SPIA 586a: Machine Learning for Policy Analysis
Week 5: Prediction in a Counterfactual Population

Brandon Stewart[1]

Princeton

February 24–26, 2025

[1] Some causal inference material drawn from: Matt Blackwell, Justin Grimmer, Jens Hainmueller, Erin Hartman, and Kosuke Imai.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - prediction in a new population
  - random forests/boosting

Where We've Been and Where We're Going...

- Last Week
  - prediction in a new population
  - random forests/boosting
- This Week
  - prediction in a counterfactual population (aka causal inference)
  - neural networks

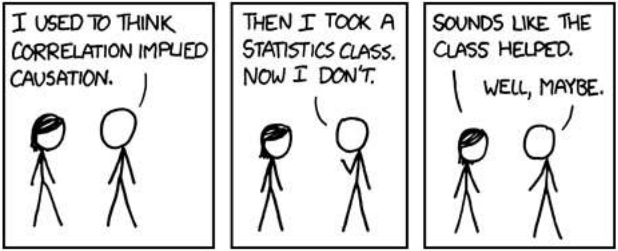
Where We've Been and Where We're Going...

- Last Week
  - prediction in a new population
  - random forests/boosting
- This Week
  - prediction in a counterfactual population (aka causal inference)
  - neural networks
- Next Week
  - experimental data

Where We've Been and Where We're Going...

- Last Week
  - prediction in a new population
  - random forests/boosting
- This Week
  - prediction in a counterfactual population (aka causal inference)
  - neural networks
- Next Week
  - experimental data
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Causal Questions

Causal Questions

- Will advertising this product to this customer make a sale that otherwise wouldn't have happened?

Causal Questions

- Will advertising this product to this customer make a sale that otherwise wouldn't have happened?
- Will implementing this policy help people?

Causal Questions

- Will advertising this product to this customer make a sale that otherwise wouldn't have happened?
- Will implementing this policy help people?
- Would sending police to this location lower crime?

Causal Questions

- Will advertising this product to this customer make a sale that otherwise wouldn't have happened?
- Will implementing this policy help people?
- Would sending police to this location lower crime?
- Does getting an A in this class help me get a job?

What is Causal Inference?

What is Causal Inference?

- Loosely, causal inference is a statement about counterfactuals—it is a statement about the difference between what would happen with an intervention and what would happen without the intervention.

What is Causal Inference?

- Loosely, causal inference is a statement about counterfactuals—it is a statement about the difference between what would happen with an intervention and what would happen without the intervention.
- We can predict what didn't (or wouldn't) happen but, by definition, this comes from a different distribution what did (or would).
- The core puzzle of causal inference is how you get the information to predict outcomes under both states of the intervention.

What is Causal Inference?

- Loosely, causal inference is a statement about counterfactuals—it is a statement about the difference between what would happen with an intervention and what would happen without the intervention.
- We can predict what didn't (or wouldn't) happen but, by definition, this comes from a different distribution what did (or would).
- The core puzzle of causal inference is how you get the information to predict outcomes under both states of the intervention.

Key Pieces

Key Pieces

- Intervention/Treatment

Key Pieces

- Intervention/Treatment
- Outcome

Key Pieces

- Intervention/Treatment
- Outcome
- Confounders (the common causes of treatment and outcome)

Fundamental Problem of Causal Inference

Fundamental Problem of Causal Inference

- The hard thing about counterfactuals is that we never get to see all of them: Fundamental Problem of Causal Inference (Holland 1986).

Fundamental Problem of Causal Inference

- The hard thing about counterfactuals is that we never get to see all of them: Fundamental Problem of Causal Inference (Holland 1986).
- Assumptions and careful design are the only way out of this problem because we never get to see the truth (unlike last week, we can't empirically check that we are doing okay).

Fundamental Problem of Causal Inference

- The hard thing about counterfactuals is that we never get to see all of them: Fundamental Problem of Causal Inference (Holland 1986).
- Assumptions and careful design are the only way out of this problem because we never get to see the truth (unlike last week, we can't empirically check that we are doing okay).
- When it works though it can be a powerful view into the things that we care the most about.

Fundamental Problem of Causal Inference

- The hard thing about counterfactuals is that we never get to see all of them: Fundamental Problem of Causal Inference (Holland 1986).
- Assumptions and careful design are the only way out of this problem because we never get to see the truth (unlike last week, we can't empirically check that we are doing okay).
- When it works though it can be a powerful view into the things that we care the most about.
- By convention we often care the counterfactual levels we care about treated and control and we often consider only binary treatment variables because continuous variables are often even more complicated!

Why Can't I Just Use Prediction?

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\left(\text{Earnings if College} - \text{Earnings if not College}\right)$$

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\left(\text{Earnings if College} - \text{Earnings if not College}\right)$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\,(\text{Earnings if College} - \text{Earnings if not College})$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.
  - We have: Earnings if College

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\left(\text{Earnings if College} - \text{Earnings if not College}\right)$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.
  - We have: Earnings if College
  - We need: Earnings if not College

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\left(\text{Earnings if College} - \text{Earnings if not College}\right)$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.
  - We have: Earnings if College
  - We need: Earnings if not College
  - To predict Earnings if not College, we only have people who actually didn't go to college

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\,(\text{Earnings if College} - \text{Earnings if not College})$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.
  - We have: Earnings if College
  - We need: Earnings if not College
  - To predict Earnings if not College, we only have people who actually didn't go to college
- When is this okay?

Why Can't I Just Use Prediction?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\left(\text{Earnings if College} - \text{Earnings if not College}\right)$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.
    - We have: Earnings if College
    - We need: Earnings if not College
    - To predict Earnings if not College, we only have people who actually didn't go to college
- When is this okay?
- The key trick of causal inference: finding sets of data where distribution of observed is equal to distribution of counterfactual.

Secret Sauce: Separating Identification and Estimation

- Identification: How much can you learn about the estimand if you have an infinite amount of data? (i.e., do we have data from the right distribution to learn with?)
- Estimation: How much can you learn about the estimand from a finite sample? (i.e. can we do that learning in practice?)

Secret Sauce: Separating Identification and Estimation

- Identification: How much can you learn about the estimand if you have an infinite amount of data? (i.e., do we have data from the right distribution to learn with?)
- Estimation: How much can you learn about the estimand from a finite sample? (i.e. can we do that learning in practice?)
- Identification precedes estimation, if identified all we have to do is predict well.

Secret Sauce: Separating Identification and Estimation

- Identification: How much can you learn about the estimand if you have an infinite amount of data? (i.e., do we have data from the right distribution to learn with?)
- Estimation: How much can you learn about the estimand from a finite sample? (i.e. can we do that learning in practice?)
- Identification precedes estimation, if identified all we have to do is predict well.

The role of assumptions:

- Often identification requires (hopefully minimal) assumptions

Secret Sauce: Separating Identification and Estimation

- Identification: How much can you learn about the estimand if you have an infinite amount of data? (i.e., do we have data from the right distribution to learn with?)
- Estimation: How much can you learn about the estimand from a finite sample? (i.e. can we do that learning in practice?)
- Identification precedes estimation, if identified all we have to do is predict well.

The role of assumptions:

- Often identification requires (hopefully minimal) assumptions
- ML helps with the estimation.

Secret Sauce: Separating Identification and Estimation

- Identification: How much can you learn about the estimand if you have an infinite amount of data? (i.e., do we have data from the right distribution to learn with?)
- Estimation: How much can you learn about the estimand from a finite sample? (i.e. can we do that learning in practice?)
- Identification precedes estimation, if identified all we have to do is predict well.

The role of assumptions:

- Often identification requires (hopefully minimal) assumptions
- ML helps with the estimation.
- Law of Decreasing Credibility (Manski): The credibility of inference decreases with the strength of the assumptions maintained

Potential Outcomes – Aspirin Example

Definitions:

Potential Outcomes – Aspirin Example

Definitions:

T_i: Unit assigned to:

$$T_i = \begin{cases} 1 & \text{Receive Aspirin} \\ 0 & \text{Receive Placebo} \end{cases}$$

$(T_i = 1)$ $(T_i = 0)$

Potential Outcomes – Aspirin Example

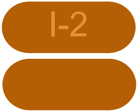
Definitions:

T_i: Unit assigned to:

$$T_i = \begin{cases} 1 & \text{Receive Aspirin} \\ 0 & \text{Receive Placebo} \end{cases}$$

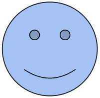
Y_i: Outcome for unit i – Patient has headache, or not

$(T_i = 1)$



$(T_i = 0)$

$(Y_i = 1)$



$(Y_i = 0)$

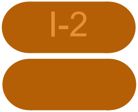
Potential Outcomes – Aspirin Example

Definitions:

T_i: Unit assigned to:

$$T_i = \begin{cases} 1 & \text{Receive Aspirin} \\ 0 & \text{Receive Placebo} \end{cases}$$

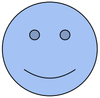
$(T_i = 1)$



$(T_i = 0)$

Y_i: Outcome for unit i – Patient has headache, or not

$(Y_i = 1)$



$(Y_i = 0)$

Potential outcomes for unit i:

$$Y_i(T_i) = \begin{cases} Y_i(1) & \text{Headache (or not) for unit } i \text{ with Aspirin} \\ Y_i(0) & \text{Headache (or not) for unit } i \text{ with placebo} \end{cases}$$

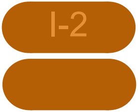
Potential Outcomes – Aspirin Example

Definitions:

T_i: Unit assigned to:

$$T_i = \begin{cases} 1 & \text{Receive Aspirin} \\ 0 & \text{Receive Placebo} \end{cases}$$

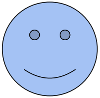
$(T_i = 1)$



$(T_i = 0)$

Y_i: Outcome for unit i – Patient has headache, or not

$(Y_i = 1)$



$(Y_i = 0)$

Potential outcomes for unit i:

$$Y_i(T_i) = \begin{cases} Y_i(1) & \text{Headache (or not) for unit } i \text{ with Aspirin} \\ Y_i(0) & \text{Headache (or not) for unit } i \text{ with placebo} \end{cases}$$

τ_i: The treatment effect

$$\tau_i = Y_i(1) - Y_i(0)$$

Illustrated potential outcomes here and later courtesy of Erin Hartman

Causal Inference is a Missing Data Problem

Example: Asprin's Impact on Headaches

Patient i		Pill T_i	Headache Status Y_i(PLACEBO)	Y_i(I-2)	Y_i	Age X_{1i}	Academic X_{2i}
1						25	Y
2						55	N
3						62	Y
4						80	N
5						32	Y
6						45	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n						71	N

Causal Inference is a Missing Data Problem

Example: Asprin's Impact on Headaches

Patient i		Pill T_i	Headache Status Y_i(PLACEBO)	Y_i(I-2)	Y_i	Age X_{1i}	Academic X_{2i}
1						25	Y
2						55	N
3						62	Y
4						80	N
5						32	Y
6						45	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n						71	N

Step 1: (Randomly) sample units

Causal Inference is a Missing Data Problem

Example: Asprin's Impact on Headaches

Patient i		Pill T_i	Headache Status Y_i(PLACEBO)	Y_i(I-2)	Y_i	Age X_{1i}	Academic X_{2i}
1		I-2				25	Y
2						55	N
3						62	Y
4		PLACEBO				80	N
5		I-2				32	Y
6						45	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n		PLACEBO				71	N

Step 2: Randomly assign treatment

Causal Inference is a Missing Data Problem

Example: Asprin's Impact on Headaches

Patient i		Pill T_i	Headache Status Y_i(PLACEBO)	Y_i(I-2)	Y_i	Age X_{1i}	Academic X_{2i}
1		I-2				25	Y
2						55	N
3						62	Y
4		PLACEBO				80	N
5		I-2				32	Y
6						45	B
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n		PLACEBO				71	N

Step 3: Measure revealed potential outcome

Prediction Policy Problems via Athey

Prediction Policy Problems via Athey

- Policy problems often involve tradeoffs and reactions of people—both cause complications.

Prediction Policy Problems via Athey

- Policy problems often involve tradeoffs and reactions of people—both cause complications.
- For example on city inspectors, we must worry about:
  - out of domain prediction
  - behavior of establishment being inspected (i.e. responsiveness to fines)
  - cost of fixing the problem

Prediction Policy Problems via Athey

- Policy problems often involve tradeoffs and reactions of people—both cause complications.
- For example on city inspectors, we must worry about:
  - out of domain prediction
  - behavior of establishment being inspected (i.e. responsiveness to fines)
  - cost of fixing the problem
- Causal problem: what is the expected improvement in overall quantity of units in the city under a new inspector allocation regime?

Prediction Policy Problems via Athey

- Policy problems often involve tradeoffs and reactions of people—both cause complications.
- For example on city inspectors, we must worry about:
  - out of domain prediction
  - behavior of establishment being inspected (i.e. responsiveness to fines)
  - cost of fixing the problem
- Causal problem: what is the expected improvement in overall quantity of units in the city under a new inspector allocation regime?
- Ebay advertisement data:
  - correlation: 1400% ROI
  - causation: -63% ROI

Prediction Policy Problems via Athey

- Policy problems often involve tradeoffs and reactions of people—both cause complications.
- For example on city inspectors, we must worry about:
  - out of domain prediction
  - behavior of establishment being inspected (i.e. responsiveness to fines)
  - cost of fixing the problem
- Causal problem: what is the expected improvement in overall quantity of units in the city under a new inspector allocation regime?
- Ebay advertisement data:
  - correlation: 1400% ROI
  - causation: -63% ROI
- Key challenge: people who clicked would have purchased anyway

Prediction Policy Problems via Athey

- Policy problems often involve tradeoffs and reactions of people—both cause complications.
- For example on city inspectors, we must worry about:
  - out of domain prediction
  - behavior of establishment being inspected (i.e. responsiveness to fines)
  - cost of fixing the problem
- Causal problem: what is the expected improvement in overall quantity of units in the city under a new inspector allocation regime?
- Ebay advertisement data:
  - correlation: 1400% ROI
  - causation: -63% ROI
- Key challenge: people who clicked would have purchased anyway
- The difficulty in all these cases is that we can't observe the causal part.

How do the Frameworks Help?

How do the Frameworks Help?

- Core idea: address confounding with pretreatment covariates X (there are other ways too!)

How do the Frameworks Help?

- Core idea: address confounding with pretreatment covariates X (there are other ways too!)
- One strategy: run separate analyses within levels of X and combine them.

How do the Frameworks Help?

- Core idea: address confounding with pretreatment covariates X (there are other ways too!)
- One strategy: run separate analyses within levels of X and combine them.
- What if we only have a few observations per level of X? (curse of dimensionality!)

How do the Frameworks Help?

- Core idea: address confounding with pretreatment covariates X (there are other ways too!)
- One strategy: run separate analyses within levels of X and combine them.
- What if we only have a few observations per level of X? (curse of dimensionality!)
- That's what all the machine learning methods have addressed! They are model-based ways of smoothing over the different groups.

How do the Frameworks Help?

- Core idea: address confounding with pretreatment covariates X (there are other ways too!)
- One strategy: run separate analyses within levels of X and combine them.
- What if we only have a few observations per level of X? (curse of dimensionality!)
- That's what all the machine learning methods have addressed! They are model-based ways of smoothing over the different groups.
- In other words: if we just include the right variables and don't include the wrong variables we can do prediction and be fine!

How do the Frameworks Help?

- Core idea: address confounding with pretreatment covariates X (there are other ways too!)
- One strategy: run separate analyses within levels of X and combine them.
- What if we only have a few observations per level of X? (curse of dimensionality!)
- That's what all the machine learning methods have addressed! They are model-based ways of smoothing over the different groups.
- In other words: if we just include the right variables and don't include the wrong variables we can do prediction and be fine!
- Mechanically this ends up looking easy: predict the outcome with $T = 1$ and predict the outcome with $T = 0$ then take the difference! (more later!)

What are the “Right” Variables?

What are the "Right" Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.

What are the “Right” Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.
- How do we know which variables are which? (Use the Frameworks!)

What are the "Right" Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.
- How do we know which variables are which? (Use the Frameworks!)
  - Potential Outcomes: Those such that $Y(0), Y(1) \perp\!\!\!\perp T | X$ $\rightsquigarrow$ intuitively, the pre-treatment things such that it is as though you have random assignments within levels of X

What are the "Right" Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.
- How do we know which variables are which? (Use the Frameworks!)
  - Potential Outcomes: Those such that $Y(0), Y(1) \perp\!\!\!\perp T | X$
    $\rightsquigarrow$ intuitively, the pre-treatment things such that it is as though you have random assignments within levels of X
  - DAG Framework: Those that block all backdoor paths
    $\rightsquigarrow$ draw your causal model and the algorithm will tell you which sets of variables.

What are the “Right” Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.
- How do we know which variables are which? (Use the Frameworks!)
  - Potential Outcomes: Those such that $Y(0), Y(1) \perp\!\!\!\perp T | X$
    $\leadsto$ intuitively, the pre-treatment things such that it is as though you have random assignments within levels of X
  - DAG Framework: Those that block all backdoor paths
    $\leadsto$ draw your causal model and the algorithm will tell you which sets of variables.
- In both cases the rough intuition: common sources of variation between treatment and outcome (i.e. it is okay to omit things that just cause treatment and just cause the outcome)

What are the "Right" Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.
- How do we know which variables are which? (Use the Frameworks!)
  - Potential Outcomes: Those such that $Y(0), Y(1) \perp\!\!\!\perp T | X$
    $\rightsquigarrow$ intuitively, the pre-treatment things such that it is as though you have random assignments within levels of X
  - DAG Framework: Those that block all backdoor paths
    $\rightsquigarrow$ draw your causal model and the algorithm will tell you which sets of variables.
- In both cases the rough intuition: common sources of variation between treatment and outcome (i.e. it is okay to omit things that just cause treatment and just cause the outcome)
- Don't include post-treatment variables unless you really know what you are doing.

What are the “Right” Variables?

- Generally speaking we want to control for/condition on/use as predictors pre-treatment confounders and not colliders.
- How do we know which variables are which? (Use the Frameworks!)
  - Potential Outcomes: Those such that $Y(0), Y(1) \perp\!\!\!\perp T | X$
    $\rightsquigarrow$ intuitively, the pre-treatment things such that it is as though you have random assignments within levels of X
  - DAG Framework: Those that block all backdoor paths
    $\rightsquigarrow$ draw your causal model and the algorithm will tell you which sets of variables.
- In both cases the rough intuition: common sources of variation between treatment and outcome (i.e. it is okay to omit things that just cause treatment and just cause the outcome)
- Don't include post-treatment variables unless you really know what you are doing.
- NB: analyzing within a subset of data is the same as conditioning on the variable defining the subset!

An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On nonlethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

From "Bloody Sunday" on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

Electronically published April 22, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 3]

It is the most surprising result of my career.
— Roland Fryer

An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On nonlethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

From "Bloody Sunday" on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

Electronically published April 22, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 3]

It is the most surprising result of my career.
— Roland Fryer

An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On non-lethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

From "Bloody Sunday" on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

Electronically published April 22, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 3]

1210

TheUpshot

DATA DIVE

Surprising New Evidence Shows Bias in Police Use of Force but Not in Shootings

It is the most surprising result of my career.
— Roland Fryer

An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On nonlethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

From "Bloody Sunday" on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

Electronically published April 22, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 3]

1210

It is the most surprising result of my career.
— Roland Fryer

An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On nonlethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

From "Bloody Sunday" on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

Electronically published April 22, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 3]

1210

Evidence:

Claim:

Why wrong:

Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim:

Why wrong:

Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim: Police are unbiased

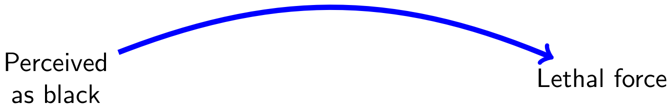
Why wrong:

Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim: Police are unbiased

Why wrong:

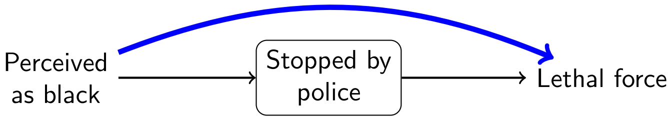


Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim: Police are unbiased

Why wrong:

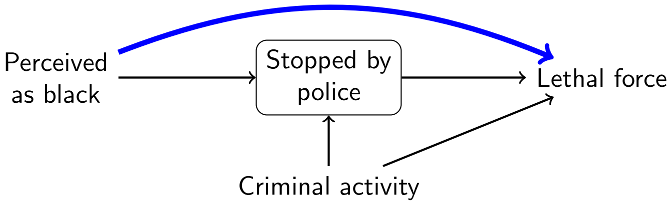


Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

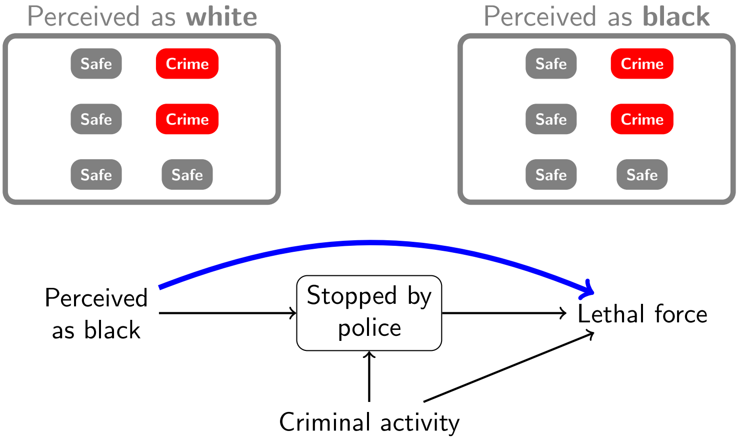
Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim: Police are unbiased

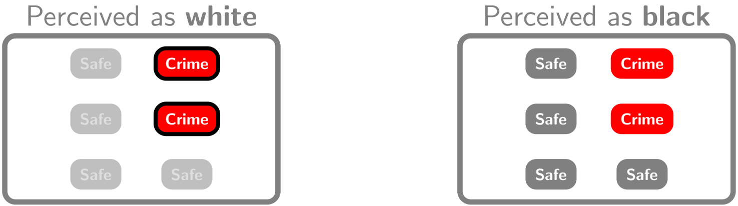
Why wrong:

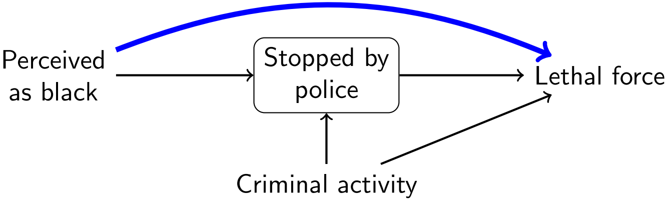


Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

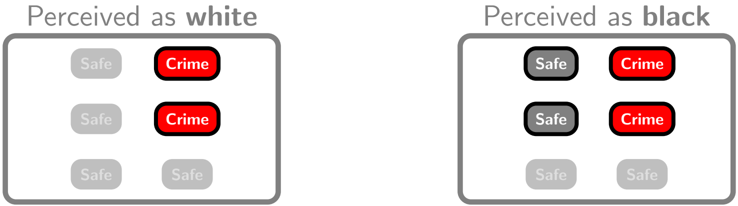


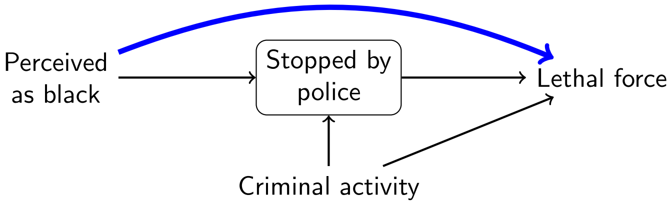
Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.





Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.





Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim: Police are unbiased

Why wrong: Causal sample selection. Among those stopped, the white civilians may be more dangerous.

Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

Claim: Police are unbiased

Why wrong: Causal sample selection. Among those stopped, the white civilians may be more dangerous.

Fryer responds:

Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

~~**Claim:** Police are unbiased~~

Why wrong: Causal sample selection. Among those stopped, the white civilians may be more dangerous.

Fryer responds:

Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

~~**Claim:** Police are unbiased~~

Why wrong: Causal sample selection. Among those stopped, the white civilians may be more dangerous.

Fryer responds: "We use the term 'racial differences' 114 times in lieu of the more prescriptive wording—'racial discrimination.' We use the phrase 'conditional on an interaction' 20 times...I am not sure how many more ways we would have needed to caveat our results to satisfy [the critics]."

Evidence: Police use lethal force at the same rate against black and white civilians who are stopped.

~~**Claim:** Police are unbiased~~

Why wrong: Causal sample selection. Among those stopped, the white civilians may be more dangerous.

Fryer responds: "We use the term 'racial differences' 114 times in lieu of the more prescriptive wording—'racial discrimination.' We use the phrase 'conditional on an interaction' 20 times...I am not sure how many more ways we would have needed to caveat our results to satisfy [the critics]."

Caveats are the problem. Estimand is only defined in terms of the regression model being fit.

Why is This a Problem?

Why is This a Problem?

What is your estimand?

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

Why is This a Problem?

What is your estimand?

A common answer:

The **purpose** of the statistical analysis

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

The model is an approximation

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

So β_1 is an approximation to...

The model is an approximation

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

The model is an approximation

So β_1 is an approximation to...

Something. I just haven't said what.

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

The model is an approximation

So β_1 is an approximation to...

Something. I just haven't said what.

Is it a good approximation?

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

The model is an approximation

So β_1 is an approximation to...

Something. I just haven't said what.

Is it a good approximation?

Epistemological crisis

Why is This a Problem?

What is your estimand?

The **purpose** of the
statistical analysis
stated **outside**
of the model

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis stated **outside** of the model

A **unit-specific quantity**

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis stated **outside** of the model

Y_i

A **unit-specific quantity**

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis stated **outside** of the model

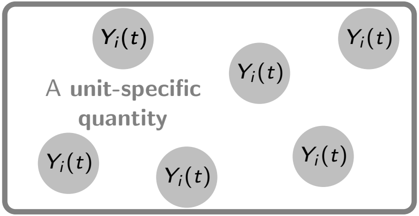
$Y_i(t)$

A **unit-specific quantity**

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis stated **outside** of the model

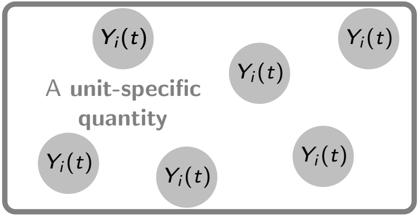


Aggregated over a **target population**

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis stated **outside** of the model



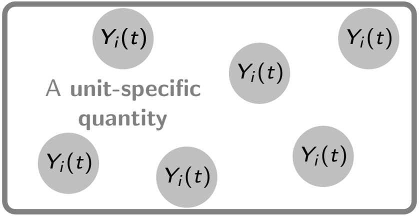
Aggregated over a **target population**

What was Fryer's estimand?

Why is This a Problem?

What is your estimand?

The **purpose** of the statistical analysis stated **outside** of the model



Aggregated over a **target population**

What was Fryer's estimand?

It could have been a lot of things (and that is the problem)

Research framework

Theory or
general goal

Estimation
strategies

Research framework

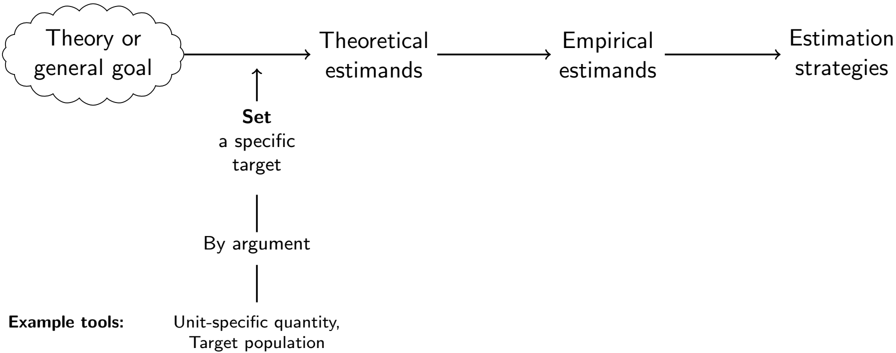
Theory or general goal $\longrightarrow$ Theoretical estimands

Estimation strategies

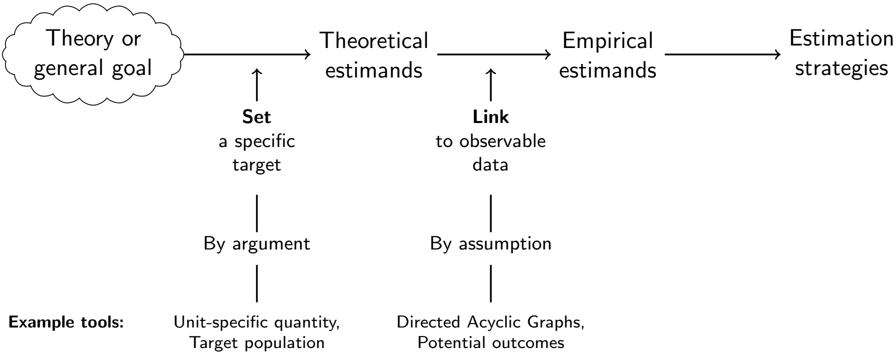
Research framework

Theory or general goal $\longrightarrow$ Theoretical estimands $\longrightarrow$ Empirical estimands $\longrightarrow$ Estimation strategies

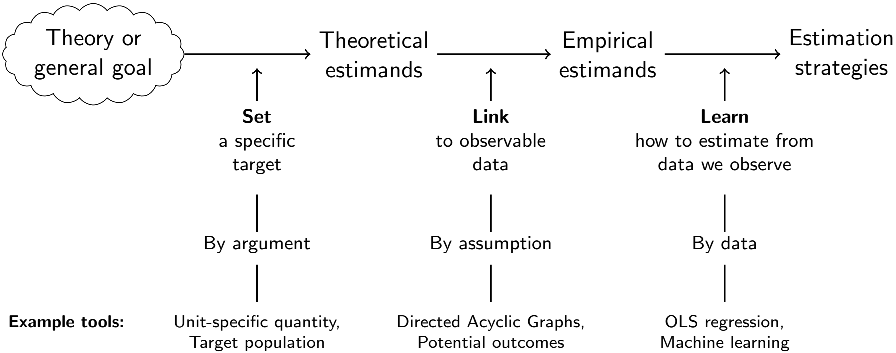
Research framework



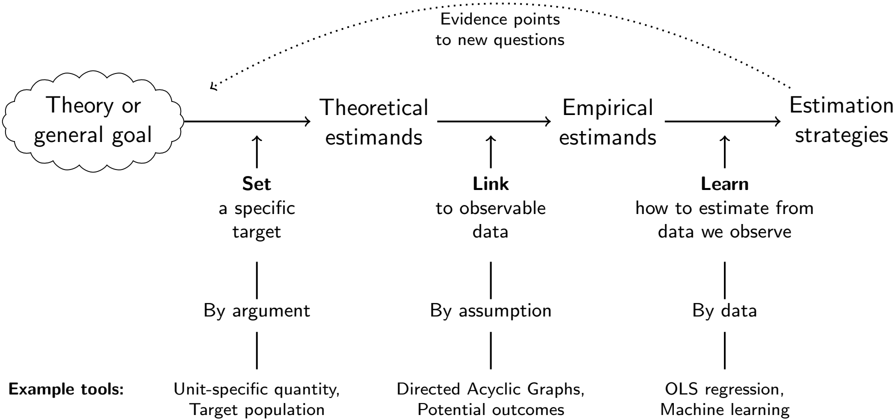
Research framework

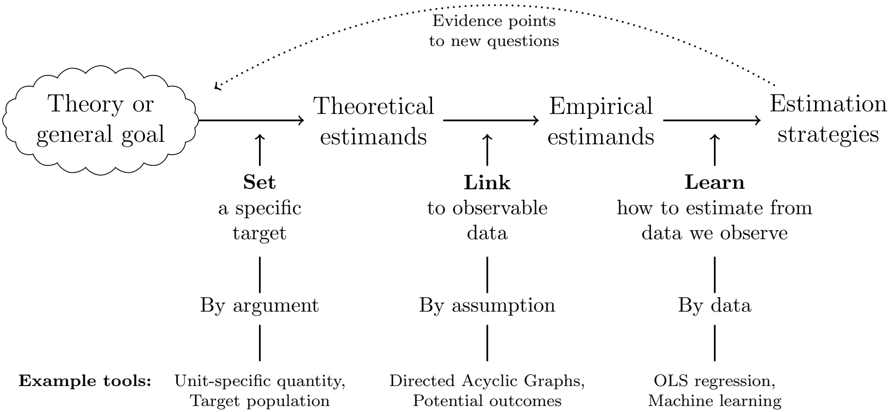


Research framework



Research framework




Evidence points
to new questions
Theory or
general goal
Theoretical
estimands
Empirical
estimands
Estimation
strategies
Set
a specific
target
Link
to observable
data
Learn
how to estimate from
data we observe
By argument
By assumption
By data
Example tools:
Unit-specific quantity,
Target population
Directed Acyclic Graphs,
Potential outcomes
OLS regression,
Machine learning

Theoretical Estimands (By Argument)

Theoretical Estimands (By Argument)

(1) Unit-Specific Quantity

Theoretical Estimands (By Argument)

(1) Unit-Specific Quantity

(2) Aggregated Over a Target Population

Theoretical Estimands (By Argument)

(1) Unit-Specific Quantity

(2) Aggregated Over a Target Population

$$\frac{1}{n}\sum_{i=1}^{n} Y_i$$

Mean over every i
among U.S. adults
(target population)

Whether each i
is employed
(unit-specific quantity)

Theoretical Estimands (By Argument)

(1) Unit-Specific Quantity

(2) Aggregated Over a Target Population

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean over every } i\\ \text{among U.S. adults}}} \Bigg(\underbrace{Y_i(1)}_{\substack{\text{Employment if}\\ \text{enrolled in}\\ \text{job training}}} - \underbrace{Y_i(0)}_{\substack{\text{Employment if}\\ \text{not enrolled in}\\ \text{job training}}} \Bigg)$$

(target population)

(unit-specific quantity)

Many Options

Estimand name	Mathematical statement	DAG	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n}\sum_i \Big(Y_i(d') - Y_i(d)\Big)$	$D \rightarrow Y$	Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x}\sum_{i:X_i=x} \Big(Y_i(d') - Y_i(d)\Big)$	$X \rightarrow D \rightarrow Y$, $X \rightarrow Y$	Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n}\sum_i \Big(\big(Y_i(a',d') - Y_i(a',d)\big) - \big(Y_i(a,d') - Y_i(a,d)\big)\Big)$	$A \rightarrow Y$, $D \rightarrow Y$	Vanderweele (2015)	Joint treatment effect
Controlled direct effect	$\frac{1}{n}\sum_i \Big(Y_i(d',m) - Y_i(d,m)\Big)$	$D \rightarrow M \rightarrow Y$, $D \rightarrow Y$	Acharya et al. (2016)	Mediation (Illustrations: Example 2)
Natural direct effect	$\frac{1}{n}\sum_i \Big(Y_i(d',M_i(d)) - Y_i(d,M_i(d))\Big)$	$D \rightarrow M \rightarrow Y$, $D \rightarrow Y$	Imai et al. (2011)	Mediation (Part B of the Online Supplement)
Effect of time-varying treatment	$\frac{1}{n}\sum_i \Big(Y_i(d_1',d_2') - Y_i(d_1,d_2)\Big)$	$D_1 \rightarrow D_2 \rightarrow Y$, $D_1 \rightarrow Y$	Wodtke et al. (2011)	Cumulative effect

Subtleties with Pager (2003)

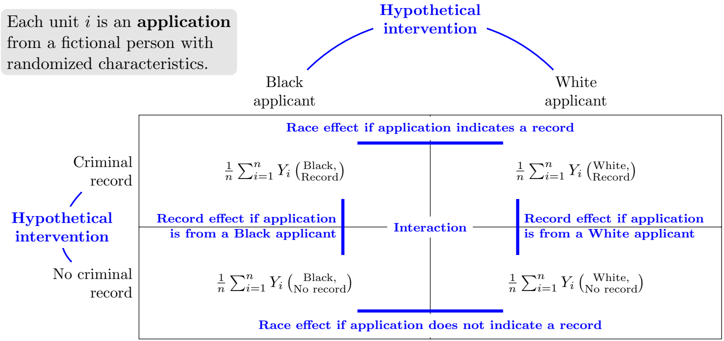
- Pager (2003) explores “the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities”

Subtleties with Pager (2003)

- Pager (2003) explores "the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities"
- Key insight: Each unit i is an application, not a person

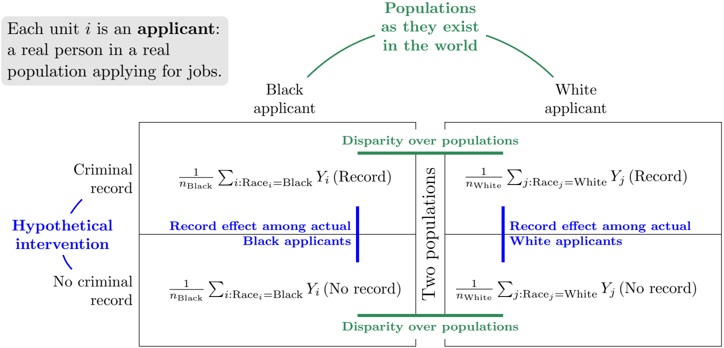
Subtleties with Pager (2003)

- Pager (2003) explores "the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities"
- Key insight: Each unit i is an application, not a person



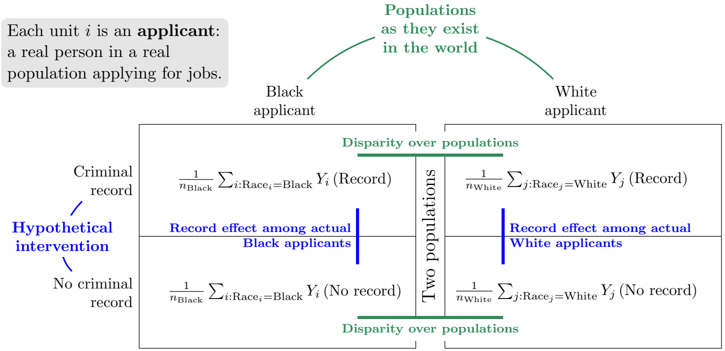
Subtleties with Pager (2003)

- Pager (2003) explores "the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities"
- Key insight: Each unit i is an application, not a person



Subtleties with Pager (2003)

- Pager (2003) explores "the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities"
- Key insight: Each unit *i* is an application, not a person



Different estimands suggest different policy and theory.

Target Population (Angrist and Evans 1998)

Target Population (Angrist and Evans 1998)

Effect of motherhood
on employment

Target Population (Angrist and Evans 1998)

Effect of motherhood
on employment

First two births
are the same sex

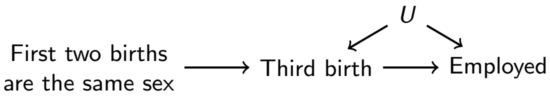
Target Population (Angrist and Evans 1998)

Effect of motherhood
on employment

$$\text{First two births are the same sex} \longrightarrow \text{Third birth}$$

Target Population (Angrist and Evans 1998)

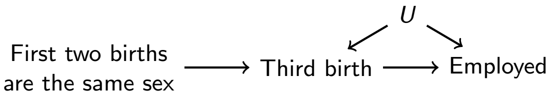
Effect of motherhood
on employment



Target Population (Angrist and Evans 1998)

Vague estimand

Effect of motherhood
on employment

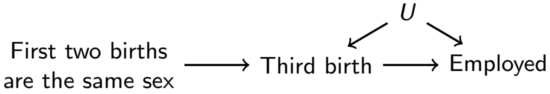


Target Population (Angrist and Evans 1998)

Vague estimand

Effect of motherhood
on employment

Precise estimand



Target Population (Angrist and Evans 1998)

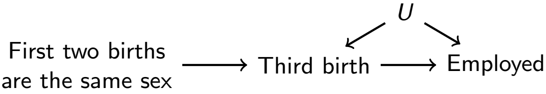
unit-specific quantity

Vague estimand

Effect of motherhood on employment

Precise estimand

Effect of having **3 vs. 2 children**



Target Population (Angrist and Evans 1998)

Vague estimand

Effect of motherhood on employment

Precise estimand

Effect of having 3 vs. 2 children [among those with at least two children who would have a third birth if and only if the first two were of the same sex]

target population

$$\begin{tikzcd}
 & & U \arrow[dl] \arrow[dr] & \\
\text{First two births are the same sex} \arrow[r] & \text{Third birth} \arrow[rr] & & \text{Employed}
\end{tikzcd}$$

Target Population (Angrist and Evans 1998)

Precise estimand

Effect of having 3 vs. 2 children
among those with at least two children who would have a third birth if and only if the first two were of the same sex

$\approx$ 4% of all mothers

Target Population (Angrist and Evans 1998)

Precise estimand

Effect of having 3 vs. 2 children
among those with at least two children who would have a third birth if and only if the first two were of the same sex

$\approx$ 4% of all mothers

You have to argue either:

1)

2)

Target Population (Angrist and Evans 1998)

Precise estimand

Effect of having 3 vs. 2 children
among those with at least two children who would have a third birth if and only if the first two were of the same sex

$\approx$ 4% of all mothers

You have to argue either:

1) That estimand matters for theory, or

2)

Target Population (Angrist and Evans 1998)

Precise estimand

Effect of having 3 vs. 2 children among those with at least two children who would have a third birth if and only if the first two were of the same sex

$\approx$ 4% of all mothers

You have to argue either:

1) That estimand matters for theory, or
2) It speaks to some broader estimand

Empirical Estimands (Assumption)

Link the Theoretical Estimand to Observable Data

Empirical Estimands (Assumption)

Link the Theoretical Estimand to Observable Data

$$\tau = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean} \\ \text{over } \textbf{all} \\ \text{applications}}} \left(Y_i \underbrace{\binom{\text{White,}}{\text{Record}}}_{\substack{\text{Potential} \\ \text{outcome under} \\ \text{one condition}}} - Y_i \underbrace{\binom{\text{Black,}}{\text{No record}}}_{\substack{\text{Potential} \\ \text{outcome under} \\ \text{another condition}}} \right)$$

Empirical Estimands (Assumption)

Link the Theoretical Estimand to Observable Data

Factual outcomes

$$\theta = \frac{1}{n_{\text{WR}}} \sum_{i \in \mathcal{S}_{\text{WR}}} Y_i - \frac{1}{n_{\text{BN}}} \sum_{i \in \mathcal{S}_{\text{BN}}} Y_i$$

Mean among applications assigned to the condition (**W**hite, **R**ecord)

Mean among applications assigned to the condition (**B**lack, **N**o record)

Empirical Estimands (Assumption)

Link the Theoretical Estimand to Observable Data

Factual outcomes

$$\theta = \frac{1}{n_{\text{WR}}} \sum_{i \in \mathcal{S}_{\text{WR}}} Y_i - \frac{1}{n_{\text{BN}}} \sum_{i \in \mathcal{S}_{\text{BN}}} Y_i$$

Mean among applications assigned to the condition (**W**hite, **R**ecord)

Mean among applications assigned to the condition (**B**lack, **N**o record)

This includes causal identification, missing data, etc.

Example: Berkeley Graduate Admissions

Example: Berkeley Graduate Admissions

Example: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973

Example: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:

Example: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate

Example: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate

Example: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate
- Evidence of discrimination toward women in admissions?

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!
- How? Overall admission rates are lower for the departments women apply to.

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!
- How? Overall admission rates are lower for the departments women apply to.
- Marginal relationships (admissions and gender) $\neq$ conditional relationship given third variable (department)

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Determining whether discrimination because of sex or ethnic identity is being practiced against persons seeking passage from one social status or locus to another is an important problem in our society today. It is legally important and morally important. It is also often quite difficult. This article is an exploration of some of the issues of measurement and assessment involved in one example of the general problem, by means of which we hope to shed some light on the difficulties. We decision to admit or to deny admission. The question we wish to pursue is whether the decision to admit or to deny was influenced by the sex of the applicant. We cannot know with any certainty the influences on the evaluators in the Graduate Admissions Office, or on the faculty reviewing committees, or on any other administrative personnel participating in the chain of actions that led to a decision on an individual application. We can, however, say that if the admissions decision and the sex by using a familiar statistic, chi-square. As already noted, we are aware of the pitfalls ahead in this naive approach, but we intend to stumble into every one of them for didactic reasons.

We must first make clear two assumptions that underlie consideration of the data in this contingency table approach. Assumption 1 is that in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of "sex bias" meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on. Theoretically one could test the assumption, for example, by examining presumably unbiased estimators of academic qualification such as Graduate Record Examination scores, undergraduate grade point averages, and so on. There are, however, enormous practical difficulties in this. We therefore predicate our discussion on the validity of assumption 1.

Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187, no. 4175 (1975): 398-404.

Bias?

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately.

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined...

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,but apparently from prior screening at earlier levels of the educational system.

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.' (Bickel et al 1975, 403, emphasis mine)

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.' (Bickel et al 1975, 403, emphasis mine)

Bias?

Bias?

- There is an important leap to their meaning for a particular policy argument.

Bias?

- There is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.

Bias?

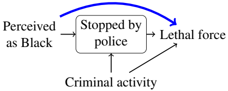
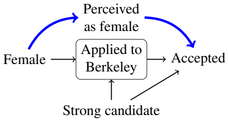
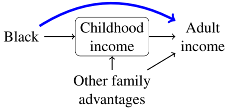
- There is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.

Bias?

- There is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.

Bias?

- There is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.
- This general pattern repeats in many debates, often because of the limits of data collection.

Study	Empirical regularity	Misleading conclusion	Directed Acyclic Graph
Fryer (2019)	Among those they stop, police shoot the same proportion of Black individuals as White individuals.	Police do not discriminate against Black individuals when using lethal force.	
Bickel et al. (1975)	Among those who apply, Berkeley departments admit a higher proportion of women than of men.	Admissions committees do not discriminate against women.	
Chetty et al. (2020)	Among those with equal childhood incomes, Black and White women earn similar amounts as adults.	Equalizing childhood incomes would eliminate the racial gap in women's adult incomes.	

Estimation (By Data)

Learn the Empirical Estimand Using Data

Estimation (By Data)

Learn the Empirical Estimand Using Data

Mean over entire sample

Expected outcome among cases with the covariate values $\vec{x}_i$ of unit i who are factually treated ($D = 1$)

$$\theta = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 0)$$

Mean over entire sample

Expected outcome among cases with the covariate values $\vec{x}_i$ of unit i who are factually untreated ($D = 0$)

Estimation (By Data)

Learn the Empirical Estimand Using Data

Mean over
entire sample

Regression prediction
at observed covariate values $\vec{x}_i$
with treatment set to $D = 1$

$$\hat{\theta} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1)$$
$$- \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 0)$$

Mean over
entire sample

Regression prediction
at observed covariate values $\vec{x}_i$
with treatment set to $D = 0$

Pal and Waldfogel (2016) estimate the family gap in pay.

Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive?

Pal and Waldfogel (2016) estimate the family gap in pay.

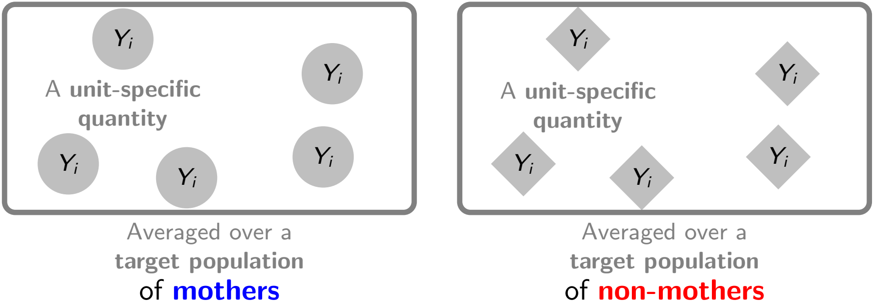
Is the theoretical estimand descriptive?

"the differential in hourly wages
between women with children
and women without children"

Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive?

"the differential in hourly wages
between women with children
and women without children"



Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive? Is it causal?

Pal and Waldfogel (2016) estimate the family gap in pay.

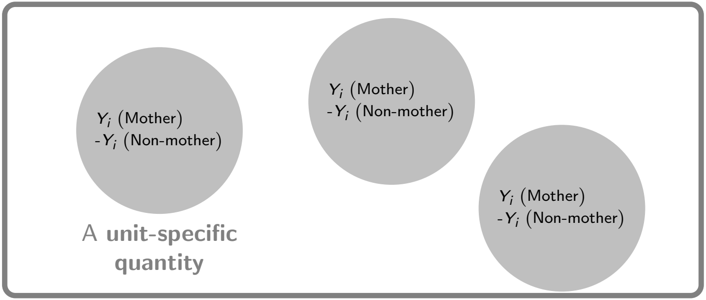
Is the theoretical estimand descriptive? Is it causal?

"causal estimation techniques"

Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive? Is it causal?

"causal estimation techniques"

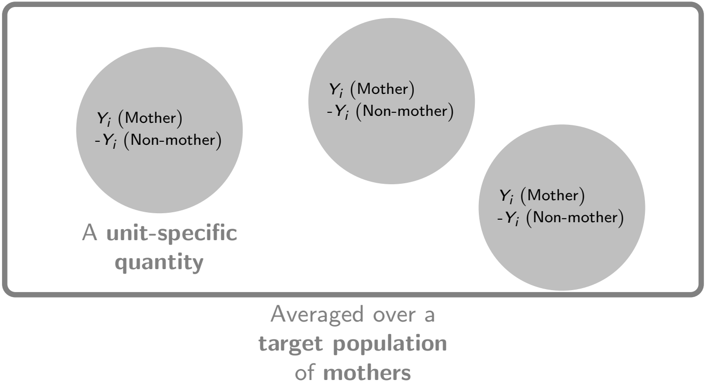


Averaged over a **target population** of **mothers**

Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive? Is it causal?

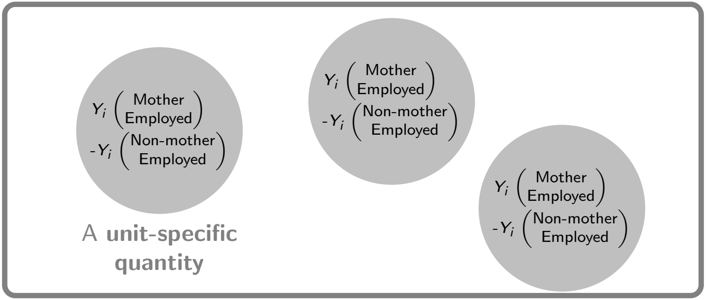
Added complexity: Wages are undefined for the non-employed.



Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive? Is it causal?

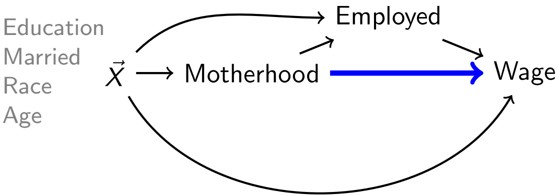
Added complexity: Wages are undefined for the non-employed.



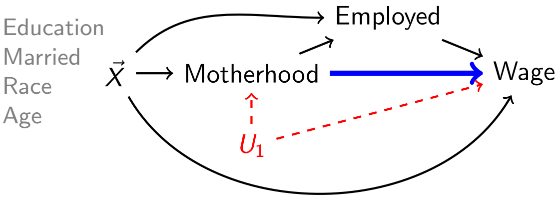
Averaged over a **target population** of **mothers**

Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\\text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\\text{Employed}\end{pmatrix}$

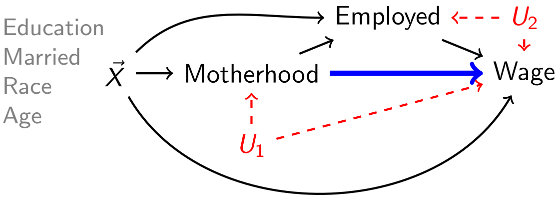
Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\ \text{Employed}\end{pmatrix}$



Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\ \text{Employed}\end{pmatrix}$



Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\ \text{Employed}\end{pmatrix}$



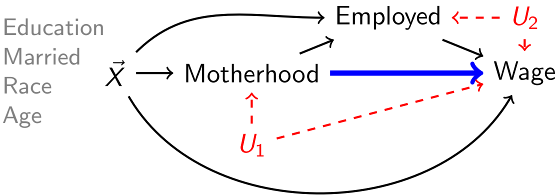
Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\ \text{Employed}\end{pmatrix}$

Potential outcomes

$Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$

All women
with a high school degree,
married, white, age 30

Focus on
one $\vec{X} = \vec{x}_i$



$$\mathbb{E}\left(Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}\,\middle|\,\vec{X}=\vec{x}_i\right)$$

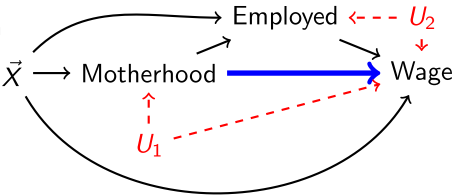
Potential outcomes

$Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$

All women
with a high school degree,
married, white, age 30

Focus on
one $\vec{X}=\vec{x}_i$

Education
Married
Race
Age



$$\mathbb{E}\left(Y_i\begin{pmatrix}\text{Mother,}\\\text{Employed}\end{pmatrix}\,\middle|\,\vec{X}=\vec{x}_i\right)$$

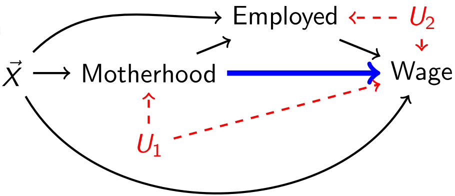
Potential outcomes

$Y_i\begin{pmatrix}\text{Mother,}\\\text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\\text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\\text{Employed}\end{pmatrix}$

All women
with a high school degree,
married, white, age 30

Focus on
one $\vec{X}=\vec{x}_i$

Education
Married
Race
Age



$$\mathbb{E}\left(Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix} \middle| \vec{X}=\vec{x}_i\right) \quad ? \quad \mathbb{E}\left(Y_i \middle| \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,}\\ \text{Employment} & = & \text{Employed,}\\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Potential outcomes

$Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$

All women
with a high school degree,
married, white, age 30

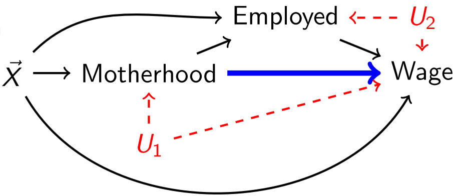
Realized outcomes

Y_i Y_i

Employed mothers
with a high school degree,
married, white, age 30

Focus on one $\vec{X} = \vec{x}_i$

Education
Married
Race
Age



$$\mathbb{E}\left(Y_i\left(\begin{array}{c}\text{Mother,}\\ \text{Employed}\end{array}\right) \,\middle|\, \vec{X}=\vec{x}_i\right) \underset{\text{By the DAG}}{=} \mathbb{E}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,}\\ \text{Employment} & = & \text{Employed,}\\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i\end{array}\right)$$

Potential outcomes

$Y_i\left(\begin{array}{c}\text{Mother,}\\ \text{Employed}\end{array}\right)$ $Y_i\left(\begin{array}{c}\text{Mother,}\\ \text{Employed}\end{array}\right)$ $Y_i\left(\begin{array}{c}\text{Mother,}\\ \text{Employed}\end{array}\right)$

All women
with a high school degree,
married, white, age 30

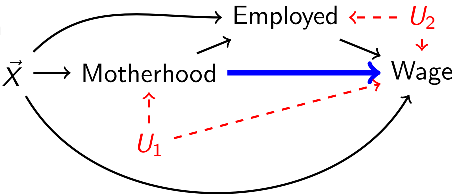
Realized outcomes

Y_i Y_i

Employed mothers
with a high school degree,
married, white, age 30

Focus on one $\vec{X}=\vec{x}_i$

Education
Married
Race
Age



$$\mathbb{E}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Realized outcomes

Y_i

Y_i

Employed mothers
with a high school degree,
married, white, age 30

This can be estimated by **machine learning**! $\longrightarrow$ $\mathbb{E}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$

Realized outcomes

Y_i

Y_i

Employed mothers
with a high school degree,
married, white, age 30

This can be estimated by **machine learning**!

Any prediction algorithm that minimizes squared errors

$$\mathbb{E}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Realized outcomes

Y_i

Y_i

Employed mothers
with a high school degree,
married, white, age 30

This can be estimated by **machine learning**!

$$\mathbb{E}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Any prediction algorithm that minimizes squared errors

Generalized Additive Model

Realized outcomes

Y_i

Y_i

Employed mothers
with a high school degree, married, white, age 30

This can be estimated by **machine learning**!

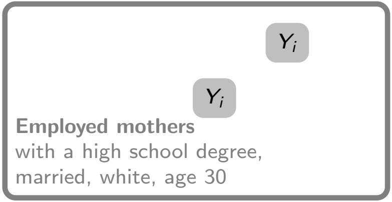
$$\mathbb{E}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Any prediction algorithm that minimizes squared errors

Generalized Additive Model

Random Forest

Realized outcomes



This can be estimated by **machine learning**!

$$\mathbb{E}\left(Y_i \;\middle|\; \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Any prediction algorithm that minimizes squared errors

Ordinary Least Squares

Generalized Additive Model

Random Forest

Realized outcomes

Y_i

Y_i

Employed mothers with a high school degree, married, white, age 30

Mechanics: How **predictive algorithms** estimate the **estimand**

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

$$\hat{\mathbb{E}}\left(Y_i \,\middle|\, \begin{array}{lcr} \text{Motherhood} & = & \texttt{Mother}, \\ \text{Employment} & = & \texttt{Employed}, \\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i \end{array}\right)$$

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

$$\hat{Y}_i\begin{pmatrix}\texttt{Mother},\\ \texttt{Employed}\end{pmatrix} = \hat{\mathbb{E}}\left(Y_i \;\middle|\; \begin{array}{lcr}\text{Motherhood} & = & \texttt{Mother},\\ \text{Employment} & = & \texttt{Employed},\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

$$\hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix} = \hat{\mathbb{E}}\left(Y_i \,\middle|\, \begin{array}{rcl}\text{Motherhood} & = & \texttt{Non-mother},\\ \text{Employment} & = & \texttt{Employed},\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

$$\hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix} = \hat{\mathbb{E}}\left(Y_i \,\middle|\, \begin{array}{rcc}\text{Motherhood} & = & \texttt{Non-mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

$$\hat{Y}_i\begin{pmatrix}\texttt{Mother,}\\ \texttt{Employed}\end{pmatrix} - \hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix}$$

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

$$\hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix} = \hat{\mathbb{E}}\left(Y_i \,\middle|\, \begin{array}{lcc}\text{Motherhood} & = & \texttt{Non-mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

3) Average over the target population

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i\begin{pmatrix}\texttt{Mother,}\\ \texttt{Employed}\end{pmatrix} - \hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix}\right)$$

Mechanics: How **predictive algorithms** estimate the **estimand**

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

$$\hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix} = \hat{\mathbb{E}}\left(Y_i \,\middle|\, \begin{array}{lcr}\text{Motherhood} & = & \texttt{Non-mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

3) Average over the target population

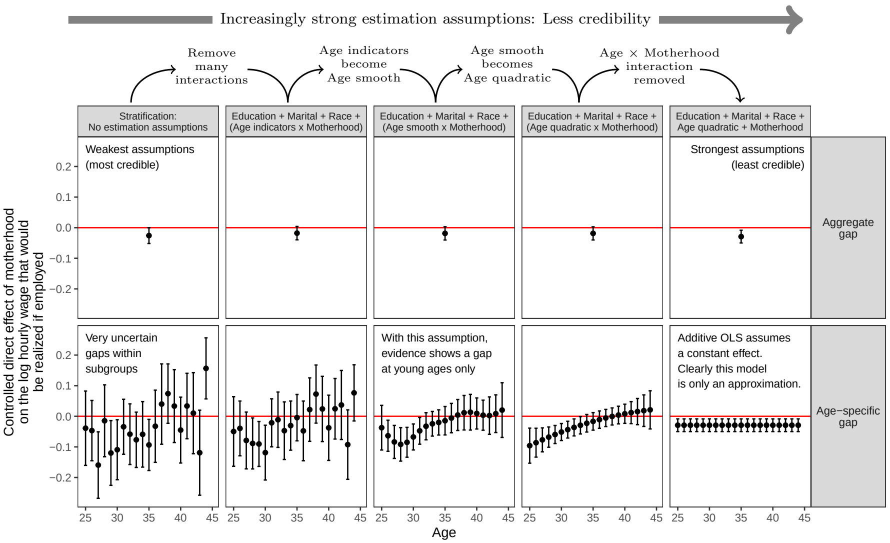
$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i\begin{pmatrix}\texttt{Mother,}\\ \texttt{Employed}\end{pmatrix} - \hat{Y}_i\begin{pmatrix}\texttt{Non-mother,}\\ \texttt{Employed}\end{pmatrix}\right)$$

This is called an **imputation estimator**

Hahn, 1998

Abadie & Imbens 2006

Also called the parametric g-formula in biostatistics, Hernán & Robins 2020


Increasingly strong estimation assumptions: Less credibility
Remove many interactions
Age indicators become Age smooth
Age smooth becomes Age quadratic
Age × Motherhood interaction removed
Stratification: No estimation assumptions
Education + Marital + Race + (Age indicators x Motherhood)
Education + Marital + Race + (Age smooth x Motherhood)
Education + Marital + Race + (Age quadratic x Motherhood)
Education + Marital + Race + Age quadratic + Motherhood
Weakest assumptions (most credible)
Strongest assumptions (least credible)
Aggregate gap
Very uncertain gaps within subgroups
With this assumption, evidence shows a gap at young ages only
Additive OLS assumes a constant effect. Clearly this model is only an approximation.
Age−specific gap
Controlled direct effect of motherhood on the log hourly wage that would be realized if employed
Age
0.2
0.1
0.0
−0.1
−0.2
25
30
35
40
45

Missing Data

Missing Data

- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!

Missing Data

- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!

Missing Data

- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!

Missing Data

- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!
- Missing data has an analogous set of assumptions and nomenclature to causal inference.

Missing Data

- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!
- Missing data has an analogous set of assumptions and nomenclature to causal inference.
- In this class we will be relatively easy going about it, but be very cautious of high levels of missingness or where selection into missingness feels highly correlated with the outcome or treatment.

Missing Data

- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!
- Missing data has an analogous set of assumptions and nomenclature to causal inference.
- In this class we will be relatively easy going about it, but be very cautious of high levels of missingness or where selection into missingness feels highly correlated with the outcome or treatment.
- Simple approaches based on imputation aren't guaranteed to work without further assumptions the same way prediction isn't guaranteed to work for causal inference!

Kleinberg et. al

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released
  - Problem: we don't have Failure to Appear (FTA) for people who are detained.

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released
  - Problem: we don't have Failure to Appear (FTA) for people who are detained.
- An interesting and complex case. It is "easier" because one counterfactual is known—if the defendant is detained they cannot fail to appear.

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released
  - Problem: we don't have Failure to Appear (FTA) for people who are detained.
- An interesting and complex case. It is "easier" because one counterfactual is known—if the defendant is detained they cannot fail to appear.
- Three requirements: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.

It is telling that in our application, much of the work happens after the prediction function was estimated.

Data and Process

Data and Process

- 1.4 million cases between November 2008 and November 2013

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and “Pure Hold-out”

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and "Pure Hold-out"
- Pure Hold-out: all of some judges, last six months, some from all remaining judges

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and “Pure Hold-out”
- Pure Hold-out: all of some judges, last six months, some from all remaining judges
- Train Gradient-Boosted Trees, 5-fold CV to select hyper parameters, retrain on all data afterwards.

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and "Pure Hold-out"
- Pure Hold-out: all of some judges, last six months, some from all remaining judges
- Train Gradient-Boosted Trees, 5-fold CV to select hyper parameters, retrain on all data afterwards.
- Don't train on any demographic other than age. Does this help? hurt?

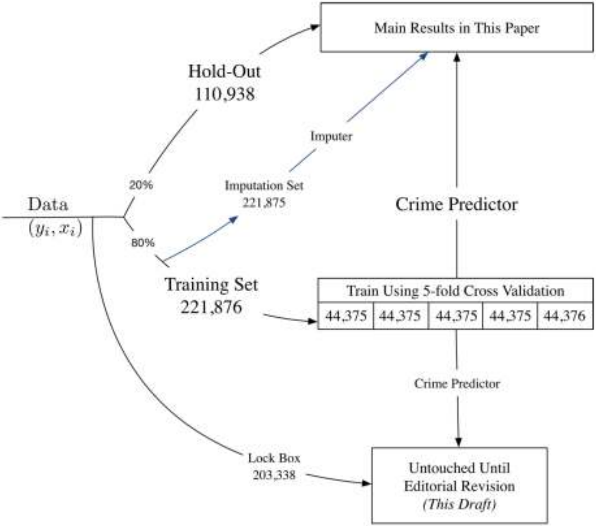


FIGURE I

Partition of New York City Data (2008–13) into Data Sets Used for Prediction and Evaluation

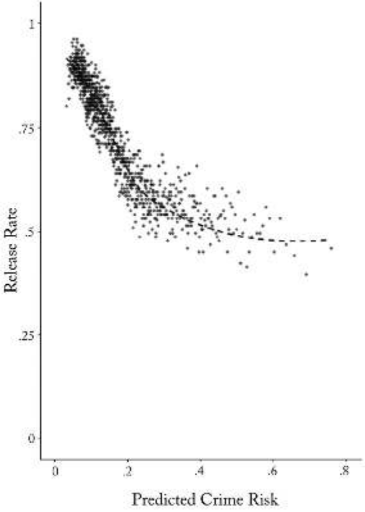


FIGURE II

How Machine Predictions of Crime Risk Relate to Judge Release Decisions and Actual Crime Rates

TABLE II

COMPARING LOGISTIC REGRESSION TO MACHINE-LEARNING PREDICTIONS OF CRIME RISK

Predicted risk percentile	ML/ logit overlap	Average observed crime rate for cases identified as high risk by: Both ML and logit	ML only	Logit only	All ML cases	All logit cases
1%	30.6%	0.6080	0.5440	0.3996	0.5636	0.4633
		(0.0309)	(0.0209)	(0.0206)	(0.0173)	(0.0174)
5%	59.9%	0.4826	0.4090	0.3040	0.4531	0.4111
		(0.0101)	(0.0121)	(0.0114)	(0.0078)	(0.0077)
10%	65.9%	0.4134	0.3466	0.2532	0.3907	0.3589
		(0.0067)	(0.0090)	(0.0082)	(0.0054)	(0.0053)
25%	72.9%	0.3271	0.2445	0.1608	0.3048	0.2821
		(0.0038)	(0.0058)	(0.0049)	(0.0032)	(0.0031)

Notes. The table shows the results of fitting a machine-learning (ML) algorithm or a logistic regression to our training data set, to identify the highest-risk observations in our test set. Both machine-learning and logistic regression models are trained on the outcome of failure to appear (FTA). Each row presents statistics for the top part of the predicted risk distribution indicated in the first column: top 25% ($N = 20{,}423$); 10% (8,173); 5% (4,087); and 1% (818). The second column shows the share of cases in the top X% of the predicted risk distribution that overlap between the set identified by ML and the set identified by logistic regression. The subsequent columns report the average crime rate observed among the released defendants within the top X% of the predicted risk distribution as identified by both ML and logit, ML only, and logit only, and all top X% identified by ML (whether or not they are also identified by logistic regression) and top X% identified by logit.

5. In practice algorithms would be decision aids, not decision makers. Our calculations simply highlight the scope of the potential gains. Understanding the determinants of compliance with prediction tools is beyond the scope of this article, though recent work has begun to focus on it (Dietvorst, Simmons, and Masey 2015; Yeomans et al. 2016; Logg 2017).

28. This logistic regression uses the same set of covariates as are given to the machine-learning algorithm, using a standard linear additive functional form. Interestingly, estimating a model that adds all two-way interactions between the predictors causes the logistic regression model to overfit the training data set and predict poorly out of sample, even in a pared-down model with just 5 or 10 covariates. Many of these interactions seem to be picking up statistical noise in the training set, although the machine-learning algorithm is able to find the interactivity that reflects real signal.

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.
- Key takeaway: even for a relatively straightforward problem, there is a lot to think about!

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.
- Key takeaway: even for a relatively straightforward problem, there is a lot to think about!
- Hopefully it is clear why the next third of the class is data—huge considerations that arise from how we came to observe the data we had.

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.
- Key takeaway: even for a relatively straightforward problem, there is a lot to think about!
- Hopefully it is clear why the next third of the class is data—huge considerations that arise from how we came to observe the data we had.
- Big data alone is not sufficient we have to understand the task and the process!

Sidenote: Missing Data

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.
- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.
- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.
- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.
- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!
- Missing data has an analogous set of assumptions and nomenclature to causal inference.

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.
- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!
- Missing data has an analogous set of assumptions and nomenclature to causal inference.
- In this class we will be relatively easy going about it, but be very cautious of high levels of missingness or where selection into missingness feels highly correlated with the outcome or treatment.

Sidenote: Missing Data

- In Kleinberg et al. the core problem is selective missing data.
- It is often said that causal inference is a missing data problem, but missing data is also a causal inference problem!
- This is why I punted on it earlier in the semester—formal approaches to missing data are really hard!
- In many cases the missingness may be relatively benign (the selection process isn't always as strong as it is with treatments) but there are no guarantees!
- Missing data has an analogous set of assumptions and nomenclature to causal inference.
- In this class we will be relatively easy going about it, but be very cautious of high levels of missingness or where selection into missingness feels highly correlated with the outcome or treatment.
- Simple approaches based on imputation aren't guaranteed to work without further assumptions the same way prediction isn't guaranteed to work for causal inference!

This Week in Review

This Week in Review

- Prediction in a counterfactual population

This Week in Review

- Prediction in a counterfactual population

Next Week:

- Salganik Chapter 4 of *Bit by Bit*: "Running Experiments"

The Core Idea: Linear/Logistic Regression with Adaptive Non-Linear Basis Functions

The Core Idea: Linear/Logistic Regression with Adaptive Non-Linear Basis Functions

$$\hat{y} = \hat{\beta}_0 + \sum_{m=1}^{M} \hat{\beta}_m \hat{\phi}_m(X)$$

Behind the Hype

Behind the Hype

- appear in earnest in the mid 1980's

Behind the Hype

- appear in earnest in the mid 1980's
- universal approximator

Behind the Hype

- appear in earnest in the mid 1980's
- universal approximator
- diminishing use in 1990's

Behind the Hype

- appear in earnest in the mid 1980's
- universal approximator
- diminishing use in 1990's
- reborn in $\approx$2010 as deep learning

High Level Summary

High Level Summary

- "just" a nonlinear model

High Level Summary

- "just" a nonlinear model
- tons of knobs to tune

High Level Summary

- “just” a nonlinear model
- tons of knobs to tune
- very powerful, but very finicky

High Level Summary

- "just" a nonlinear model
- tons of knobs to tune
- very powerful, but very finicky
- impressive performance on complex data that can defy intuition

Single Layer Feedforward Neural Network

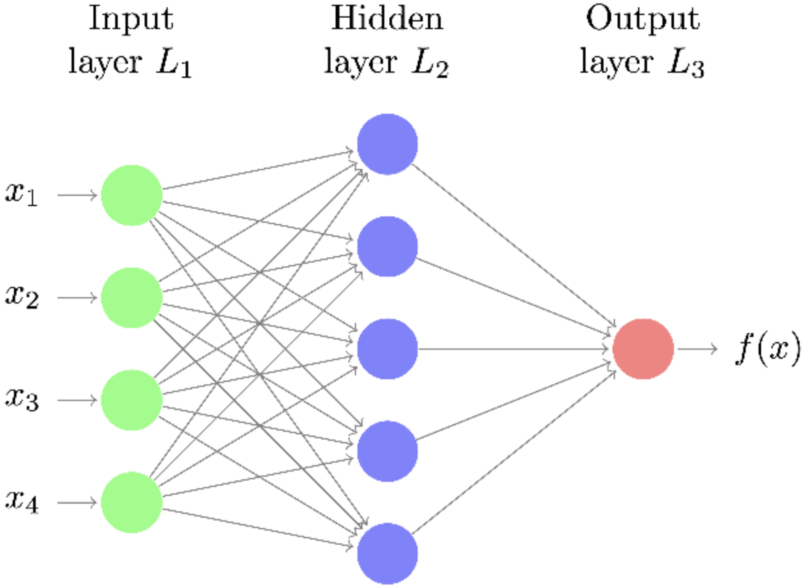


Image Credit: Efron and Hastie *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*

Single Layer Feedforward Neural Network

$$z_{i,\ell} \;=\; g\left(w_0^{(1)} + \sum_{j=1}^{4} w_{\ell,j}^{(1)} x_j\right)$$

Single Layer Feedforward Neural Network

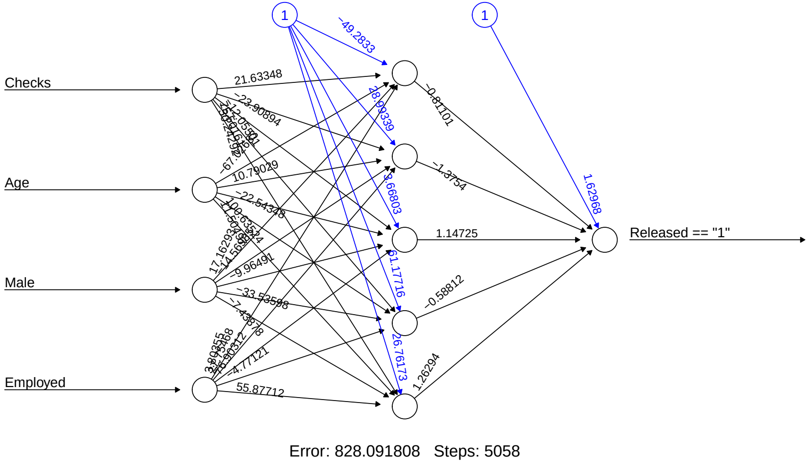
$$
\begin{aligned}
z_{i,\ell} &= g\left(w_0^{(1)} + \sum_{j=1}^{4} w_{\ell,j}^{(1)} x_j\right) \\
\hat{Y}_i &= h\left(w_0^{(2)} + \sum_{\ell=1}^{5} w_{\ell,j}^{(2)} z_{i,\ell}\right)
\end{aligned}
$$

Single Layer Feedforward Neural Network

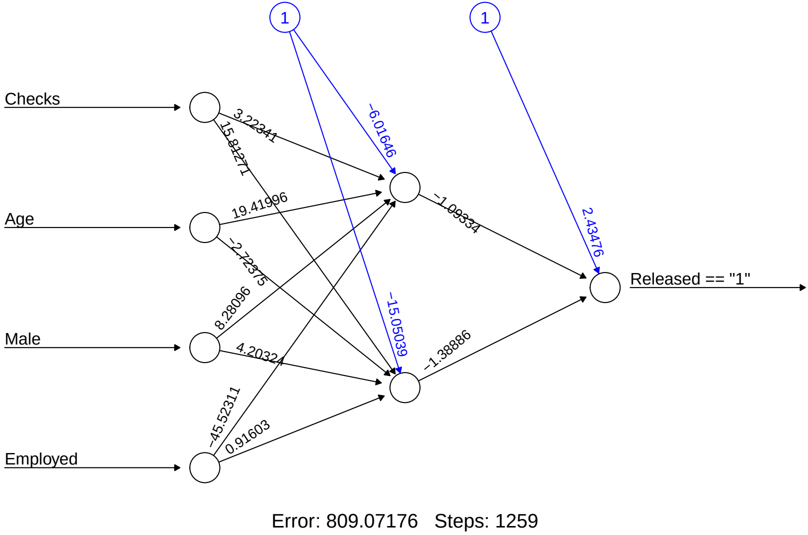
$$
\begin{aligned}
z_{i,\ell} &= g\left(w_0^{(1)} + \sum_{j=1}^{4} w_{\ell,j}^{(1)} x_j\right) \\
\hat{Y}_i &= h\left(w_0^{(2)} + \sum_{\ell=1}^{5} w_{\ell,j}^{(2)} z_{i,\ell}\right)
\end{aligned}
$$

$g()$ must be non-linear (called the activation function), $h()$ will generally be the identity, logistic, or softmax.

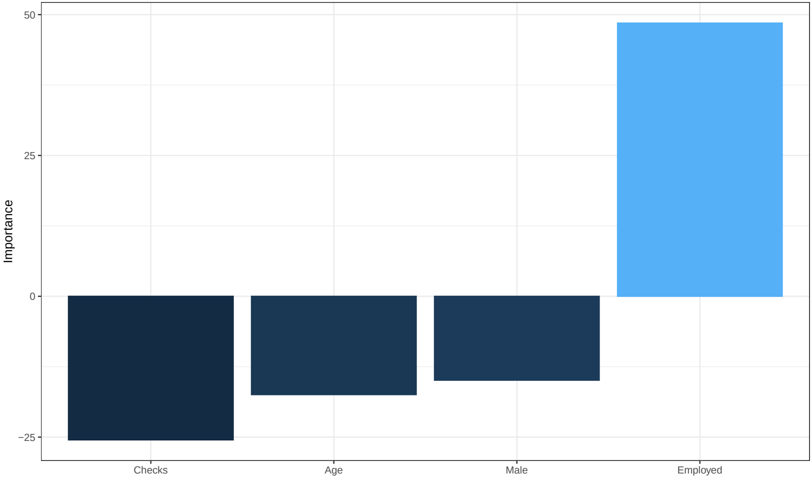
Single Layer Feedforward Neural Network



Single Layer Feedforward Neural Network



Single Layer Feedforward Neural Network



Optimization

- Much harder than other models!

Optimization

- Much harder than other models!
- Stochastic gradient descent! (including fancier variants)

Optimization

- Much harder than other models!
- Stochastic gradient descent! (including fancier variants)
- Careful initialization needed.

Optimization

- Much harder than other models!
- Stochastic gradient descent! (including fancier variants)
- Careful initialization needed.
- All handled by modern software packages.

Mini-batch Stochastic Gradient Descent

- Core Idea: update each step with a small piece of the data rather than the full thing.

Mini-batch Stochastic Gradient Descent

- Core Idea: update each step with a small piece of the data rather than the full thing.
- An epoch is a sequence of steps which account for one full pass through the data.

Mini-batch Stochastic Gradient Descent

- Core Idea: update each step with a small piece of the data rather than the full thing.
- An epoch is a sequence of steps which account for one full pass through the data.
- So if we have 100K data points and batch sizes of 100, there are 1000 steps per epoch.

Mini-batch Stochastic Gradient Descent

- Core Idea: update each step with a small piece of the data rather than the full thing.
- An epoch is a sequence of steps which account for one full pass through the data.
- So if we have 100K data points and batch sizes of 100, there are 1000 steps per epoch.
- Sometimes convenient to specify train time in epochs rather than steps.

Why Do We Need a Non-linearity?

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

$$\hat{Y} = w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right)$$

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

What happens if $g()$ is the identity function?

$$\hat{Y} = w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right)$$

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

What happens if $g()$ is the identity function?

$$
\begin{aligned}
\hat{Y} &= w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right) \\
&= \left(w_0^{(2)} + w_1^{(2)} w_0^{(1)}\right) + w_1^{(2)} w_1^{(1)} x_1 + w_1^{(2)} w_2^{(1)} x_2
\end{aligned}
$$

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

What happens if $g()$ is the identity function?

$$
\begin{aligned}
\hat{Y} &= w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right) \\
&= \left(w_0^{(2)} + w_1^{(2)} w_0^{(1)}\right) + w_1^{(2)} w_1^{(1)} x_1 + w_1^{(2)} w_2^{(1)} x_2
\end{aligned}
$$

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

What happens if $g()$ is the identity function?

$$
\begin{aligned}
\hat{Y} &= w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right) \\
&= \left(w_0^{(2)} + w_1^{(2)} w_0^{(1)}\right) + w_1^{(2)} w_1^{(1)} x_1 + w_1^{(2)} w_2^{(1)} x_2 \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2
\end{aligned}
$$

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

What happens if $g()$ is the identity function?

$$
\begin{aligned}
\hat{Y} &= w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right) \\
&= \left(w_0^{(2)} + w_1^{(2)} w_0^{(1)}\right) + w_1^{(2)} w_1^{(1)} x_1 + w_1^{(2)} w_2^{(1)} x_2 \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2
\end{aligned}
$$

Why Do We Need a Non-linearity?

Imagine a simple model with 2 covariates, 1 hidden node in 1 layer for a regression problem.

What happens if $g()$ is the identity function?

$$
\begin{aligned}
\hat{Y} &= w_0^{(2)} + w_1^{(2)} g\left(w_0^{(1)} + w_1^{(1)} x_1 + w_2^{(1)} x_2\right) \\
&= \left(w_0^{(2)} + w_1^{(2)} w_0^{(1)}\right) + w_1^{(2)} w_1^{(1)} x_1 + w_1^{(2)} w_2^{(1)} x_2 \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2
\end{aligned}
$$

It reduces to a linear model!

Does It Have to Be That Non-linearity?

Does It Have to Be That Non-linearity?

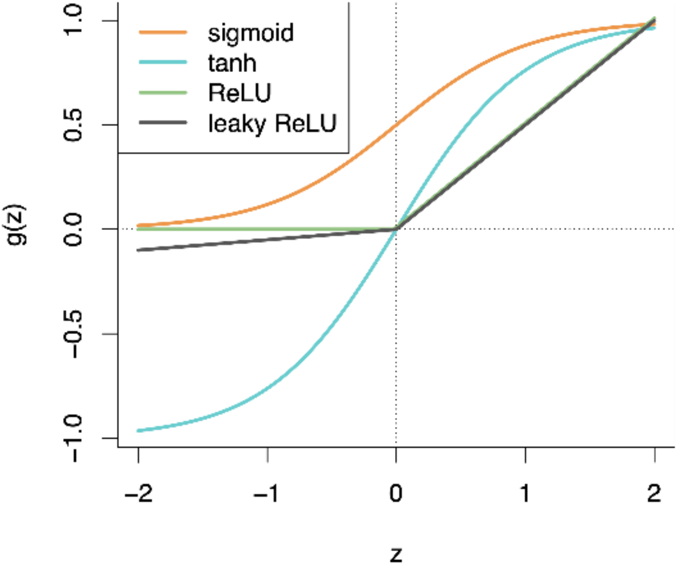


Image Credit: Efron and Hastie *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*

Regularization

Many variants of regularization

Regularization

Many variants of regularization

- "Weight Decay" (aka ridge penalty)

Regularization

Many variants of regularization

- "Weight Decay" (aka ridge penalty)
- Weight Sharing

Regularization

Many variants of regularization

- "Weight Decay" (aka ridge penalty)
- Weight Sharing
- Early Stopping

Regularization

Many variants of regularization

- "Weight Decay" (aka ridge penalty)
- Weight Sharing
- Early Stopping
- Data Augmentation

Regularization

Many variants of regularization

- "Weight Decay" (aka ridge penalty)
- Weight Sharing
- Early Stopping
- Data Augmentation
- Dropout

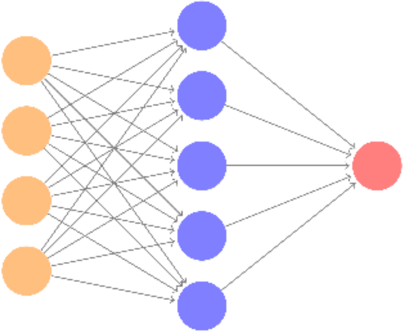
Regularization

Many variants of regularization

- "Weight Decay" (aka ridge penalty)
- Weight Sharing
- Early Stopping
- Data Augmentation
- Dropout

Some are approximately equivalent to ridge penalty.

Dropout



Dropout

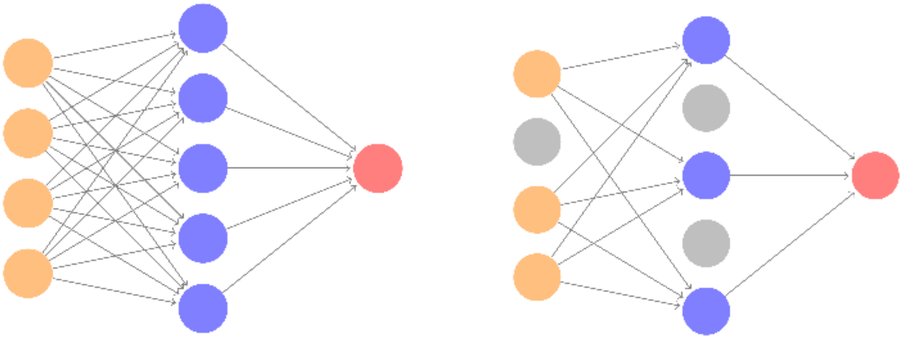


Image Credit: G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning, with applications in R: Second Edition*

Implementation

Implementation

- `neuralnet`: simple neural networks from before deep learning

Implementation

- `neuralnet`: simple neural networks from before deep learning
- `keras`: a frontend for Tensorflow

Implementation

- `neuralnet`: simple neural networks from before deep learning
- `keras`: a frontend for Tensorflow
- `torch` and `torchvision`: an R version of PyTorch

Implementation in torch

Implementation in torch

1) Standardize the inputs.

Implementation in `torch`

1) Standardize the inputs.
2) Use `nn_module()` to define the model structure.

Implementation in `torch`

1) Standardize the inputs.
2) Use `nn_module()` to define the model structure.
3) `setup` for the loss, optimizer, and reporting.

Implementation in `torch`

1) Standardize the inputs.
2) Use `nn_module()` to define the model structure.
3) `setup` for the loss, optimizer, and reporting.
4) `set_hparams` to define inputs to initialize.

Implementation in `torch`

1) Standardize the inputs.
2) Use `nn_module()` to define the model structure.
3) `setup` for the loss, optimizer, and reporting.
4) `set_hparams` to define inputs to initialize.
5) `fit` to actually estimate the model.

Implementation in `torch`

1) Standardize the inputs.
2) Use `nn_module()` to define the model structure.
3) `setup` for the loss, optimizer, and reporting.
4) `set_hparams` to define inputs to initialize.
5) `fit` to actually estimate the model.

Two important notes:

- returns pointers vs. regular R objects
- we use fast settings to decrease computation time

1) Standardize the Inputs

```
x <- scale(model.matrix(Salary ~ . - 1, data = Gitters))
y <- Gitters$Salary
```

2) Define Model Structure

```
modnn <- nn_module(
  initialize = function(input_size) {
    self$hidden <- nn_linear(input_size, 50)
    self$activation <- nn_relu()
    self$dropout <- nn_dropout(0.4)
    self$output <- nn_linear(50, 1)
  },
  forward = function(x) {
    x %>%
      self$hidden() %>%
      self$activation() %>%
      self$dropout() %>%
      self$output()
  }
)
```

3-4) Define Other Model Components

```
modnn <- modnn %>%
  setup(
    loss = nn_mse_loss(),
    optimizer = optim_rmsprop,
    metrics = list(luz_metric_mae())
  ) %>%
  set_hparams(input_size = ncol(x))
```

5) Fit the Model

```
fitted <- modnn %>%
  fit(
    data = list(x[-testid, ],
    matrix(y[-testid], ncol = 1)),
    valid_data = list(x[testid, ],
    matrix(y[testid], ncol = 1)),
    epochs = 20
  )
```

What Can We Do With This?

- Call `predict(fitted, newdata = )` to get predictions.

What Can We Do With This?

- Call `predict(fitted, newdata = )` to get predictions.
- New! Use `as_array()` to convert the pointer to a regular R object!

What Can We Do With This?

- Call `predict(fitted, newdata = )` to get predictions.
- New! Use `as_array()` to convert the pointer to a regular R object!
- Use `plot(fitted)` to show error metrics during optimization.

The Core Idea

- Including many layers in the neural network can be effective.

The Core Idea

- Including many layers in the neural network can be effective.
- Narrowing the number of hidden nodes as layers get further from data.

The Core Idea

- Including many layers in the neural network can be effective.
- Narrowing the number of hidden nodes as layers get further from data.
- Many innovations to make the above happen!

The Core Idea

- Including many layers in the neural network can be effective.
- Narrowing the number of hidden nodes as layers get further from data.
- Many innovations to make the above happen!
- When does this work?

The Core Idea

- Including many layers in the neural network can be effective.
- Narrowing the number of hidden nodes as layers get further from data.
- Many innovations to make the above happen!
- When does this work?
  complex data and high signal-to-noise ratio.

Multiple Layer Feedforward Neural Network

Multiple Layer Feedforward Neural Network

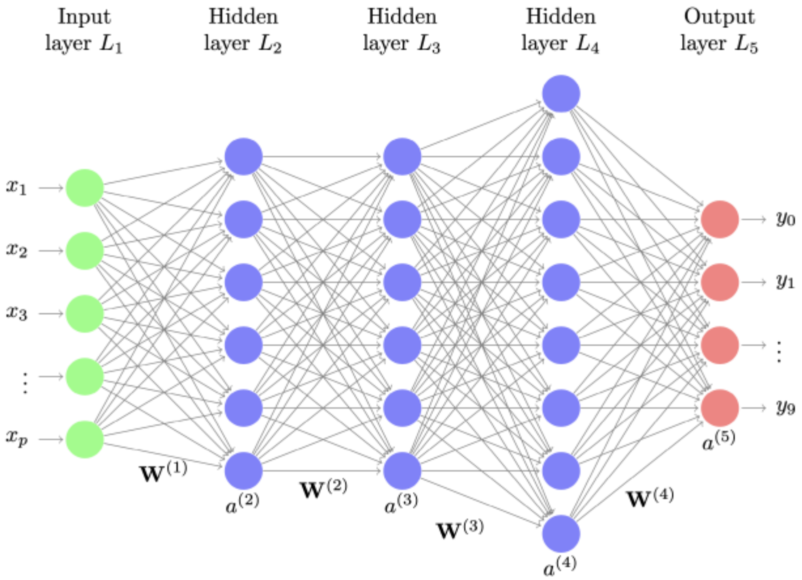
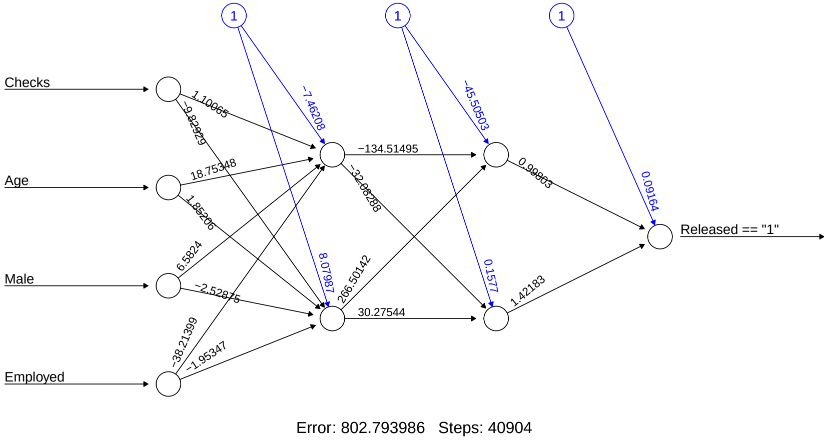


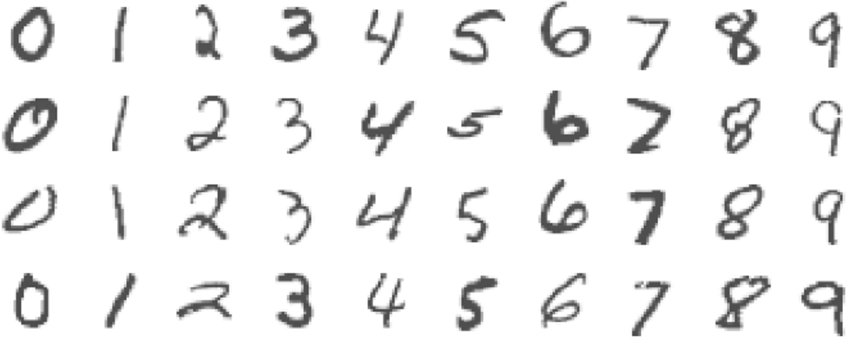
Image Credit: Efron and Hastie *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*

Multiple Layer Feedforward Neural Network

Multiple Layer Feedforward Neural Network



Example: MNIST Digits



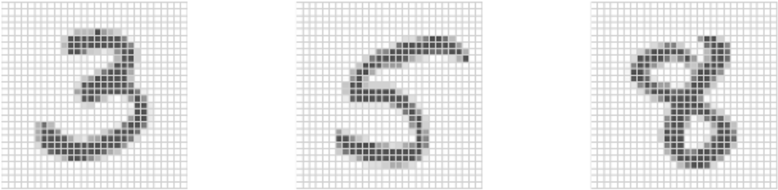


Image Credit: G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning, with applications in R: Second Edition*

Implementation using torch

```
M_NN_struct <- nn_module(
  initialize = function() {
      self$linear1 <- nn_linear(in_features = 28*28, out_features = 256)
      self$linear2 <- nn_linear(in_features = 256, out_features = 128)
      self$linear3 <- nn_linear(in_features = 128, out_features = 10)

      self$drop1 <- nn_dropout(p = 0.4)
      self$drop2 <- nn_dropout(p = 0.3)

      self$activation <- nn_relu()
  },
  forward = function(x) {
     x %>%
     self$linear1() %>%
     self$activation() %>%
     self$drop1() %>%
     self$linear2() %>%
     self$activation() %>%
     self$drop2() %>%
     self$linear3()
}
)
```

Implementation using torch

```
modelnn <- modelnn %>%
  setup(
    loss = nn_cross_entropy_loss(),
    optimizer = optim_rmsprop,
    metrics = list(luz_metric_accuracy())
  )
```

Alternative Architectures

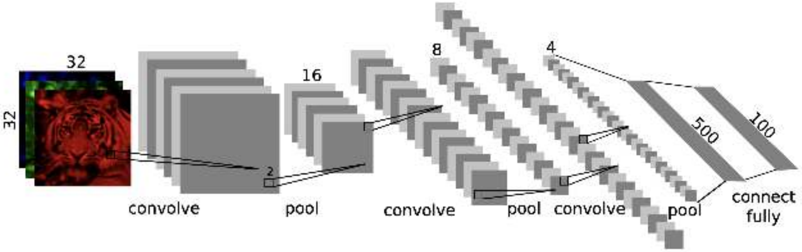


Image Credit: Efron and Hastie *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*

Convolutional Neural Networks (CNNs)

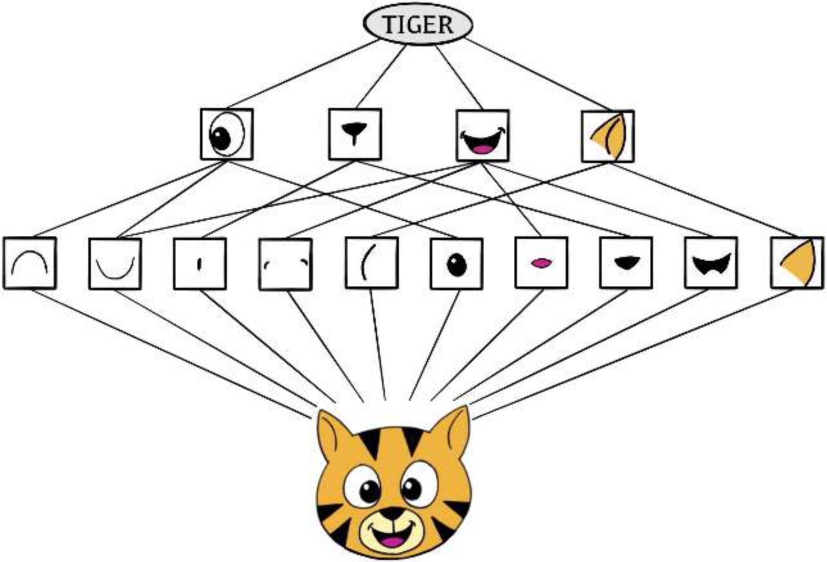
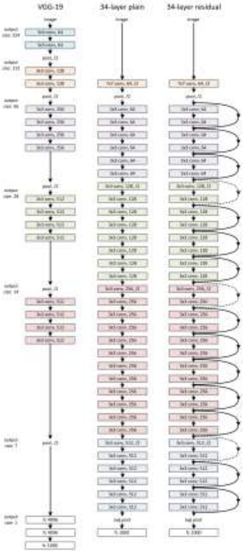


Image Credit: G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning, with applications in R: Second Edition*

Resnet Architecture



He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

A Story About Architecture

A Story About Architecture

Turing Award Won by 3 Pioneers in Artificial Intelligence

From left, Yann LeCun, Geoffrey Hinton and Yoshua Bengio. The researchers worked on key developments for neural networks, which are reshaping how computer systems are built. From left, Facebook, via Associated Press; Aaron Vincent Elkaim for The New York Times; Chad Buchanan/Getty Images

But Also a Story About Data

But Also a Story About Data

Fei-Fei Li

From Wikipedia, the free encyclopedia

The native form of this personal name is Li Feifei. *This article uses Western name order when mentioning individuals.*

Fei-Fei Li (simplified Chinese: 李飞飞; traditional Chinese: 李飛飛; born 1976) is a Chinese-born American computer scientist, non-profit executive, and writer. She is the Sequoia Capital Professor of Computer Science at Stanford University.[2] Li is a Co-Director of the Stanford Institute for Human-Centered Artificial Intelligence, and a Co-Director of the Stanford Vision and Learning Lab.[3][4] She served as the director of the Stanford Artificial Intelligence Laboratory (SAIL)[5] from 2013 to 2018. In 2017, she co-founded AI4ALL, a nonprofit organization working to increase diversity and inclusion in the field of artificial intelligence.[6][7] Her research expertise includes artificial intelligence (AI), machine learning, deep learning, computer vision and cognitive neuroscience.[8] She was the leading scientist and principal investigator of ImageNet.[9]

Fei-Fei Li

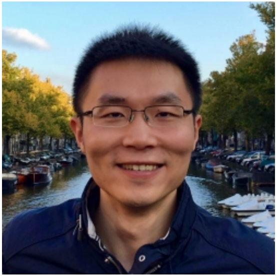
Li at AI for Good in 2017

Born	1976 (age 44–45)[1] Beijing
Nationality	American
Alma mater	Princeton University (B.A. in Physics)

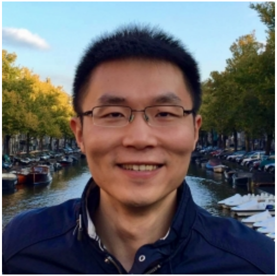
Princeton Connections!

Princeton Connections!

Princeton Connections!



Princeton Connections!



Tuning Parameters

Tuning Parameters

- number of layers

Tuning Parameters

- number of layers
- number of nodes per layer

Tuning Parameters

- number of layers
- number of nodes per layer
- regularization

Tuning Parameters

- number of layers
- number of nodes per layer
- regularization
- optimization technique

Tuning Parameters

- number of layers
- number of nodes per layer
- regularization
- optimization technique
- architecture

Powerful but difficult to use

Powerful but difficult to use

If you have ever tried to train a NN on your own, then you may know how painful and uncertain it can be. I am not talking about following the existing tutorials and demos, when all hyperparameters are already tuned for you, but about taking some data and creating something from scratch. Even with modern DL [deep learning] high-level toolkits, where all best practices such as proper weights initialization and optimizers' betas, gammas, and other options are set to sane defaults, and tons of other stuff hidden under the hood, there are still lots of decisions that you can make, hence lots of things can go wrong. As a result, your network almost never works from the first run and this is something that you should get used to.

(Lapan 2018, 66)

Summary

Summary

- neural networks

Summary

- neural networks
- deep learning as an extension of basic neural networks

Summary

- neural networks
- deep learning as an extension of basic neural networks

Going Deeper:

G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning, with applications in R: Second Edition* Chapter 10

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression					
GAMs					
Ridge/Lasso					
Dec. Trees					
Rand. Forests					
Boosting					
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs					
Ridge/Lasso					
Dec. Trees					
Rand. Forests					
Boosting					
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs	Y				
Ridge/Lasso					
Dec. Trees					
Rand. Forests					
Boosting					
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs	Y				
Ridge/Lasso	N				
Dec. Trees					
Rand. Forests					
Boosting					
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs	Y				
Ridge/Lasso	N				
Dec. Trees	Y				
Rand. Forests					
Boosting					
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs	Y				
Ridge/Lasso	N				
Dec. Trees	Y				
Rand. Forests	Y				
Boosting					
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs	Y				
Ridge/Lasso	N				
Dec. Trees	Y				
Rand. Forests	Y				
Boosting	Y				
Neural Nets					

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N				
GAMs	Y				
Ridge/Lasso	N				
Dec. Trees	Y				
Rand. Forests	Y				
Boosting	Y				
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y				
Ridge/Lasso	N				
Dec. Trees	Y				
Rand. Forests	Y				
Boosting	Y				
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y	N			
Ridge/Lasso	N				
Dec. Trees	Y				
Rand. Forests	Y				
Boosting	Y				
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y	N			
Ridge/Lasso	N	N			
Dec. Trees	Y				
Rand. Forests	Y				
Boosting	Y				
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y	N			
Ridge/Lasso	N	N			
Dec. Trees	Y	Y			
Rand. Forests	Y				
Boosting	Y				
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y	N			
Ridge/Lasso	N	N			
Dec. Trees	Y	Y			
Rand. Forests	Y	Y			
Boosting	Y				
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y	N			
Ridge/Lasso	N	N			
Dec. Trees	Y	Y			
Rand. Forests	Y	Y			
Boosting	Y	Y			
Neural Nets	Y				

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N			
GAMs	Y	N			
Ridge/Lasso	N	N			
Dec. Trees	Y	Y			
Rand. Forests	Y	Y			
Boosting	Y	Y			
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N			
Ridge/Lasso	N	N			
Dec. Trees	Y	Y			
Rand. Forests	Y	Y			
Boosting	Y	Y			
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N	Low		
Ridge/Lasso	N	N			
Dec. Trees	Y	Y			
Rand. Forests	Y	Y			
Boosting	Y	Y			
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N	Low		
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y			
Rand. Forests	Y	Y			
Boosting	Y	Y			
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N	Low		
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y			
Boosting	Y	Y			
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N	Low		
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y			
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N	Low		
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y	High		
Neural Nets	Y	Y			

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None		
GAMs	Y	N	Low		
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y	High		
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low		
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y	High		
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low		
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y	High		
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low		
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y	High		
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium		
Boosting	Y	Y	High		
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High		
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High		

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	N
Ridge/Lasso	N	N	Low	Y	
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	N
Ridge/Lasso	N	N	Low	Y	N/Y
Dec. Trees	Y	Y	Low	N	
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	N
Ridge/Lasso	N	N	Low	Y	N/Y
Dec. Trees	Y	Y	Low	N	Y
Rand. Forests	Y	Y	Medium	N	
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	N
Ridge/Lasso	N	N	Low	Y	N/Y
Dec. Trees	Y	Y	Low	N	Y
Rand. Forests	Y	Y	Medium	N	$\sim$N
Boosting	Y	Y	High	N	
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	N
Ridge/Lasso	N	N	Low	Y	N/Y
Dec. Trees	Y	Y	Low	N	Y
Rand. Forests	Y	Y	Medium	N	$\sim$N
Boosting	Y	Y	High	N	$\sim$N
Neural Nets	Y	Y	Very High	Y	

Review of Methods

	Nonlinear	Interactions	Tuning	Smooth	Var Sel
Regression	N	N	None	Y	N
GAMs	Y	N	Low	Y	N
Ridge/Lasso	N	N	Low	Y	N/Y
Dec. Trees	Y	Y	Low	N	Y
Rand. Forests	Y	Y	Medium	N	$\sim$N
Boosting	Y	Y	High	N	$\sim$N
Neural Nets	Y	Y	Very High	Y	N

Things to Consider When Picking A Method

Things to Consider When Picking A Method

- How hard is the problem?

Things to Consider When Picking A Method

- How hard is the problem?
- How much data do I have?

Things to Consider When Picking A Method

- How hard is the problem?
- How much data do I have?
- How many predictors do you have? How Informative are they?

Things to Consider When Picking A Method

- How hard is the problem?
- How much data do I have?
- How many predictors do you have? How Informative are they?
- How much prior knowledge do you have?

Things to Consider When Picking A Method

- How hard is the problem?
- How much data do I have?
- How many predictors do you have? How Informative are they?
- How much prior knowledge do you have?
- How smooth is the output function?

Things to Consider When Picking A Method

- How hard is the problem?
- How much data do I have?
- How many predictors do you have? How Informative are they?
- How much prior knowledge do you have?
- How smooth is the output function?
- Who do I have to explain this too?

The Big Assumption: Ignorability

Identification by randomization:

The Big Assumption: Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)

The Big Assumption: Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)
- Randomization therefore means treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, i.e.

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

The Big Assumption: Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)
- Randomization therefore means treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, i.e.

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

- This is sometimes called unconfoundedness or ignorability

The Big Assumption: Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)
- Randomization therefore means treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, i.e.

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

- This is sometimes called unconfoundedness or ignorability
- Another way of thinking of it: The distributions of the potential outcomes $(Y_i(1),\ Y_i(0))$ are the same for the treatment and control group.

The Big Assumption: Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)
- Randomization therefore means treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, i.e.

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

- This is sometimes called unconfoundedness or ignorability
- Another way of thinking of it: The distributions of the potential outcomes $(Y_i(1), Y_i(0))$ are the same for the treatment and control group.
- Yet another way of thinking of it: The treatment and control group are exchangeable, or balanced (on observables and unobservables) on average

How do we proceed?

Identification by conditional independence:

- If treatment is <u>not randomized</u>, then treatment may be related underlying characteristics, observed and unobserved, which are related to the potential outcomes

How do we proceed?

Identification by conditional independence:

- If treatment is not randomized, then treatment may be related underlying characteristics, observed and unobserved, which are related to the potential outcomes
- Therefore, we need to assume that treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, conditional on some pre-treatment characteristics X, i.e.

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

How do we proceed?

Identification by conditional independence:

- If treatment is not randomized, then treatment may be related underlying characteristics, observed and unobserved, which are related to the potential outcomes
- Therefore, we need to assume that treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, conditional on some pre-treatment characteristics X, i.e.

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

- Conditioning set should yield $Y_i(0), Y_i(1)$ and T_i conditionally independent. (This is next week's topic).

How do we proceed?

Identification by conditional independence:

- If treatment is not randomized, then treatment may be related underlying characteristics, observed and unobserved, which are related to the potential outcomes
- Therefore, we need to assume that treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, conditional on some pre-treatment characteristics X, i.e.

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

- Conditioning set should yield $Y_i(0), Y_i(1)$ and T_i conditionally independent. (This is next week's topic).
- This is conditional ignorability.

Graphical Models

Graphical Models

- A general framework for representing causal relationships based on directed acyclic graphs (DAG)

Graphical Models

- A general framework for representing causal relationships based on directed acyclic graphs (DAG)
- A way to encode assumptions about the causal structure precisely and separately from assumptions about estimation.

Graphical Models

- A general framework for representing causal relationships based on directed acyclic graphs (DAG)
- A way to encode assumptions about the causal structure precisely and separately from assumptions about estimation.
- The do-calculus and a DAG are all you need to determine whether causal quantities are identified.

Graphical Models

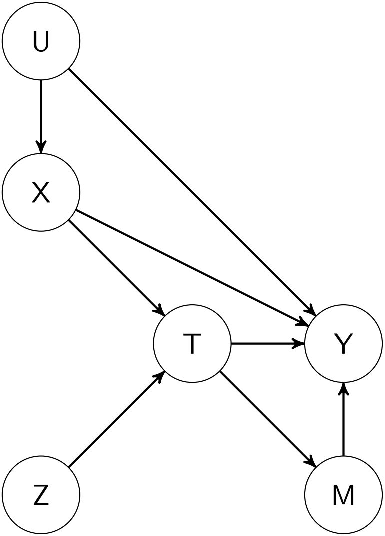
- A general framework for representing causal relationships based on directed acyclic graphs (DAG)
- A way to encode assumptions about the causal structure precisely and separately from assumptions about estimation.
- The do-calculus and a DAG are all you need to determine whether causal quantities are identified.
- Nice software that takes the graph and returns an identification strategy: DAGitty at `http://dagitty.net` (also accessible directly in `R`).

Graphical Models

- A general framework for representing causal relationships based on directed acyclic graphs (DAG)
- A way to encode assumptions about the causal structure precisely and separately from assumptions about estimation.
- The do-calculus and a DAG are all you need to determine whether causal quantities are identified.
- Nice software that takes the graph and returns an identification strategy: DAGitty at `http://dagitty.net` (also accessible directly in `R`).

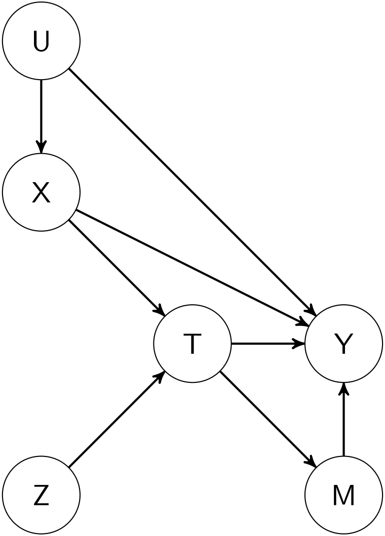
Code Idea: causal reasoning before statistical reasoning

Components of a DAG



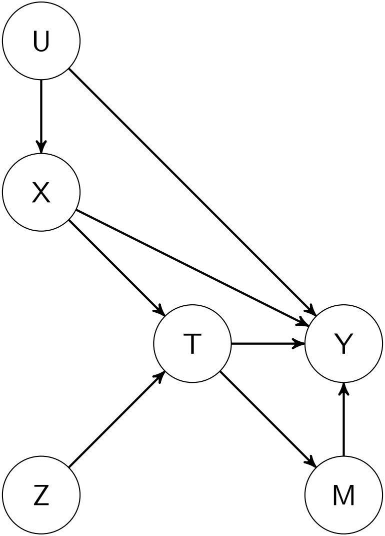
- nodes represent variables
  (unobserved typically called U or V)

Components of a DAG



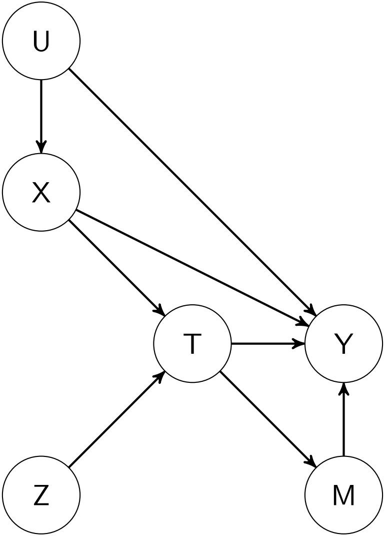
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects

Components of a DAG



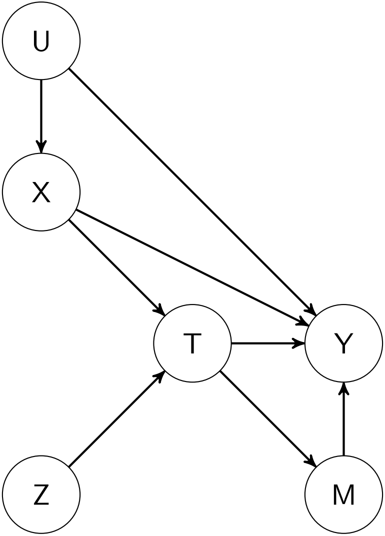
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables

Components of a DAG



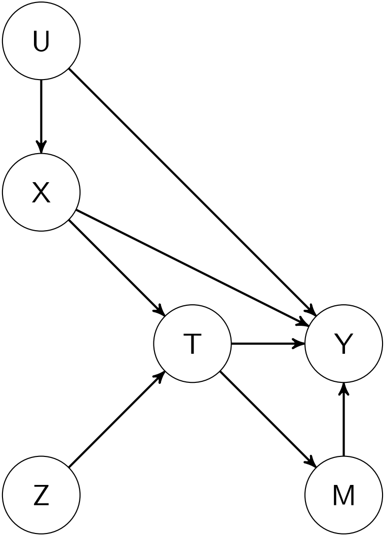
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables
- absence of arrows represents no causal effect

Components of a DAG



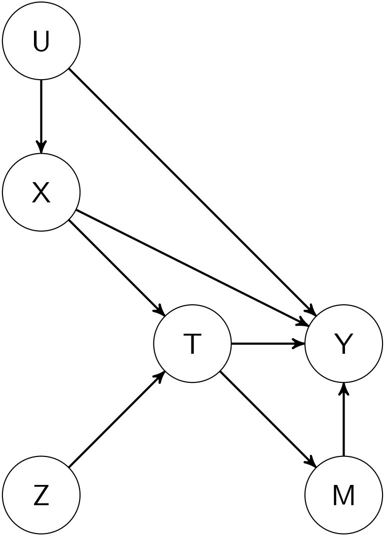
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables
- absence of arrows represents no causal effect
- positioning conveys no mathematical meaning but often is oriented left-to-right with causal ordering for readability.

Components of a DAG



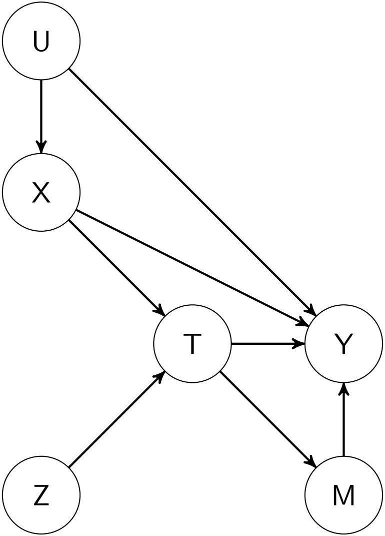
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables
- absence of arrows represents no causal effect
- positioning conveys no mathematical meaning but often is oriented left-to-right with causal ordering for readability.
- dashed lines are used in context dependent ways

Components of a DAG



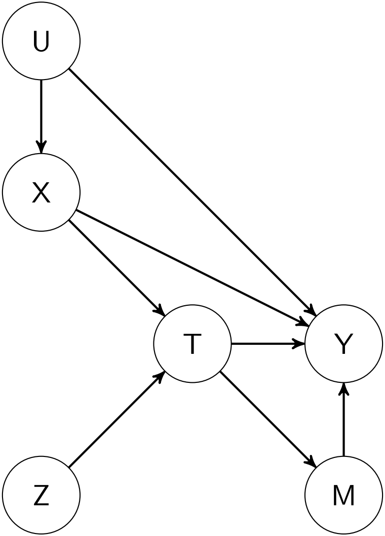
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables
- absence of arrows represents no causal effect
- positioning conveys no mathematical meaning but often is oriented left-to-right with causal ordering for readability.
- dashed lines are used in context dependent ways
- all relationships are non-parametric

Relationships in a DAG



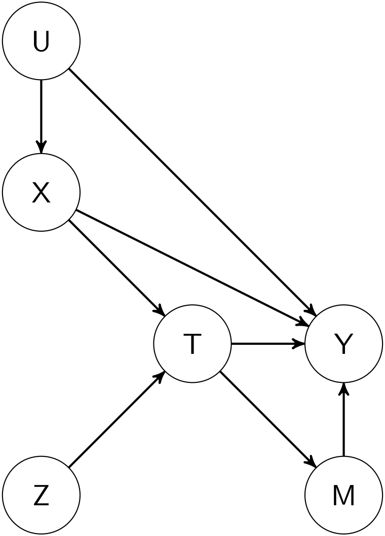
- Parents (Children): directly causing (caused by) a node

Relationships in a DAG



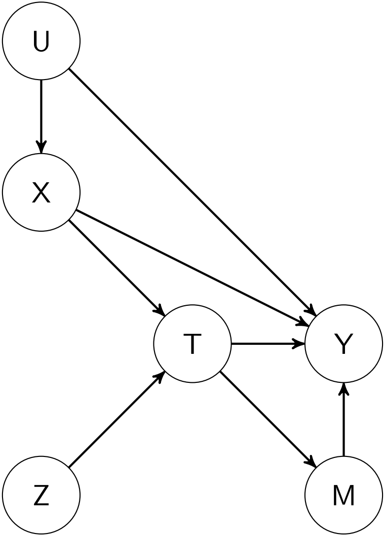
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node

Relationships in a DAG



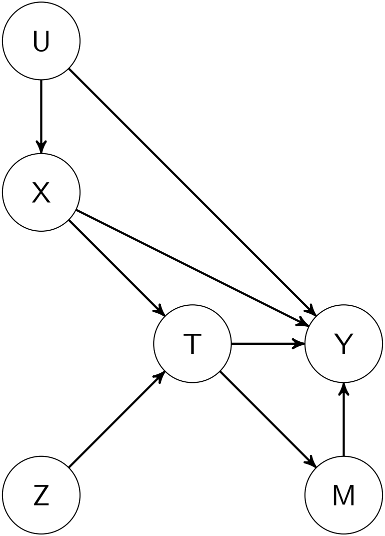
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node
- Path: a route that connects the variables (path is causal when all arrows point the same way)

Relationships in a DAG



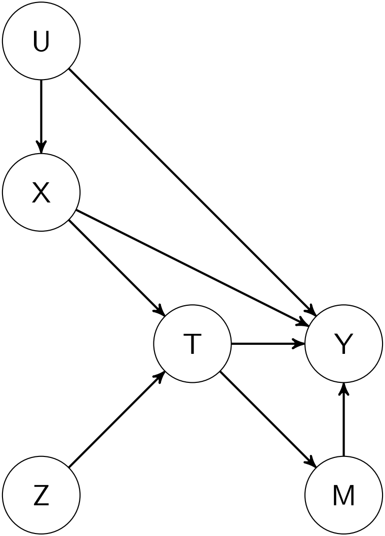
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node
- Path: a route that connects the variables (path is causal when all arrows point the same way)
- Acyclic implies that there are no cycles and a variable can't cause itself

Relationships in a DAG



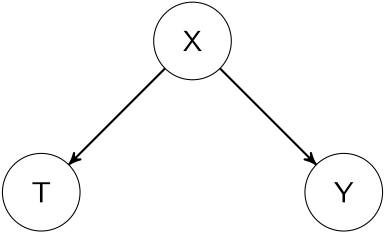
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node
- Path: a route that connects the variables (path is causal when all arrows point the same way)
- Acyclic implies that there are no cycles and a variable can't cause itself
- Causal Markov assumption: condition on its direct causes, a variable is independent of its non-descendents.

Relationships in a DAG



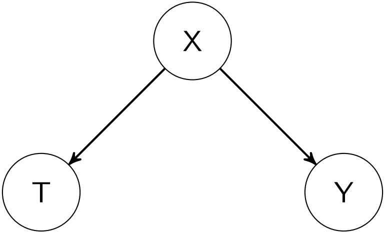
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node
- Path: a route that connects the variables (path is causal when all arrows point the same way)
- Acyclic implies that there are no cycles and a variable can't cause itself
- Causal Markov assumption: condition on its direct causes, a variable is independent of its non-descendents.
- We will talk in depth about two types of relationships: confounders and colliders

Confounders



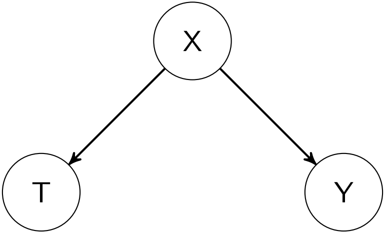
- X is a confounder (or common cause)

Confounders



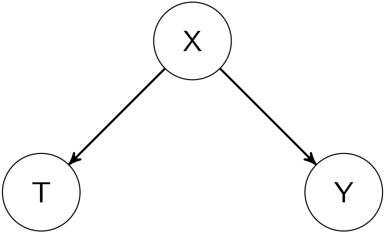
- X is a confounder (or common cause)
- Even without a causal effect or directed edge between T and Y they will have a marginal associational relationship

Confounders



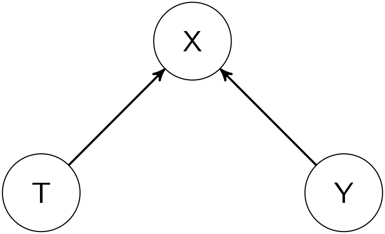
- X is a confounder (or common cause)
- Even without a causal effect or directed edge between T and Y they will have a marginal associational relationship
- Conditional on X, T and Y are unrelated in this graph.

Confounders



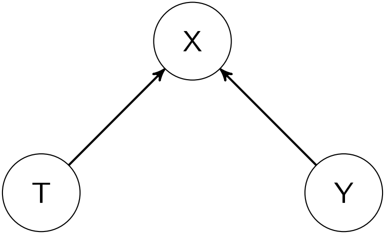
- X is a confounder (or common cause)
- Even without a causal effect or directed edge between T and Y they will have a marginal associational relationship
- Conditional on X, T and Y are unrelated in this graph.
- We can think of conditioning on a confounder as blocking the flow of association.

Colliders



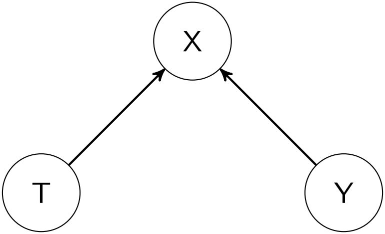
- X is now a collider because two arrows point into it

Colliders



- X is now a collider because two arrows point into it
- In this scenario T and Y are not marginally associated

Colliders



- X is now a collider because two arrows point into it
- In this scenario T and Y are not marginally associated
- If we control for X they become associated and create a connection between T and Y

Colliders are scary because you can induce dependence

ANNUAL REVIEWS **Further**

Click here for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable

Felix Elwert[1] and Christopher Winship[2]

[1]Department of Sociology, University of Wisconsin, Madison, Wisconsin 53706; email: elwert@wisc.edu

[2]Department of Sociology, Harvard University, Cambridge, Massachusetts 02138; email: cwinship@wjh.harvard.edu

Annu. Rev. Sociol. 2014. 40:31–53

First published online as a Review in Advance on June 2, 2014

The *Annual Review of Sociology* is online at soc.annualreviews.org

This article's doi: 10.1146/annurev-soc-071913-043455

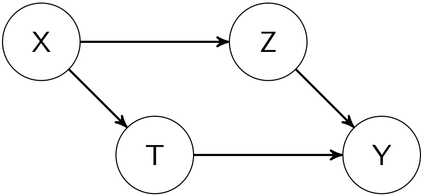
Keywords

causality, directed acyclic graphs, identification, confounding, selection

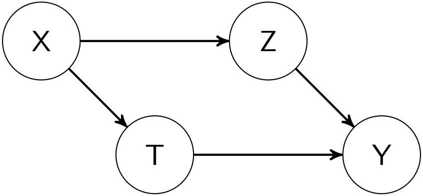
Abstract

Endogenous selection bias is a central problem for causal inference. Recognizing the problem, however, can be difficult in practice. This article introduces a purely graphical way of characterizing endogenous selection bias and of understanding its consequences (Hernán et al. 2004). We use causal graphs (direct acyclic graphs, or DAGs) to highlight that endogenous selection bias stems from conditioning (e.g., controlling, stratifying, or selecting) on a so-called collider variable, i.e., a variable that is itself caused by two other variables, one that is (or is associated with) the treatment and another that is (or is associated with) the outcome. Endogenous selection bias can result from direct conditioning on the outcome variable, a post-outcome variable, a post-treatment variable, and even a pre-treatment variable. We highlight the difference between endogenous selection bias, common-cause confounding, and overcontrol bias and discuss numerous examples from social stratification, cultural sociology, social network analysis, political sociology, social demography, and the sociology of education.

From Confounders to Back-Door Paths

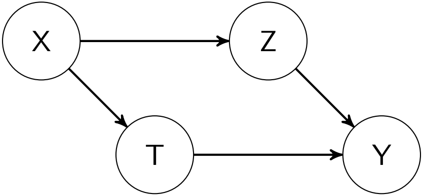


From Confounders to Back-Door Paths



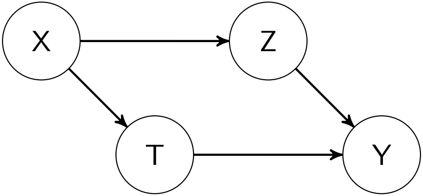
- Identify causal effect of T on Y by conditioning on X, Z or X and Z

From Confounders to Back-Door Paths



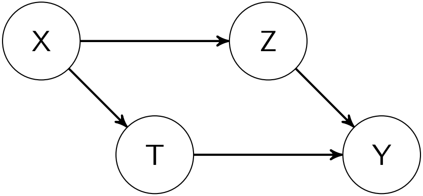
- Identify causal effect of T on Y by conditioning on X, Z or X and Z
- We can formalize this logic with the idea of a back-door path

From Confounders to Back-Door Paths



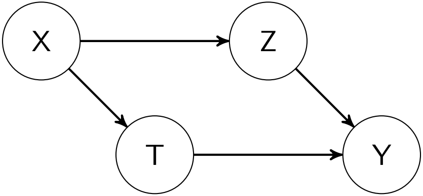
- Identify causal effect of T on Y by conditioning on X, Z or X and Z
- We can formalize this logic with the idea of a back-door path
- A back-door path is "a path between any causally ordered sequence of two variables that begins with a directed edge that points to the first variable." (Morgan and Winship 2013)

From Confounders to Back-Door Paths



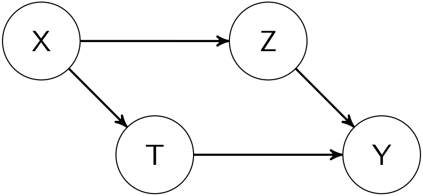
- Identify causal effect of T on Y by conditioning on X, Z or X and Z
- We can formalize this logic with the idea of a back-door path
- A back-door path is "a path between any causally ordered sequence of two variables that begins with a directed edge that points to the first variable." (Morgan and Winship 2013)
- Two paths from T to Y here:
  1. $T \rightarrow Y$ (directed or causal path)
  2. $T \leftarrow X \rightarrow Z \rightarrow Y$ (back-door path)

From Confounders to Back-Door Paths



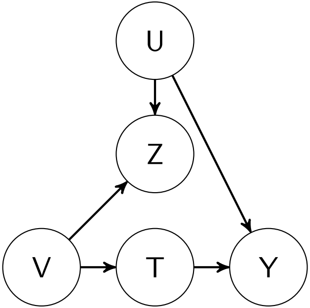
- Identify causal effect of T on Y by conditioning on X, Z or X and Z
- We can formalize this logic with the idea of a back-door path
- A back-door path is "a path between any causally ordered sequence of two variables that begins with a directed edge that points to the first variable." (Morgan and Winship 2013)
- Two paths from T to Y here:
  1. $T \rightarrow Y$ (directed or causal path)
  2. $T \leftarrow X \rightarrow Z \rightarrow Y$ (back-door path)
- Observed marginal association between T and Y is a composite of these two paths and thus does not identify the causal effect of T on Y

From Confounders to Back-Door Paths



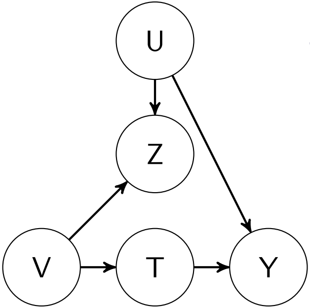
- Identify causal effect of T on Y by conditioning on X, Z or X and Z
- We can formalize this logic with the idea of a back-door path
- A back-door path is "a path between any causally ordered sequence of two variables that begins with a directed edge that points to the first variable." (Morgan and Winship 2013)
- Two paths from T to Y here:
  1. $T \rightarrow Y$ (directed or causal path)
  2. $T \leftarrow X \rightarrow Z \rightarrow Y$ (back-door path)
- Observed marginal association between T and Y is a composite of these two paths and thus does not identify the causal effect of T on Y
- We want to block the back-door path to leave only the causal effect

Colliders and Back-Door Paths



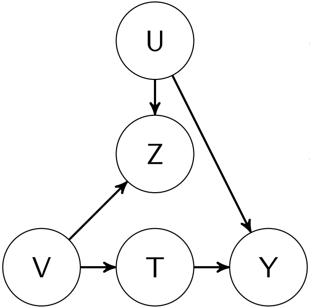
- Z is a collider and it lies along a back-door path from T to Y

Colliders and Back-Door Paths



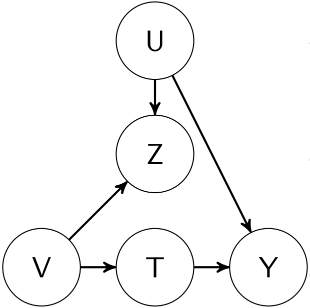
- Z is a collider and it lies along a back-door path from T to Y
- Conditioning on a collider on a back-door path does not help and in fact causes new associations

Colliders and Back-Door Paths



- Z is a collider and it lies along a back-door path from T to Y
- Conditioning on a collider on a back-door path does not help and in fact causes new associations
- Here we are fine unless we condition on Z which opens a path $T \leftarrow V \leftrightarrow U \rightarrow Y$ (this particular case is called M-bias)

Colliders and Back-Door Paths



- Z is a collider and it lies along a back-door path from T to Y
- Conditioning on a collider on a back-door path does not help and in fact causes new associations
- Here we are fine unless we condition on Z which opens a path $T \leftarrow V \leftrightarrow U \rightarrow Y$ (this particular case is called M-bias)
- So how do we know which back-door paths to block?

Blocking backdoor paths: College and earnings

What do we need to include to block all backdoor paths between college and earnings?

$$\begin{tikzcd}
 & T \arrow[r] & Y \\
X \arrow[ur] \arrow[urr] & &
\end{tikzcd}$$

Ability,

Blocking backdoor paths: College and earnings

What do we need to include to block all backdoor paths between college and earnings?

$$\begin{tikzcd}
& T \arrow[r] & Y \\
X \arrow[ur] \arrow[urr] & &
\end{tikzcd}$$

Ability, parents' income,

Blocking backdoor paths: College and earnings

What do we need to include to block all backdoor paths between college and earnings?

$$
\begin{tikzcd}
& T \arrow[r] & Y \\
X \arrow[ur] \arrow[urr] & &
\end{tikzcd}
$$

Ability, parents' income, parents' education,

Blocking backdoor paths: College and earnings

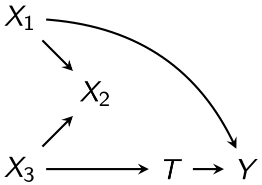
What do we need to include to block all backdoor paths between college and earnings?

$$
\begin{tikzcd}
 & T \arrow[r] & Y \\
X \arrow[ur] \arrow[urr] & &
\end{tikzcd}
$$

Ability, parents' income, parents' education, extended family who pay for college and help you find a job, neighborhood characteristics that affect high school quality and also the availability of local jobs, ... lots of things!

Non-causal paths: Part 2

Now consider this graph. Is there an unblocked backdoor path from T to Y?



Non-causal paths: Part 2

Now consider this graph. Is there an unblocked backdoor path from T to Y?

$$\begin{tikzcd}
X_1 \arrow[dr] \arrow[rrrdd, bend left] & & & \\
& X_2 & & \\
X_3 \arrow[ur] \arrow[rr] & & T \arrow[r] & Y
\end{tikzcd}$$

No need to condition! X_2 already blocks this path. it is a collider.

Colliders: Be careful!

$$\begin{tikzcd}[row sep=small] X_1 \arrow[dr] & \\ & Y \\ X_2 \arrow[ur] & \end{tikzcd}$$

Y is a collider. X_1 and X_2 are not associated, but they are when we hold Y constant.

What situations might produce this?

Colliders: Be careful!

$$\begin{tikzcd}
X_1 \arrow[dr] & \\
& Y \\
X_2 \arrow[ur] &
\end{tikzcd}$$

Y is a collider. X_1 and X_2 are not associated, but they are when we hold Y constant.

What situations might produce this?

- X_1 being in a car accident. X_2 is having cancer. Y is being in a hospital.

Colliders: Be careful!

$$\begin{tikzcd}
X_1 \arrow[dr] & \\
& Y \\
X_2 \arrow[ur] &
\end{tikzcd}$$

Y is a collider. X_1 and X_2 are not associated, but they are when we hold Y constant.

What situations might produce this?

- X_1 being in a car accident. X_2 is having cancer. Y is being in a hospital.
- X_1 is living in a warm climate. X_2 is being an elite swimmer. Y is going swimming in January.

Colliders: Be careful!

$$X_1 \rightarrow Y \leftarrow X_2$$

Y is a collider. X_1 and X_2 are not associated, but they are when we hold Y constant.

What situations might produce this?

- X_1 being in a car accident. X_2 is having cancer. Y is being in a hospital.
- X_1 is living in a warm climate. X_2 is being an elite swimmer. Y is going swimming in January.
- X_1 is family income. X_2 is religiosity. Y attendance at a Catholic high school.

Colliders: Be careful!

$$X_1 \rightarrow Y \leftarrow X_2$$

Y is a collider. X_1 and X_2 are not associated, but they are when we hold Y constant.

What situations might produce this?

- X_1 being in a car accident. X_2 is having cancer. Y is being in a hospital.
- X_1 is living in a warm climate. X_2 is being an elite swimmer. Y is going swimming in January.
- X_1 is family income. X_2 is religiosity. Y attendance at a Catholic high school.

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Hypothetical substantive question:

Does acting ability causally affect the probability of marriage?

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Hypothetical substantive question:

Does acting ability causally affect the probability of marriage?

Hypothetical approach: Estimate on a sample of Hollywood actors and actresses.

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Hypothetical substantive question:

Does acting ability causally affect the probability of marriage?

Hypothetical approach: Estimate on a sample of Hollywood actors and actresses.

We want to estimate:

Acting ability $\longrightarrow$ Marriage

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Hypothetical substantive question:

Does acting ability causally affect the probability of marriage?

Hypothetical approach: Estimate on a sample of Hollywood actors and actresses.

We want to estimate:

Acting ability $\longrightarrow$ Marriage

Should we worry about this design? It depends on our theory about how these variables are related. We can argue about identification with a DAG.

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

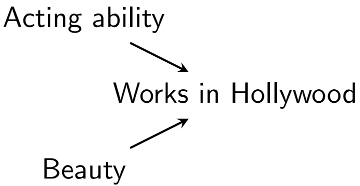
Suppose working in Hollywood is a function of two factors: acting ability and beauty. In the general population, these two are uncorrelated. However, among those who work in Hollywood, those who are bad at acting <u>must</u> be beautiful.

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Suppose working in Hollywood is a function of two factors: acting ability and beauty. In the general population, these two are uncorrelated. However, among those who work in Hollywood, those who are bad at acting must be beautiful.

True DAG

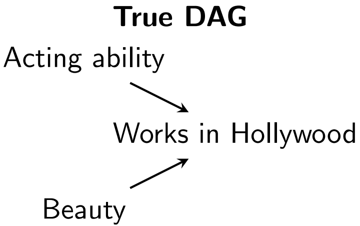


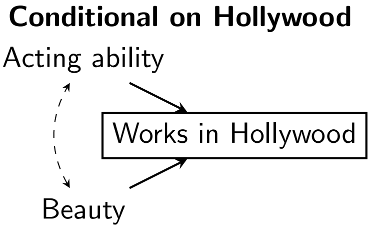
Conditional on Hollywood

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Suppose working in Hollywood is a function of two factors: acting ability and beauty. In the general population, these two are uncorrelated. However, among those who work in Hollywood, those who are bad at acting must be beautiful.

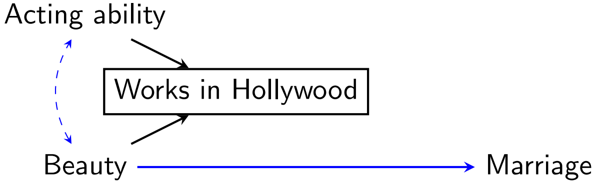




Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

This is an example of conditioning on a collider! We induce a negative association between acting ability and beauty.



Under the assumptions above, our results are driven by collider conditioning!

Thoughts on DAGs and Potential Outcomes

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a "foreign language" for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a “foreign language” for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.
- Graphical Models with DAGs are very visually appealing but the operations on the graph can be challenging

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a “foreign language” for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.
- Graphical Models with DAGs are very visually appealing but the operations on the graph can be challenging
- DAGs very helpful for thinking through identification and the entire causal process

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a “foreign language” for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.
- Graphical Models with DAGs are very visually appealing but the operations on the graph can be challenging
- DAGs very helpful for thinking through identification and the entire causal process
- Note that both are about non-parametric identification and not estimation.

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a "foreign language" for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.
- Graphical Models with DAGs are very visually appealing but the operations on the graph can be challenging
- DAGs very helpful for thinking through identification and the entire causal process
- Note that both are about non-parametric identification and not estimation.
  - Good: provides a very general framework that applies in non-linear scenarios and interactions

Thoughts on DAGs and Potential Outcomes

- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a “foreign language” for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.
- Graphical Models with DAGs are very visually appealing but the operations on the graph can be challenging
- DAGs very helpful for thinking through identification and the entire causal process
- Note that both are about non-parametric identification and not estimation.
    - Good: provides a very general framework that applies in non-linear scenarios and interactions
    - Tricky: identification results for identification only holds when variable is completely controlled for (which may be difficult!)

We Covered

We Covered

- Components of a DAG
- Confounders and Colliders

We Covered

- Components of a DAG
- Confounders and Colliders

Morgan and Winship (2014) *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.