threshold: RDD
     - Before treatment: IV
  3. How many units can we get accurate measures for?
     - One: Process tracing
     - Small-N: Comparative Case Studies
     - Large-N: Controls/Matching
  4. Are the assumptions met?

Choosing a Method

- How do we decide which causal inference strategy to use?
  1. What is the treatment assignment mechanism?
     - Randomized: field experiment
     - As-if random: natural experiment
     - Messy: Observational study
  2. Where is the as-if variation in treatment?
     - Across time: Diff-in-diff
     - Across threshold: RDD
     - Before treatment: IV
  3. How many units can we get accurate measures for?
     - One: Process tracing
     - Small-N: Comparative Case Studies
     - Large-N: Controls/Matching
  4. Are the assumptions met?
     - Parallel trends, no sorting, balance...

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
   - Online survey experiment

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
   - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
    - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
    - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?
    - Instrument peasant revolt with serfdom

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
    - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
    - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?
    - Instrument peasant revolt with serfdom
4. Do women govern differently from men?

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
   - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?
   - Instrument peasant revolt with serfdom
4. Do women govern differently from men?
   - Regression discontinuity in close elections in Brazil

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
   - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?
   - Instrument peasant revolt with serfdom
4. Do women govern differently from men?
   - Regression discontinuity in close elections in Brazil
5. Do US political contact campaigns change voters' choices?

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
   - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?
   - Instrument peasant revolt with serfdom
4. Do women govern differently from men?
   - Regression discontinuity in close elections in Brazil
5. Do US political contact campaigns change voters' choices?
   - Field experiment

Choosing a Method

1. Has experience with Obamacare increased electoral turnout?
   - Difference-in-differences between states that did/did not expand Obamacare
2. Can playing a video game as a Roma character reduce anti-Roma prejudice in Hungary?
   - Online survey experiment
3. Does peasant revolt in 19th century Russia lead to less representative local government?
   - Instrument peasant revolt with serfdom
4. Do women govern differently from men?
   - Regression discontinuity in close elections in Brazil
5. Do US political contact campaigns change voters' choices?
   - Field experiment

The Role of Theory

- To avoid data mining: We have to test plausible, relevant theories

The Role of Theory

- To avoid data mining: We have to test plausible, relevant theories
- To tell us which experiments and research designs to run

The Role of Theory

- To avoid data mining: We have to test plausible, relevant theories
- To tell us which experiments and research designs to run
- To justify assumptions (exclusion restriction, confounders)

The Role of Theory

- To avoid data mining: We have to test plausible, relevant theories
- To tell us which experiments and research designs to run
- To justify assumptions (exclusion restriction, confounders)
- To help us interpret what we have learned

The Role of Qualitative Evidence

- Vital for identifying natural experiments

The Role of Qualitative Evidence

- Vital for identifying natural experiments
- To validate assumptions (no sorting, randomization worked, SUTVA)

The Role of Qualitative Evidence

- Vital for identifying natural experiments
- To validate assumptions (no sorting, randomization worked, SUTVA)
- To understand specific analysis requirements, eg. non-compliance, clustering

The Role of Qualitative Evidence

- Vital for identifying natural experiments
- To validate assumptions (no sorting, randomization worked, SUTVA)
- To understand specific analysis requirements, eg. non-compliance, clustering
- For Process Tracing: Causal Process Observations

Limitations of Causal Methodologies

- Usually a trade-off between avoiding bias and generalizability

Limitations of Causal Methodologies

- Usually a trade-off between avoiding bias and generalizability
- Sure you have shown that D affects Y, but how?? The connection is still a black box!

Limitations of Causal Methodologies

- Usually a trade-off between avoiding bias and generalizability
- Sure you have shown that D affects Y, but how?? The connection is still a black box!
- Causal effects are probably highly heteregeneous - who cares about the average?

Limitations of Causal Methodologies

- Usually a trade-off between avoiding bias and generalizability
- Sure you have shown that D affects Y, but how?? The connection is still a black box!
- Causal effects are probably highly heteregeneous - who cares about the average?
- They only tell us about 'unusual' parts of the population (eg. RDD)

Limitations of Causal Methodologies

- Usually a trade-off between avoiding bias and generalizability
- Sure you have shown that D affects Y, but how?? The connection is still a black box!
- Causal effects are probably highly heteregeneous - who cares about the average?
- They only tell us about 'unusual' parts of the population (eg. RDD)
- Even if variable X has a causal effect, *how much* of the real world does it explain?

Limitations of Causal Methodologies

- Usually a trade-off between avoiding bias and generalizability
- Sure you have shown that D affects Y, but how?? The connection is still a black box!
- Causal effects are probably highly heteregeneous - who cares about the average?
- They only tell us about 'unusual' parts of the population (eg. RDD)
- Even if variable X has a causal effect, *how much* of the real world does it explain?
- Sometimes it's just not possible to show causation. That's OK!
  - We just need to recognize the evidence we have is not representative of everything that happens in the real world

Section 2

Frontiers

Frontiers of Strengthening Causal Arguments

- Writing a paper means sustaining a convincing argument

Frontiers of Strengthening Causal Arguments

- Writing a paper means sustaining a convincing argument
- Choosing and implementing an appropriate method is only the first step

Frontiers of Strengthening Causal Arguments

- Writing a paper means sustaining a convincing argument
- Choosing and implementing an appropriate method is only the first step
- We also need to show that our estimate is reliable and not a ’chance’ finding

Frontiers of Strengthening Causal Arguments

- Writing a paper means sustaining a convincing argument
- Choosing and implementing an appropriate method is only the first step
- We also need to show that our estimate is reliable and not a 'chance' finding
- More importantly, that it is evidence in support of **specific** theory

Frontiers of Strengthening Causal Arguments

- Writing a paper means sustaining a convincing argument
- Choosing and implementing an appropriate method is only the first step
- We also need to show that our estimate is reliable and not a ’chance’ finding
- More importantly, that it is evidence in support of **specific** theory
- You don’t want to publish a paper that someone contradicts next week!

Robustness Tests

- In general, we will trust our estimate more if it doesn't change even when we change our model
  - Not just direction and significance, but in the substantive effect size
- Alternative covariates/matching procedures

Robustness Tests

- In general, we will trust our estimate more if it doesn't change even when we change our model
  - Not just direction and significance, but in the substantive effect size
- Alternative covariates/matching procedures
- Alternative bandwidths/functional forms

Robustness Tests

- In general, we will trust our estimate more if it doesn't change even when we change our model
  - Not just direction and significance, but in the substantive effect size
- Alternative covariates/matching procedures
- Alternative bandwidths/functional forms
- Alternative (but conceptually equivalent) measures of key variables

Robustness Tests

- In general, we will trust our estimate more if it doesn't change even when we change our model
  - Not just direction and significance, but in the substantive effect size
- Alternative covariates/matching procedures
- Alternative bandwidths/functional forms
- Alternative (but conceptually equivalent) measures of key variables
- Alternative samples (dropping outliers etc.)

Robustness Tests

- In general, we will trust our estimate more if it doesn't change even when we change our model
  - Not just direction and significance, but in the substantive effect size
- Alternative covariates/matching procedures
- Alternative bandwidths/functional forms
- Alternative (but conceptually equivalent) measures of key variables
- Alternative samples (dropping outliers etc.)
- Various formal tests, but best to plot overlap of confidence intervals from many models

Sensitivity Analysis

- An alternative is to ask - quantitatively - how much do our results change when we alter the model or its assumptions?

Sensitivity Analysis

- An alternative is to ask - quantitatively - how much do our results change when we alter the model or its assumptions?
- One example for observational studies:
  - How much larger would **unmeasured** confounders have to be than **measured confounders** to remove the entire estimated treatment effect? (Altonji et al 2005)

Sensitivity Analysis

- An alternative is to ask - quantitatively - how much do our results change when we alter the model or its assumptions?
- One example for observational studies:
  - How much larger would **unmeasured** confounders have to be than **measured confounders** to remove the entire estimated treatment effect? (Altonji et al 2005)
  - Take a small set of covariates, run your regression and store β_R

Sensitivity Analysis

- An alternative is to ask - quantitatively - how much do our results change when we alter the model or its assumptions?
- One example for observational studies:
    - How much larger would **unmeasured** confounders have to be than **measured confounders** to remove the entire estimated treatment effect? (Altonji et al 2005)
    - Take a small set of covariates, run your regression and store β_R
    - Take a larger set of covariates, run your regression and store β_F

Sensitivity Analysis

- An alternative is to ask - quantitatively - how much do our results change when we alter the model or its assumptions?
- One example for observational studies:
  - How much larger would **unmeasured** confounders have to be than **measured confounders** to remove the entire estimated treatment effect? (Altonji et al 2005)
  - Take a small set of covariates, run your regression and store β_R
  - Take a larger set of covariates, run your regression and store β_F
  - Calculate $\frac{\beta_F}{\beta_R - beta_F}$
- Eg. Nunn and Wantchekon (2011) argue that for unmeasured confounders to explain their estimated effect of the slave trade on trust, they would have to be 3 - 11 times larger than measured confounders

Heterogeneity Tests

- We have an average treatment effect

Heterogeneity Tests

- We have an average treatment effect
- But theory may predict different groups are affected to different degrees

Heterogeneity Tests

- We have an average treatment effect
- But theory may predict different groups are affected to different degrees
- We can test for heterogeneous effects: **Conditional Average Treatment Effects (CATE)**

Heterogeneity Tests

- We have an average treatment effect
- But theory may predict different groups are affected to different degrees
- We can test for heterogeneous effects: **Conditional Average Treatment Effects (CATE)**
- $Y_i \; \beta_1 D_i + \beta_2 X_i + \beta_3 D_i * X_i + \epsilon_i$

Heterogeneity Tests

- We have an average treatment effect
- But theory may predict different groups are affected to different degrees
- We can test for heterogeneous effects: **Conditional Average Treatment Effects (CATE)**
- $Y_i \; \beta_1 D_i + \beta_2 X_i + \beta_3 D_i * X_i + \epsilon_i$
- X_i MUST be a **pre-treatment** covariate we are testing for heterogeneous effects on

Heterogeneity Tests

- We have an average treatment effect
- But theory may predict different groups are affected to different degrees
- We can test for heterogeneous effects: **Conditional Average Treatment Effects (CATE)**
- $Y_i \ \beta_1 D_i + \beta_2 X_i + \beta_3 D_i * X_i + \epsilon_i$
- X_i MUST be a **pre-treatment** covariate we are testing for heterogeneous effects on
- CRUCIAL: Our **covariate** is not randomly assigned, so the interpretation of causal effects is **not causal**, just descriptive

Heterogeneity Tests

- Ex. Ferraz and Finan (2008)
  - Audits reduce corruption, they argue due to electoral accountability

Heterogeneity Tests

- Ex. Ferraz and Finan (2008)
  - Audits reduce corruption, they argue due to electoral accountability
  - The effects should therefore be stronger where more people know about the audits

Heterogeneity Tests

- Ex. Ferraz and Finan (2008)
  - Audits reduce corruption, they argue due to electoral accountability
  - The effects should therefore be stronger where more people know about the audits
  - And for first-term Mayors with re-eletion incentives
- Are there other theories consistent with *all* of this evidence?

Heterogeneity Tests

- Ex. Ferraz and Finan (2008)
  - Audits reduce corruption, they argue due to electoral accountability
  - The effects should therefore be stronger where more people know about the audits
  - And for first-term Mayors with re-eletion incentives
- Are there other theories consistent with *all* of this evidence?
- Note this does not mean that being a first-term mayor *causes* audits to be less effective

Heterogeneity Tests

- But what if we look for heterogeneous effects on 20 variables?

Heterogeneity Tests

- But what if we look for heterogeneous effects on 20 variables?
- And then construct an appropriate theory based on the variables that show differential effects

Heterogeneity Tests

- But what if we look for heterogeneous effects on 20 variables?
- And then construct an appropriate theory based on the variables that show differential effects
- Theory first! Avoid *ex post* construction of theory and data-mining

Heterogeneity Tests

- But what if we look for heterogeneous effects on 20 variables?
- And then construct an appropriate theory based on the variables that show differential effects
- Theory first! Avoid *ex post* construction of theory and data-mining
- At least correct p-values for multiple testing

Heterogeneity Tests

- But what if we look for heterogeneous effects on 20 variables?
- And then construct an appropriate theory based on the variables that show differential effects
- Theory first! Avoid *ex post* construction of theory and data-mining
- At least correct p-values for multiple testing
- More details on this egap page

Placebo Tests

- How likely is it that our treatment effect is just a product of messy data?

Placebo Tests

- How likely is it that our treatment effect is just a product of messy data?
- Normally we test for a treatment effect where we expect one

Placebo Tests

- How likely is it that our treatment effect is just a product of messy data?
- Normally we test for a treatment effect where we expect one
- But we can also test for a treatment effect where we **don't** expect one
  - Evidence of no treatment effect supports our interpretation

Placebo Tests

- How likely is it that our treatment effect is just a product of messy data?
- Normally we test for a treatment effect where we expect one
- But we can also test for a treatment effect where we **don't** expect one
  - Evidence of no treatment effect supports our interpretation
  - Evidence of a treatment effect suggests messy data

Placebo Tests

- How likely is it that our treatment effect is just a product of messy data?
- Normally we test for a treatment effect where we expect one
- But we can also test for a treatment effect where we **don't** expect one
  - Evidence of no treatment effect supports our interpretation
  - Evidence of a treatment effect suggests messy data
- Common for regression discontinuities (alternative thresholds) and difference-in-differences (alternative times of treatment)

Table 2: **The LPT effect on the PT electoral support in presidential elections (2002-2018)**

	PT (2002)	PT (2006)	PT (2010)	PT (2014)	PT (2018)
LATE	-2.62	6.90***	4.87**	5.97***	5.59**
	(2.12)	(2.68)	(2.32)	(2.46)	(2.62)
BW est (h)	5.28	4.50	5.00	4.31	4.39
BW bias (b)	8.27	7.88	8.24	7.32	7.11
N Left	1711	1711	1711	1711	1711
N Right	3851	3851	3851	3851	3851
Eff N Left	351	303	334	289	295
Eff N Right	491	412	462	389	399
N clusters Left	523	506	521	478	466
N clusters Right	879	826	871	737	697

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. RD local linear estimates using Calonico et al. (2014b) optimal bandwidth triangular kernel selection. Robust standard errors, clustered at the municipal level, in parenthesis. Controls: the expectation of schooling years, and share of households with the mid-school degree. N Left and N Right represent the total number of observation in the left and right sides of the cutoff. Eff N Left and Eff N Right are the number of cases within the bandwidth. BW est (h) is the Bandwidth used to compute the LATE (Local Average Treatment Effect). BW bias (b) is the Bandwidth used to compute the standard errors.

Generalizability

- How ’weird’ are the units we are measuring the Local Average Treatment Effect for?

Generalizability

- How ’weird’ are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers

Generalizability

- How ’weird’ are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers
- We don’t know if any single individual is a complier

Generalizability

- How 'weird' are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers
- We don't know if any single individual is a complier
- But we can describe them **on average**

Generalizability

- How ’weird’ are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers
- We don’t know if any single individual is a complier
- But we can describe them **on average**
- The first stage of the IV regression tells us about compliance with treatment

Generalizability

- How ’weird’ are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers
- We don’t know if any single individual is a complier
- But we can describe them **on average**
- The first stage of the IV regression tells us about compliance with treatment
- Relative likelihood that a complier has covariate X equals:

Generalizability

- How 'weird' are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers
- We don't know if any single individual is a complier
- But we can describe them **on average**
- The first stage of the IV regression tells us about compliance with treatment
- Relative likelihood that a complier has covariate X equals:

$$\frac{\text{First Stage Effect for Units with Covariate X}}{\text{First Stage Effect for Everyone}}$$

Generalizability

- How 'weird' are the units we are measuring the Local Average Treatment Effect for?
- We can try to *describe* the characteristics of these compliers
- We don't know if any single individual is a complier
- But we can describe them **on average**
- The first stage of the IV regression tells us about compliance with treatment
- Relative likelihood that a complier has covariate X equals:

$$\frac{\text{First Stage Effect for Units with Covariate X}}{\text{First Stage Effect for Everyone}}$$

$$\frac{Pr(D_i = 1 \& Z_i = 1 | X_i = 1)}{Pr(D_i = 1 \& Z_i = 1)}$$

TABLE 4.4.3

Complier characteristics ratios for twins and sex composition instruments

Variable	$P[x_{1i}=1]$ (1)	Twins at Second Birth: $P[x_{1i}=1 \mid D_{1i}>D_{0i}]$ (2)	Twins at Second Birth: $P[x_{1i}=1 \mid D_{1i}>D_{0i}]/P[x_{1i}=1]$ (3)	First Two Children Are Same Sex: $P[x_{1i}=1 \mid D_{1i}>D_{0i}]$ (4)	First Two Children Are Same Sex: $P[x_{1i}=1 \mid D_{1i}>D_{0i}]/P[x_{1i}=1]$ (5)
Age 30 or older at first birth	.0029	.004	1.39	.0023	.995
Black or hispanic	.125	.103	.822	.102	.814
High school graduate	.822	.861	1.048	.815	.998
College graduate	.132	.151	1.14	.0904	.704

Notes: The table reports an analysis of complier characteristics for twins and sex composition instruments. The ratios in columns 3 and 5 give the relative likelihood that compliers have the characteristic indicated at left. Data are from the 1980 census 5 percent sample, including married mothers aged 21–35 with at least two children, as in Angrist and Evans (1998). The sample size is 254,654 for all columns.

Generalizability

- Replication is crucial to our ability to generalize

Generalizability

- Replication is crucial to our ability to generalize
- Replication in different samples from the same population

Generalizability

- Replication is crucial to our ability to generalize
- Replication in different samples from the same population
- Replication in different populations

Generalizability

- Replication is crucial to our ability to generalize
- Replication in different samples from the same population
- Replication in different populations
- Replication of different treatment implementations

Mechanisms

- To avoid the critique that experiments are a black box, and to support specific theories, we need to start testing **causal mechanisms**

Mechanisms

- To avoid the critique that experiments are a black box, and to support specific theories, we need to start testing **causal mechanisms**
- We have already seen how to use process tracing to ’test’ specific mechanisms in individual cases

Mechanisms

- To avoid the critique that experiments are a black box, and to support specific theories, we need to start testing **causal mechanisms**
- We have already seen how to use process tracing to ’test’ specific mechanisms in individual cases
- Quantitative tests also exist, exploiting ’post-treatment bias’

Mechanisms

- To avoid the critique that experiments are a black box, and to support specific theories, we need to start testing **causal mechanisms**
- We have already seen how to use process tracing to ’test’ specific mechanisms in individual cases
- Quantitative tests also exist, exploiting ’post-treatment bias’
- But require additional assumptions: **Sequential ignorability**
  - That the mediator (mechanism) is independent of potential outcomes conditional on treatment

Mechanisms

- To avoid the critique that experiments are a black box, and to support specific theories, we need to start testing **causal mechanisms**
- We have already seen how to use process tracing to ’test’ specific mechanisms in individual cases
- Quantitative tests also exist, exploiting ’post-treatment bias’
- But require additional assumptions: **Sequential ignorability**
  - That the mediator (mechanism) is independent of potential outcomes conditional on treatment
  - Hard!

Mechanisms

- One practical approach is to run two regressions that recreates our DAG:

$$M_i = \alpha_1 + \beta_1 D_i + \epsilon_1$$

$$Y_i = \alpha_3 + \beta_3 D_i + \beta_4 M_i + \epsilon_3$$

- This implies:

$$Y_i = \alpha_3 + D_i(\beta_3 + \beta_4 * \beta_1) + (\alpha_1 + \epsilon_1) * \beta_4 + \epsilon_3$$

Mechanisms

- One practical approach is to run two regressions that recreates our DAG:

$$M_i = \alpha_1 + \beta_1 D_i + \epsilon_1$$

$$Y_i = \alpha_3 + \beta_3 D_i + \beta_4 M_i + \epsilon_3$$

- This implies:

$$Y_i = \alpha_3 + D_i(\beta_3 + \beta_4 * \beta_1) + (\alpha_1 + \epsilon_1) * \beta_4 + \epsilon_3$$

- Direct effect of treatment = β_3

Mechanisms

- One practical approach is to run two regressions that recreates our DAG:

$$M_i = \alpha_1 + \beta_1 D_i + \epsilon_1$$

$$Y_i = \alpha_3 + \beta_3 D_i + \beta_4 M_i + \epsilon_3$$

- This implies:

$$Y_i = \alpha_3 + D_i(\beta_3 + \beta_4 * \beta_1) + (\alpha_1 + \epsilon_1) * \beta_4 + \epsilon_3$$

- Direct effect of treatment = β_3
- Indirect effect of treatment = $\beta_4 * \beta_1$

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?
- We can **constrain ourselves**

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?
- We can **constrain ourselves**
- Submit a Pre-Analysis Plan, eg. to egap or see BITSS

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?
- We can **constrain ourselves**
- Submit a Pre-Analysis Plan, eg. to egap or see BITSS
- Document the theory and hypotheses you're using (to avoid fitting an explanation to the data)

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?
- We can **constrain ourselves**
- Submit a Pre-Analysis Plan, eg. to egap or see BITSS
- Document the theory and hypotheses you're using (to avoid fitting an explanation to the data)
- Document the regressions you will run (to avoid data-mining)

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?
- We can **constrain ourselves**
- Submit a Pre-Analysis Plan, eg. to egap or see BITSS
- Document the theory and hypotheses you're using (to avoid fitting an explanation to the data)
- Document the regressions you will run (to avoid data-mining)
- If you need to change later, no problem! Just need to justify why

Pre-Analysis Plans

- There are a lot of tests and specifications we can run!
- How do we know what is *ex post* data-mining and what is a real test of a specific theory?
- We can **constrain ourselves**
- Submit a Pre-Analysis Plan, eg. to egap or see BITSS
- Document the theory and hypotheses you're using (to avoid fitting an explanation to the data)
- Document the regressions you will run (to avoid data-mining)
- If you need to change later, no problem! Just need to justify why
- It's transparent how far away we have come from the original test of theory