What is Your Estimand?

Defining the Target Quantity Connects Statistical Evidence to Theory

Ian Lundberg
UCLA
Sociology
ianlundberg.org

Rebecca Johnson
Dartmouth Quantitative
Social Science
rebeccajohnson.io

Brandon M. Stewart
Princeton
Sociology
brandonstewart.org

16 November 2021. Indiana University, SOC-S 651.
Paper in *American Sociological Review*. Preprint on SocArxiv. Replication code on Dataverse. Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879.

There is a question every quantitative study must answer:

There is a question every quantitative study must answer:

What is your estimand?

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

There is a question every quantitative study must answer:

What is your estimand?

A common answer:

The **purpose** of the statistical analysis

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

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

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\underset{\substack{\uparrow \\ \beta_1 \text{ is an estimand} \\ \text{that assumes a model}}}{\beta_1} + X_2\beta_2 + \epsilon$$

There is a question every quantitative study must answer:

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?

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\underset{\uparrow}{\beta_1} + X_2\beta_2 + \epsilon$$

β_1 is an estimand that assumes a model

What if the model is wrong?

The model is an approximation

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1 \underset{\uparrow}{\beta_1} + X_2\beta_2 + \epsilon$$

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

There is a question every quantitative study must answer:

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.

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

A common answer:

— We took [data source]

— We estimated β_1

$$Y = \beta_0 + X_1\underset{\uparrow}{\beta_1} + X_2\beta_2 + \epsilon$$

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?

There is a question every quantitative study must answer:

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

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the
statistical analysis

stated **outside**
of the model

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model

A **unit-specific quantity**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model

Y_i

A **unit-specific quantity**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model

$Y_i(t)$

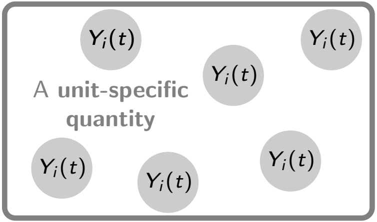
A **unit-specific quantity**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



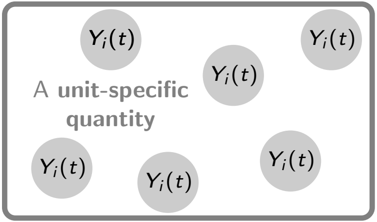
Aggregated over a **target population**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



Aggregated over a **target population**

Our framework expands **theory**,

links to transparent **evidence**,

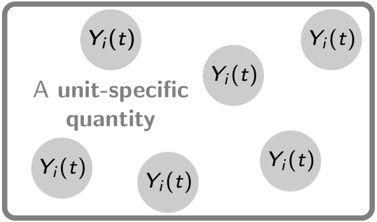
and unlocks computational **tools**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



Aggregated over a **target population**

Our framework expands **theory**,

links to transparent **evidence**,

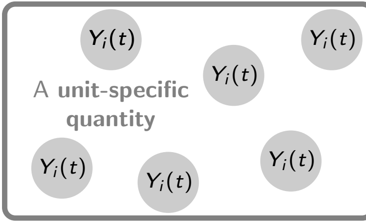
and unlocks computational **tools**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



Aggregated over a **target population**

Our framework expands **theory**,

links to transparent **evidence**,

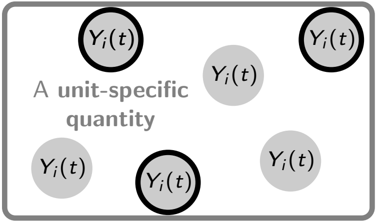
and unlocks computational **tools**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



Aggregated over a **target population**

Our framework expands **theory**,

links to transparent **evidence**,

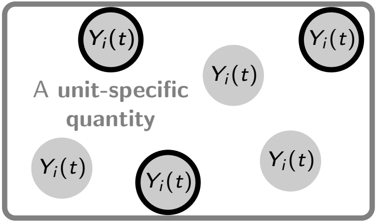
and unlocks computational **tools**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



Aggregated over a **target population**

Our framework expands **theory**,

links to transparent **evidence**,

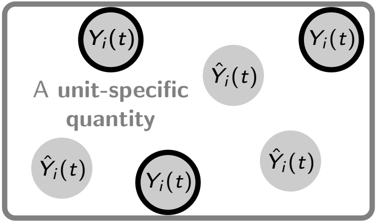
and unlocks computational **tools**

There is a question every quantitative study must answer:

What is your estimand?

The **purpose** of the statistical analysis

stated **outside** of the model



Aggregated over a **target population**

Our framework expands **theory**,

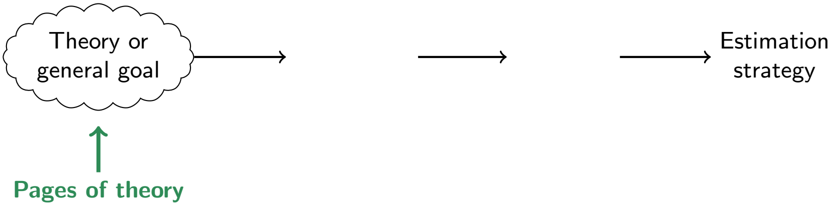
links to transparent **evidence**,

and unlocks computational **tools**

Research framework: Estimands connect theory to evidence

Theory or general goal $\longrightarrow$ $\longrightarrow$ $\longrightarrow$ Estimation strategy

Research framework: Estimands connect theory to evidence



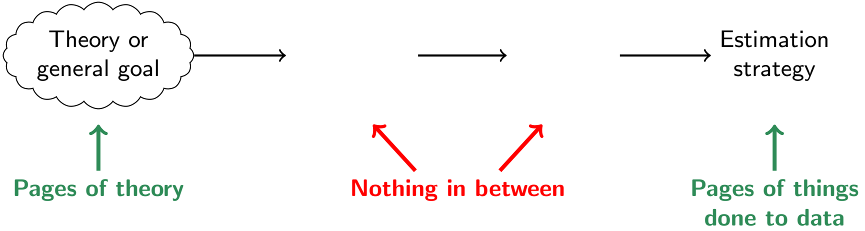
Research framework: Estimands connect theory to evidence

Theory or general goal ⟶ ⟶ ⟶ Estimation strategy

↑ **Pages of theory**

↑ **Pages of things done to data**

Research framework: Estimands connect theory to evidence



Research framework: Estimands connect theory to evidence

Theory or general goal $\longrightarrow$ Theoretical estimand $\longrightarrow$ Empirical estimand $\longrightarrow$ Estimation strategy

Research framework: Estimands connect theory to evidence

Theory or general goal ⟶ Theoretical estimand ⟶ Empirical estimand ⟶ Estimation strategy

Set a specific target

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Definition

A **unit-specific quantity** aggregated over a **target population**

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Definition	Example
A **unit-specific quantity** aggregated over a **target population**	$\frac{1}{\text{Size of U.S. adult population}} \sum_{i \text{ in U.S. adult population}} \left(\text{Employed}_i\right)$

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Definition	Example
A **unit-specific quantity** aggregated over a **target population**	$$\frac{1}{\text{Size of U.S. adult population}} \sum_{\substack{i \text{ in U.S.} \\ \text{adult population}}} \left(\underbrace{\text{Employed}_i(\text{Job training})}_{\substack{\text{Employment if received} \\ \text{job training}}} - \underbrace{\text{Employed}_i(\text{No job training})}_{\substack{\text{Employment if did not receive} \\ \text{job training}}} \right)$$

Research framework: Estimands connect theory to evidence

Theory or general goal $\longrightarrow$ Theoretical estimand $\longrightarrow$ Empirical estimand $\longrightarrow$ Estimation strategy

$\uparrow$

Set a specific target

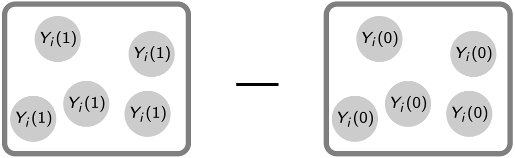
Definition	Example
A **unit-specific quantity** aggregated over a **target population**	$$\frac{1}{\text{Size of U.S. adult population}} \sum_{\substack{i \text{ in U.S.} \\ \text{adult population}}} \Bigg(\underbrace{\text{Employed}_i(\text{Job training})}_{\substack{\text{Employment if received} \\ \text{job training}}} - \underbrace{\text{Employed}_i(\text{No job training})}_{\substack{\text{Employment if did not receive} \\ \text{job training}}} \Bigg)$$

Lieberson 1987, Abbott 1988, Freedman 1991, Xie 2013, Hernán 2018

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Definition	Example
A **unit-specific quantity** aggregated over a **target population**	

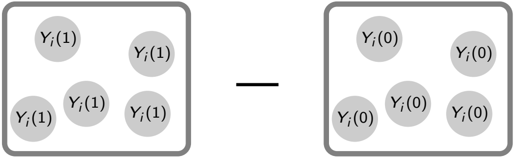
Lieberson 1987, Abbott 1988, Freedman 1991, Xie 2013, Hernán 2018

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Link to observable data

Definition	Example
A quantity involving **observable data**	

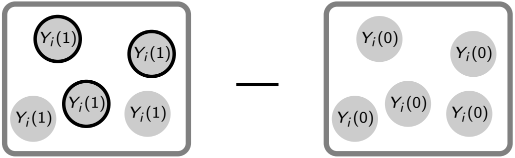
Lundberg, Johnson, and Stewart. What is Your Estimand?

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Link to observable data

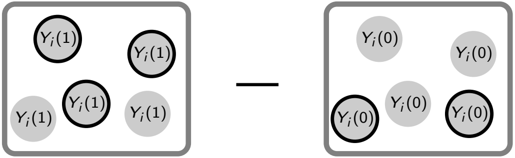
Definition	Example
A quantity involving **observable data**	

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Link to observable data

Definition	Example
A quantity involving **observable data**	

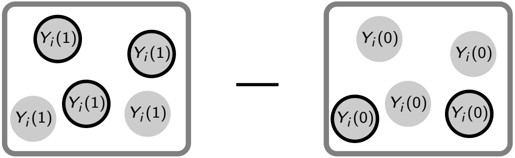
Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Link to observable data

Definition	Example
A quantity involving **observable data**	



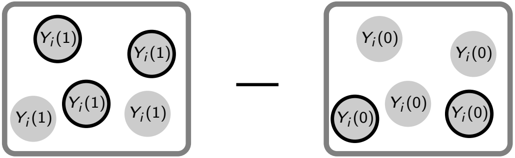
$$\begin{tikzcd} \vec{X} \arrow[r] \arrow[rr, bend right] & T \arrow[r] & Y \end{tikzcd}$$

Pearl 2009, Imbens and Rubin 2015,
Morgan and Winship 2015, Elwert and Winship 2014

Research framework: Estimands connect theory to evidence

Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target | **Link** to observable data | **Learn** from data

Definition	Example
An algorithm applied to data	

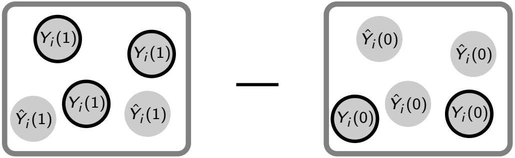
Research framework: Estimands connect theory to evidence

Theory or general goal $\longrightarrow$ Theoretical estimand $\longrightarrow$ Empirical estimand $\longrightarrow$ Estimation strategy

Set a specific target

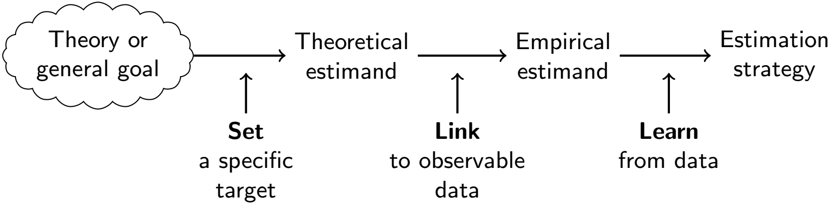
Link to observable data

Learn from data

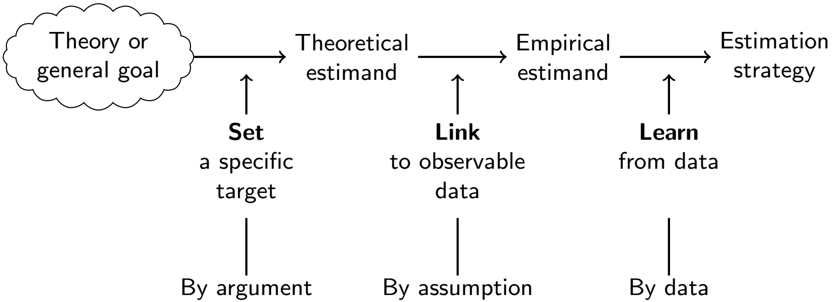
Definition	Example
An algorithm applied to data	

Young 2009, Watts 2014, Berk et al. 2019, Molina and Garip 2019

Research framework: Estimands connect theory to evidence



Research framework: Estimands connect theory to evidence



Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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]

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

> 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]

Theory or general goal → **Set** by argument → Theoretical estimand → **Link** by assumption → Empirical estimand → **Learn** by data → Estimation strategy

> 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]

TheUpshot

DATA DIVE

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

Theory or general goal → **Set** by argument → Theoretical estimand → **Link** by assumption → Empirical estimand → **Learn** by data → Estimation strategy

> 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


TheUpshot

DATA DIVE

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

OPINION | COMMENTARY

The Myth of Systemic Police Racism

WSJ | OPINION

Hold officers accountable who use excessive force. But there's no evidence of widespread racial bias.

By Heather Mac Donald
June 2, 2020 1:44 pm ET

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

> 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]

TheUpshot

DATA DIVE

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

OPINION | COMMENTARY

The Myth of Systemic Police Racism

Hold officers accountable who use excessive force. But there's no evidence of widespread racial bias.

WSJ | OPINION

By Heather Mac Donald
June 2, 2020 1:44 pm ET

Reality check: study finds no racial bias in police shootings

The Guardian

Lundberg, Johnson, and Stewart. What is Your Estimand?

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

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

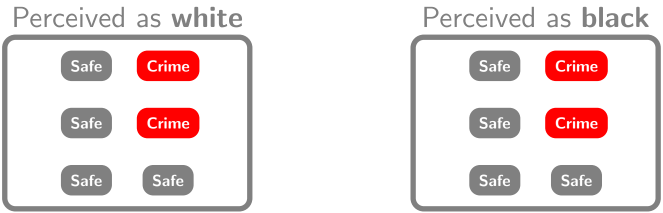
Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

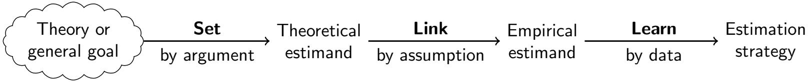
Claim: Police are unbiased

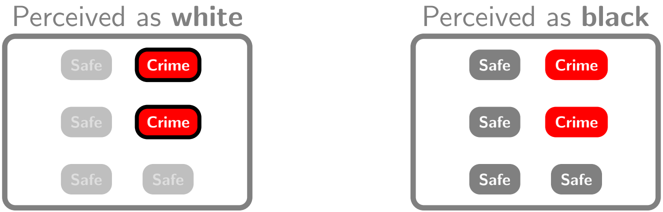
Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Theory or general goal → **Set** by argument → Theoretical estimand → **Link** by assumption → Empirical estimand → **Learn** by data → Estimation strategy

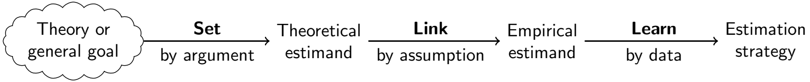


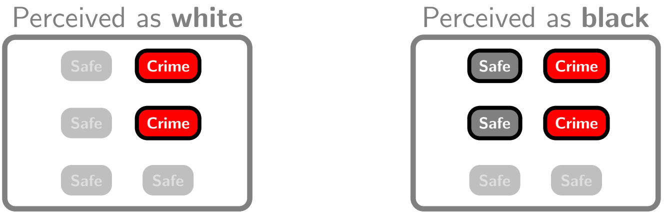
Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.



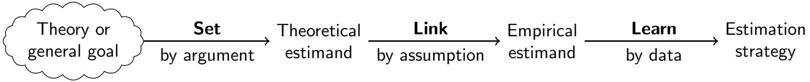


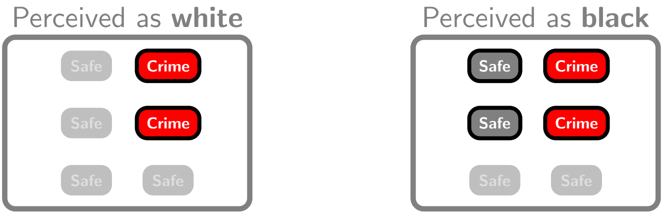
Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

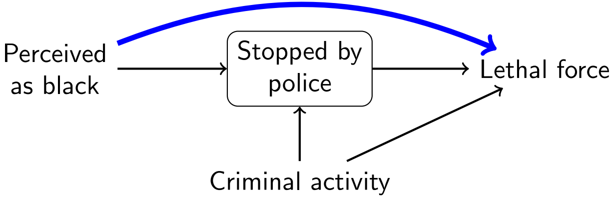




Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.







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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Claim: Police are unbiased

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Claim: Police are unbiased

Fryer responds:

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

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

Fryer responds:

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

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

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]."

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Is the theoretical estimand descriptive?

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Is the theoretical estimand descriptive?

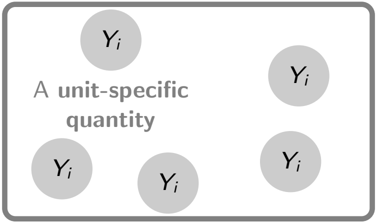
> "the differential in hourly wages
> between women with children
> and women without children"

Theory or general goal —**Set** by argument→ Theoretical estimand —**Link** by assumption→ Empirical estimand —**Learn** by data→ Estimation strategy

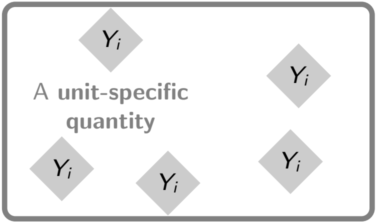
Pal and Waldfogel (2016) estimate the family gap in pay.

Is the theoretical estimand descriptive?

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



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



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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Is the theoretical estimand descriptive? Is it causal?

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Is the theoretical estimand descriptive? Is it causal?

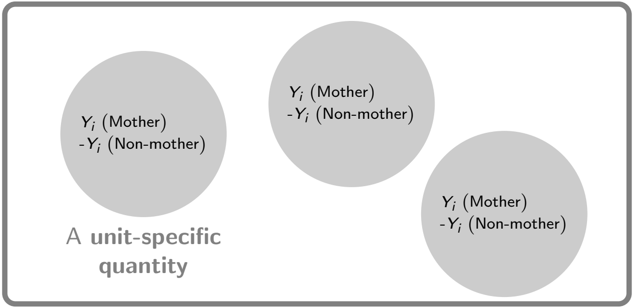
"causal estimation techniques"

Theory or general goal —**Set** by argument→ Theoretical estimand —**Link** by assumption→ Empirical estimand —**Learn** by data→ Estimation strategy

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

Is the theoretical estimand descriptive? Is it causal?

"causal estimation techniques"



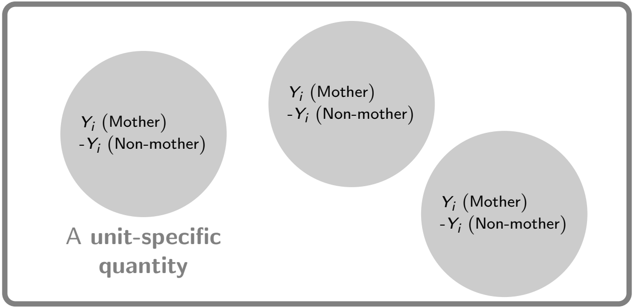
Averaged over a **target population** of **mothers**

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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.



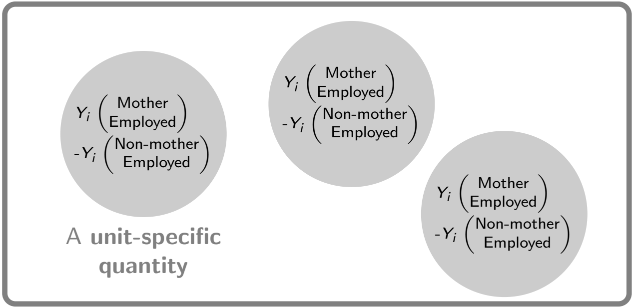
Averaged over a **target population** of **mothers**

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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.



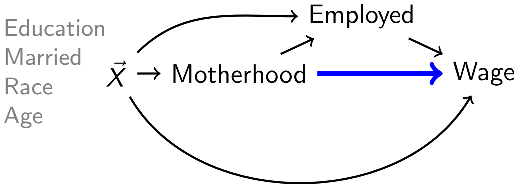
Averaged over a **target population** of **mothers**

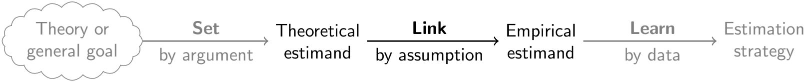
Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

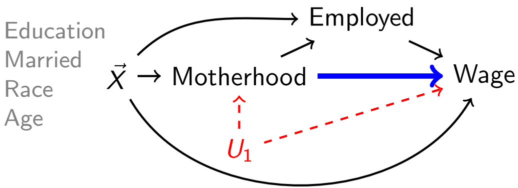
Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

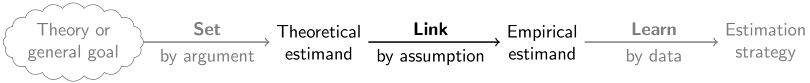
Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\\text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\\text{Employed}\end{pmatrix}$



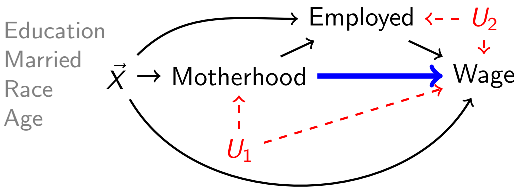


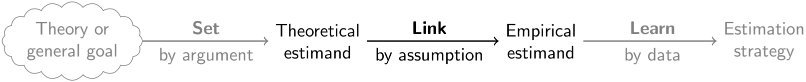
Unit-specific quantity: $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix} - Y_i\begin{pmatrix}\text{Non-mother,}\\ \text{Employed}\end{pmatrix}$





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

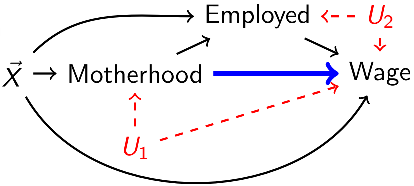
Potential outcomes

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

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

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

Education
Married
Race
Age



Lundberg, Johnson, and Stewart. What is Your Estimand?

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

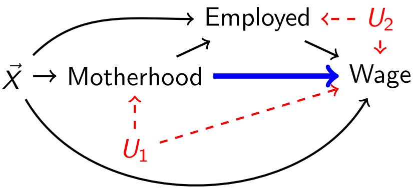
Potential outcomes

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

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

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

Education
Married
Race
Age



Theory or general goal → **Set** by argument → Theoretical estimand → **Link** by assumption → Empirical estimand → **Learn** by data → Estimation strategy

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

Potential outcomes

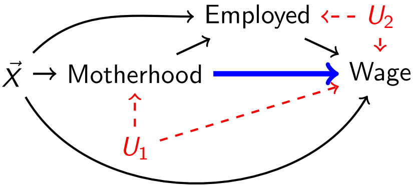
$Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$ $Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}$

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

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

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

Education
Married
Race
Age



Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Potential outcomes

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

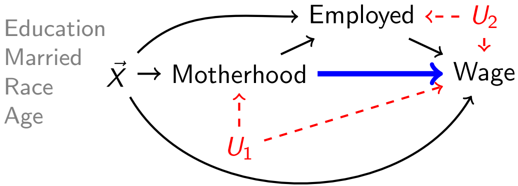
All women
with a high school degree,
married, white, age 30

Realized outcomes

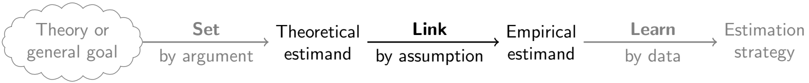
Y_i Y_i

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

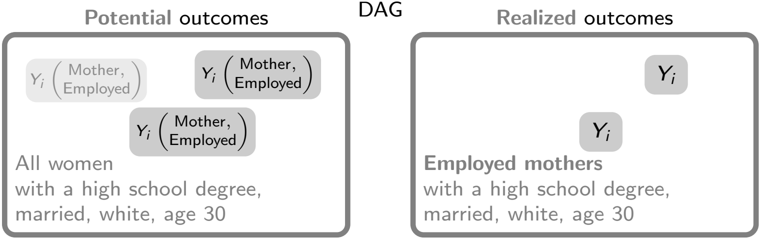
Focus on one $\vec{X}=\vec{x}_i$

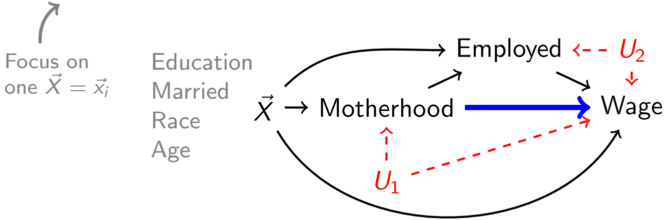


Lundberg, Johnson, and Stewart. What is Your Estimand?



$$\mathrm{E}\left(Y_i\begin{pmatrix}\text{Mother,}\\ \text{Employed}\end{pmatrix}\;\middle|\;\vec{X}=\vec{x}_i\right) \underset{\substack{\text{By the}\\ \text{DAG}}}{=} \mathrm{E}\left(Y_i\;\middle|\;\begin{matrix}\text{Motherhood} & = & \text{Mother,}\\ \text{Employment} & = & \text{Employed,}\\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i\end{matrix}\right)$$

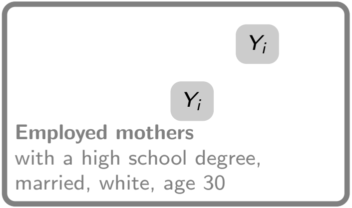




Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Realized outcomes



$$\text{Theory or general goal} \xrightarrow[\text{by argument}]{\textbf{Set}} \text{Theoretical estimand} \xrightarrow[\text{by assumption}]{\textbf{Link}} \text{Empirical estimand} \xrightarrow[\text{by data}]{\textbf{Learn}} \text{Estimation strategy}$$

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

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

Realized outcomes

Y_i

Y_i

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

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

Any prediction algorithm that minimizes squared errors

Realized outcomes

Y_i

Y_i

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

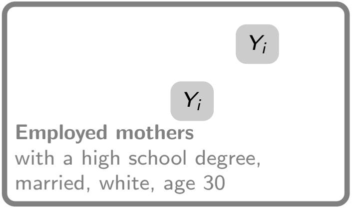
This can be estimated by **machine learning**!

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

Any prediction algorithm that minimizes squared errors

Generalized Additive Model

Realized outcomes



Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

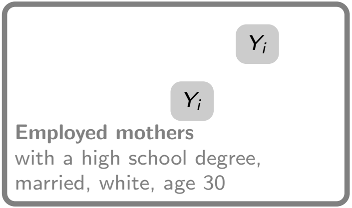
$$\mathrm{E}\left(Y_i \;\middle|\; \begin{array}{lcr} \text{Motherhood} & = & \text{Mother,} \\ \text{Employment} & = & \text{Employed,} \\ \text{Covariates } \vec{X} & = & \text{Observed } \vec{x}_i \end{array}\right)$$

Any prediction algorithm that minimizes squared errors

Generalized Additive Model

Random Forest

Realized outcomes



Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

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

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

1) Learn an algorithm to predict the outcome

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

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

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

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

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

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

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

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

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

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

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

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

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

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

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

3) Average over the target population

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

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

1) Learn an algorithm to predict the outcome

2) Predict for every unit at each treatment value

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

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

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

Choose an algorithm by **predictive performance**

Outcome	Log hourly wage
Predictors	Motherhood, age, race, education, marital status

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

Choose an algorithm by **predictive performance**

Outcome Log hourly wage

Predictors Motherhood, age, race, education, marital status

Candidate algorithms

Least flexible

Most flexible

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

Choose an algorithm by **predictive performance**

Outcome Log hourly wage

Predictors Motherhood, age, race, education, marital status

Candidate algorithms

Least flexible — OLS with a quadratic for age

Most flexible

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

Choose an algorithm by **predictive performance**

Outcome Log hourly wage

Predictors Motherhood, age, race, education, marital status

Candidate algorithms

Least flexible	OLS with a quadratic for age
	+ Interaction between age and motherhood
Most flexible	

Choose an algorithm by **predictive performance**

Outcome Log hourly wage

Predictors Motherhood, age, race, education, marital status

Candidate algorithms

Least flexible

OLS with a quadratic for age

+ Interaction between age and motherhood

+ Allow a smooth curve for age rather than quadratic

Most flexible

Choose an algorithm by **predictive performance**

Outcome Log hourly wage

Predictors Motherhood, age, race, education, marital status

Candidate algorithms

Least flexible

OLS with a quadratic for age

+ Interaction between age and motherhood

+ Allow a smooth curve for age rather than quadratic

+ Include each age as a separate indicator variable

Most flexible

Theory or general goal $\xrightarrow[\text{by argument}]{\textbf{Set}}$ Theoretical estimand $\xrightarrow[\text{by assumption}]{\textbf{Link}}$ Empirical estimand $\xrightarrow[\text{by data}]{\textbf{Learn}}$ Estimation strategy

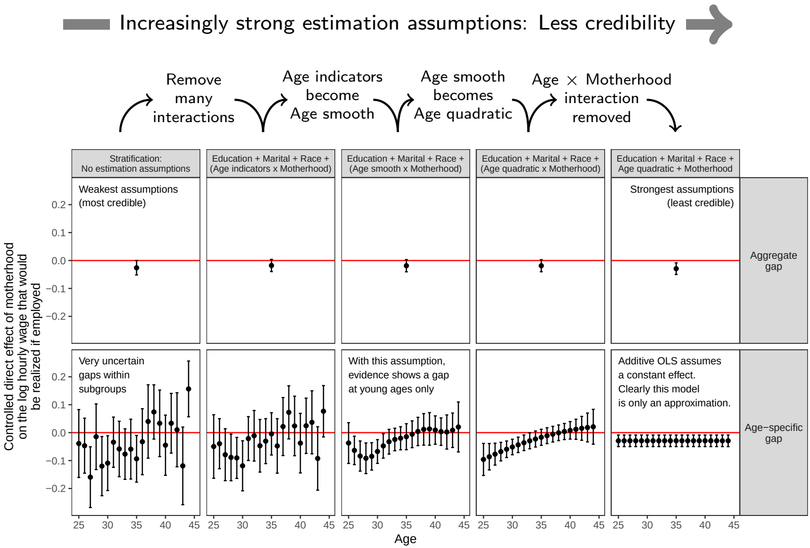
Choose an algorithm by **predictive performance**

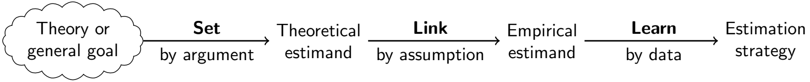
Outcome Log hourly wage

Predictors Motherhood, age, race, education, marital status

Candidate algorithms

Least flexible	OLS with a quadratic for age
	+ Interaction between age and motherhood
	+ Allow a smooth curve for age rather than quadratic
	+ Include each age as a separate indicator variable
Most flexible	+ Include all interactions among all predictors

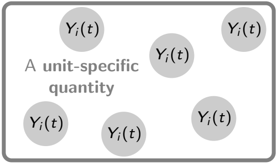

Increasingly strong estimation assumptions: Less credibility
Remove many interactions
Age indicators become Age smooth
Age smooth becomes Age quadratic
Age × Motherhood interaction removed
Stratification: No estimation assumptions
Education + Marital + Race + (Age indicators x Motherhood)
Education + 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
0.2
0.1
0.0
−0.1
−0.2
25
30
35
40
45
Age



What is your estimand?

⟵ Every quantitative study should answer this question

Defining the Target Quantity Connects Statistical Evidence to Theory



Averaged over a **target population**

Ian Lundberg

ilundberg@princeton.edu
ianlundberg.org

Rebecca Johnson

rebecca.ann.johnson@dartmouth.edu
rebeccajohnson.io

Brandon Stewart

bms4@princeton.edu
brandonstewart.org

Preprint on SocArxiv
Code on Dataverse
American Sociological Review