Predictive algorithms: Introduction to an interactive workshop

Ian Lundberg
UCLA

Including past work with
Rebecca Johnson (Georgetown)
Brandon Stewart (Princeton)

and current work with
Kristin Liao (UCLA)
ilundberg.github.io/description

We acknowledge support through facilities and resources provided by the California Center for Population Research at UCLA (CCPR), which receives core support (P2C-HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health.

Plan for today

- Estimands in quantitative social science
- Descriptive estimands: A $\hat{Y}$ view
- Intro to a computer tutorial

American Sociological Review
1–34

DOI:10.1177/00031224211004187
journals.sagepub.com/home/asr

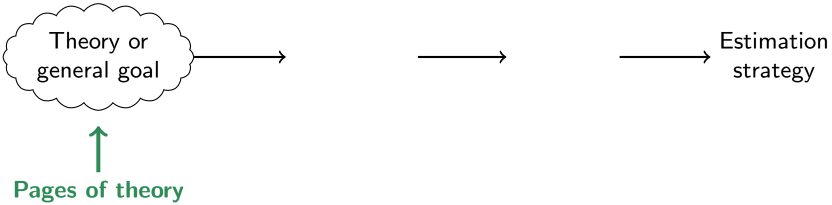
What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory

Ian Lundberg,[a] Rebecca Johnson,[b] and Brandon M. Stewart[a]

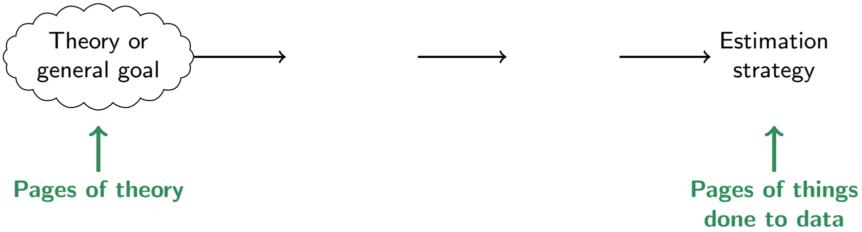
Research framework: Estimands connect theory to evidence

Theory or general goal ⟶ ⟶ ⟶ Estimation strategy

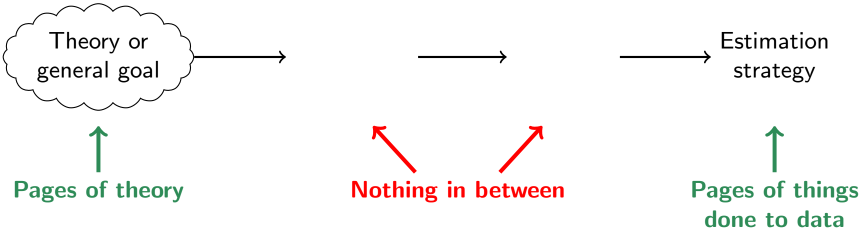
Research framework: Estimands connect theory to evidence



Research framework: Estimands connect theory to evidence



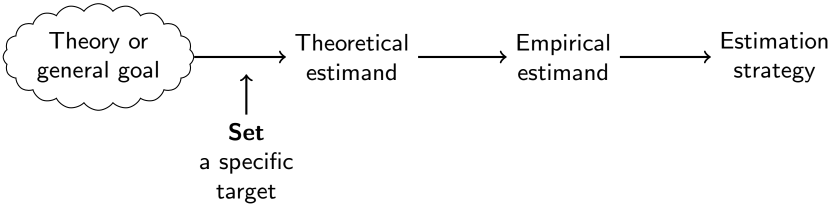
Research framework: Estimands connect theory to evidence



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



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**	$\frac{1}{\substack{\text{Size of U.S.}\\\text{adult population}}} \sum_{\substack{i \text{ in U.S.}\\\text{adult population}}} \left(\text{Employed}_i\right)$

Research framework: Estimands connect theory to evidence

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

Set a specific target

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

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**	$\frac{1}{\text{Size of U.S. adult population}} \sum_{\substack{i \text{ in U.S.} \\ \text{adult population}}} \left(\underbrace{\text{Employed}_i(\text{Job training})}_{\substack{\text{Employment if received} \\ \text{job training}}} - \underbrace{\text{Employed}_i(\text{No job training})}_{\substack{\text{Employment if did not receive} \\ \text{job training}}} \right)$

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

Research framework: Estimands connect theory to evidence

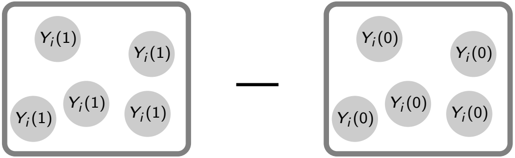
Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

Set a specific target

Definition

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

Example



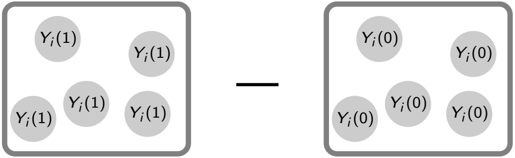
Lieberson 1987, Abbott 1988, Freedman 1991, Xie 2013, Hernán 2018

Research framework: Estimands connect theory to evidence

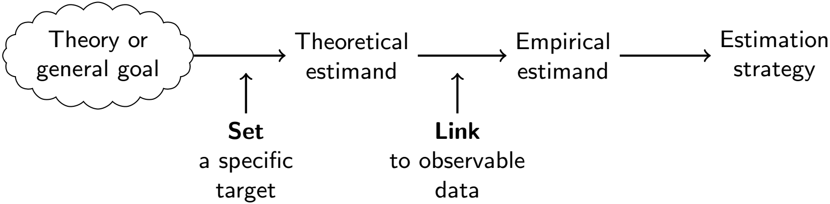
Theory or general goal → Theoretical estimand → Empirical estimand → Estimation strategy

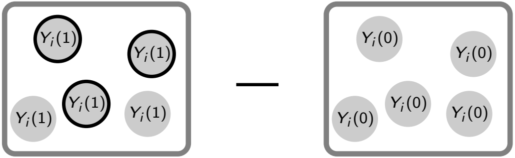
Set a specific target

Link to observable data

Definition	Example
A quantity involving **observable data**	

Research framework: Estimands connect theory to evidence



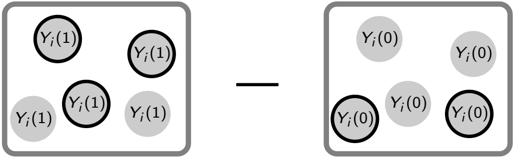
Definition	Example
A quantity involving **observable data**	

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

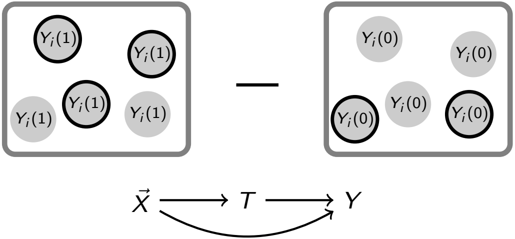
Definition	Example
A quantity involving **observable data**	

Research framework: Estimands connect theory to evidence

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

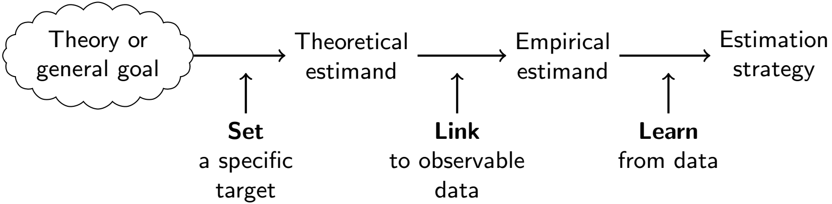
Set a specific target

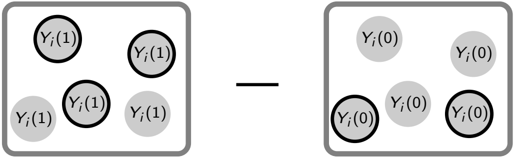
Link to observable data

Definition	Example
A quantity involving **observable data**	

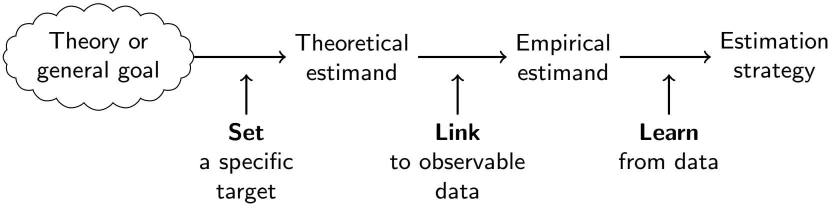
Pearl 2009, Imbens and Rubin 2015,
Morgan and Winship 2015, Elwert and Winship 2014

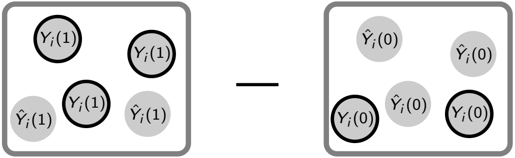
Research framework: Estimands connect theory to evidence



Definition	Example
An algorithm applied to data	

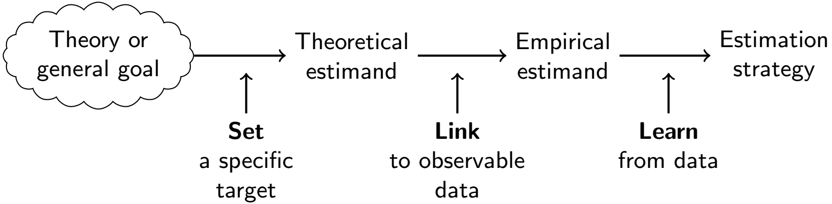
Research framework: Estimands connect theory to evidence



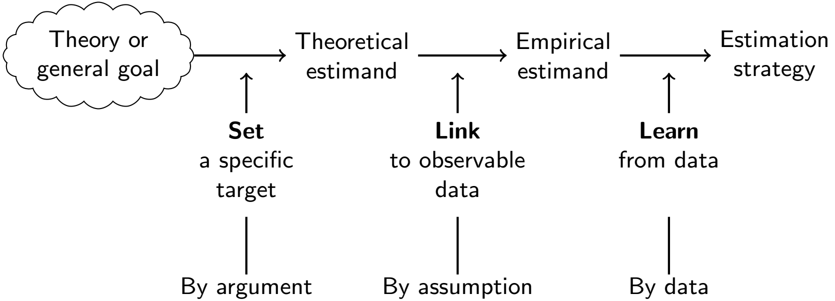
Definition	Example
An algorithm applied to data	

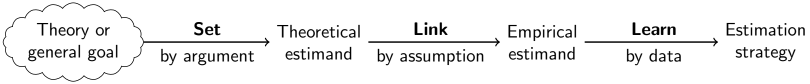
Young 2009, Watts 2014, Berk et al. 2019, Molina and Garip 2019

Research framework: Estimands connect theory to evidence

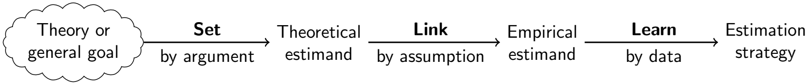


Research framework: Estimands connect theory to evidence





Angrist and Evans 1998



Effect of motherhood
on employment

Angrist and Evans 1998

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

Effect of motherhood
on employment

First two births
are the same sex

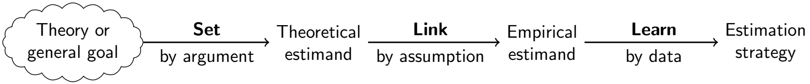
Angrist and Evans 1998

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

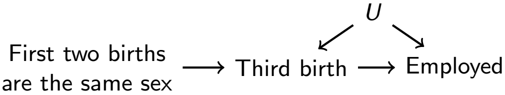
Effect of motherhood
on employment

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

Angrist and Evans 1998



Effect of motherhood
on employment



Angrist and Evans 1998

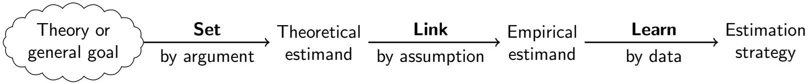
$$\begin{tikzcd} \text{Theory or general goal} \arrow[r, "\textbf{Set}", "\text{by argument}"'] & \text{Theoretical estimand} \arrow[r, "\textbf{Link}", "\text{by assumption}"'] & \text{Empirical estimand} \arrow[r, "\textbf{Learn}", "\text{by data}"'] & \text{Estimation strategy} \end{tikzcd}$$

Vague estimand

Effect of motherhood
on employment

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

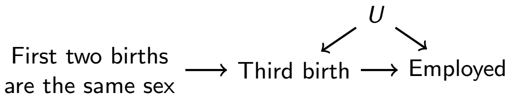
Angrist and Evans 1998



Vague estimand

Effect of motherhood on employment

Precise estimand



Angrist and Evans 1998

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

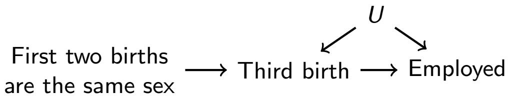
unit-specific quantity

Vague estimand

Effect of motherhood on employment

Precise estimand

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



Angrist and Evans 1998

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

Vague estimand

Effect of motherhood on employment

target population

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]

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

Angrist and Evans 1998

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

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

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)

Angrist and Evans 1998

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

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)

Angrist and Evans 1998

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

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

Angrist and Evans 1998

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

1. Set the target quantity.

Describe a population

What is the proportion employed
among U.S. resident women ages 21–35?

Describe a population

What is the proportion employed
among U.S. resident women ages 21–35?

Woman 1
Woman 2
Woman 3
Woman 4

Describe a population

What is the proportion employed
among U.S. resident women ages 21–35?

	Employed?
Woman 1	1
Woman 2	0
Woman 3	1
Woman 4	1

Describe population subgroups

What is the proportion employed
among U.S. resident women ages 21–35,
comparing mothers to non-mothers?

Describe population subgroups

What is the proportion employed
among U.S. resident women ages 21–35,
comparing mothers to non-mothers?

	Employed?
Mother 1	0
Mother 2	0
Mother 3	0
Mother 4	1

	Employed?
Non-Mother 1	1
Non-Mother 2	0
Non-Mother 3	1
Non-Mother 4	1

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

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

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

Woman 1
Woman 2
Woman 3
Woman 4

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$
Woman 1	0
Woman 2	0
Woman 3	0
Woman 4	1

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$
Woman 1	0	1
Woman 2	0	0
Woman 3	0	1
Woman 4	1	1

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$	Causal effect $Y(1) - Y(0)$
Woman 1	0	1	-1
Woman 2	0	0	0
Woman 3	0	1	-1
Woman 4	1	1	0

Why model?

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$	Causal effect $Y(1) - Y(0)$
Woman 1	0	1	-1
Woman 2	0	0	0
Woman 3	0	1	-1
Woman 4	1	1	0

Why model?

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$	Causal effect $Y(1) - Y(0)$
Woman 1	?	1	?
Woman 2	?	0	?
Woman 3	0	?	?
Woman 4	1	?	?

Why model?

Describe population subgroups

What is the proportion employed
among U.S. resident women ages 21–35,
comparing mothers to non-mothers?

	Employed?
Mother 1	0
Mother 2	0
Mother 3	0
Mother 4	1

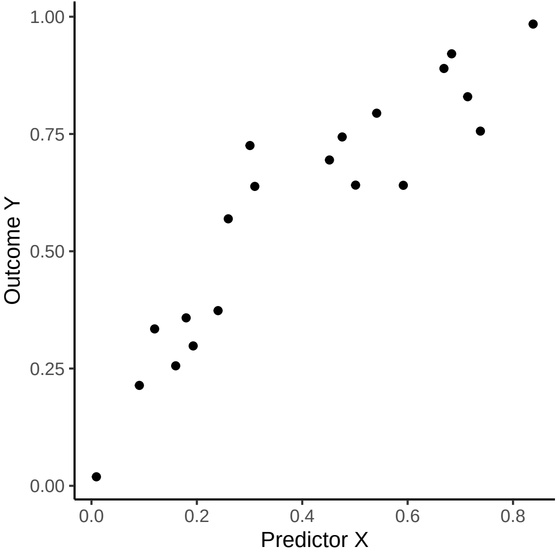
	Employed?
Non-Mother 1	1
Non-Mother 2	0
Non-Mother 3	1
Non-Mother 4	1

A $\hat{Y}$ view of description

With Kristin Liao, UCLA

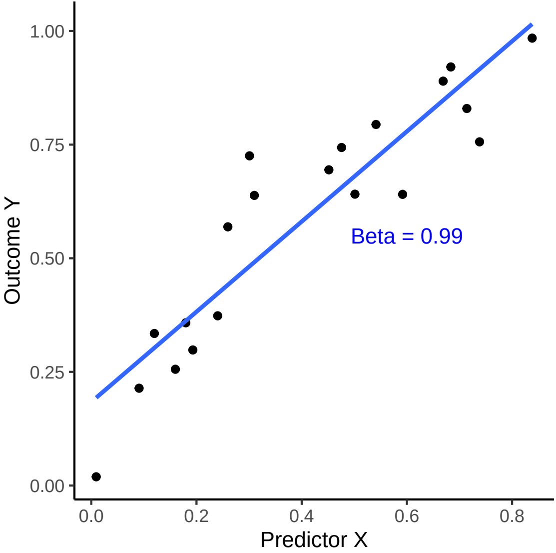
A $\hat{Y}$ view of description

With Kristin Liao, UCLA



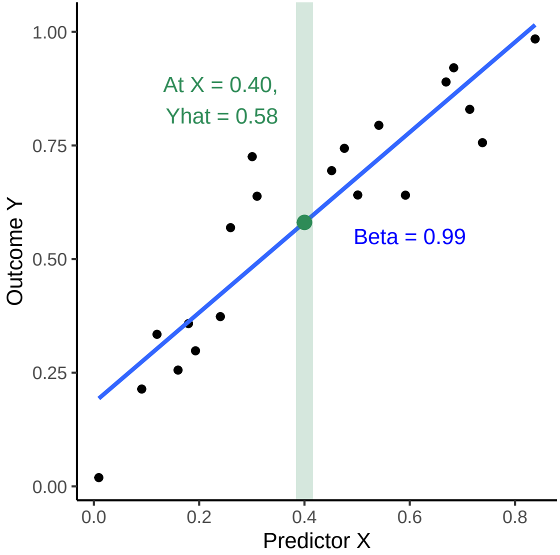
A $\hat{Y}$ view of description

With Kristin Liao, UCLA



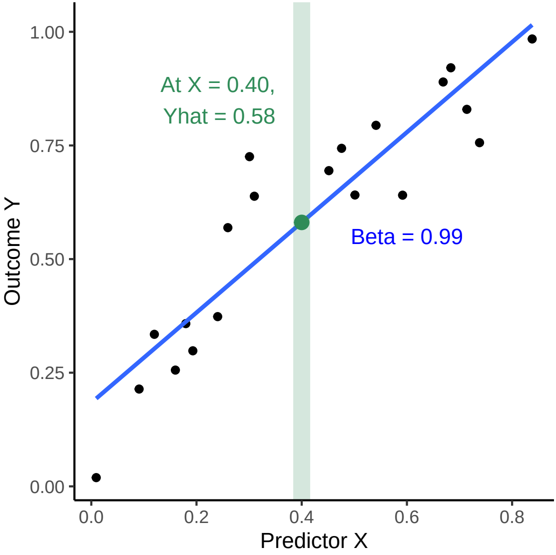
A $\hat{Y}$ view of description

With Kristin Liao, UCLA



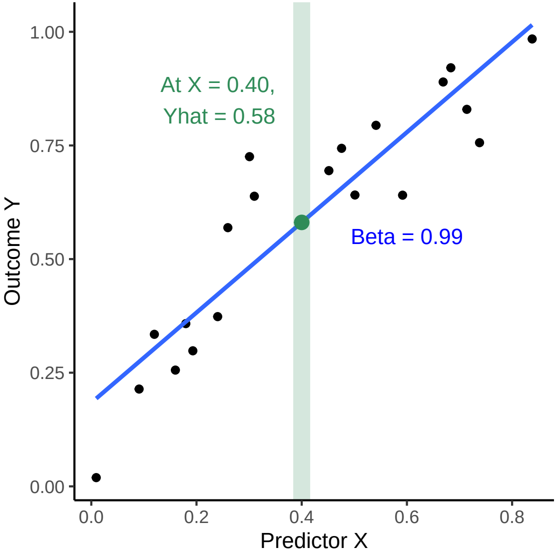
A $\hat{Y}$ view of description

With Kristin Liao, UCLA



A $\hat{Y}$ view of description

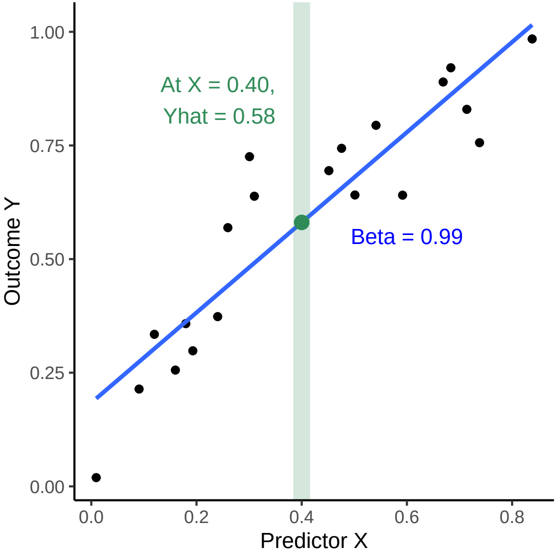
With Kristin Liao, UCLA



Why model?

A $\hat{Y}$ view of description

With Kristin Liao, UCLA

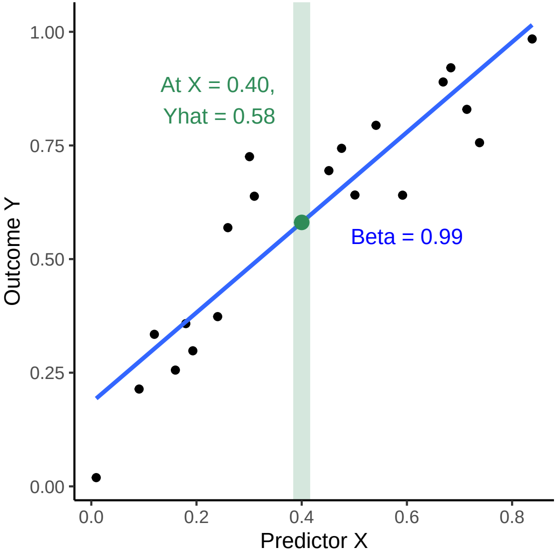


Why model?

A subgroups may have few units

A $\hat{Y}$ view of description

With Kristin Liao, UCLA



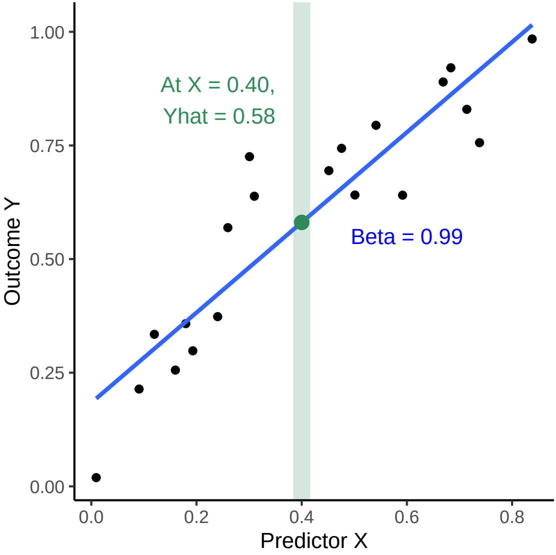
Why model?

A subgroups may have few units

Model pools information across subgroups

A $\hat{Y}$ view of description

With Kristin Liao, UCLA



Why model?

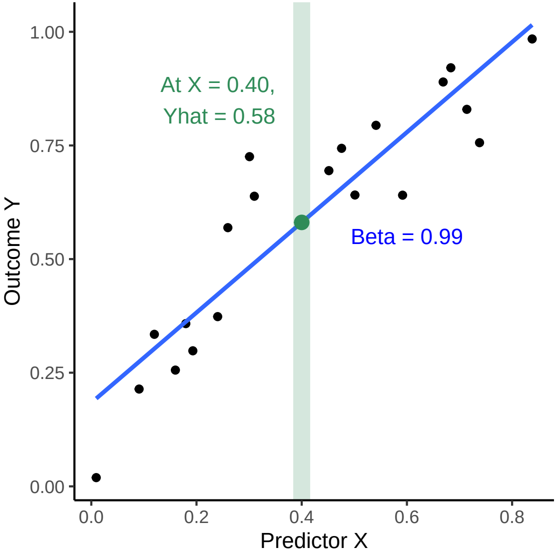
A subgroups may have few units

Model pools information across subgroups

Report $\hat{Y}$, not $\hat{\beta}$

A $\hat{Y}$ view of description

With Kristin Liao, UCLA



Why model?

A subgroups may have few units

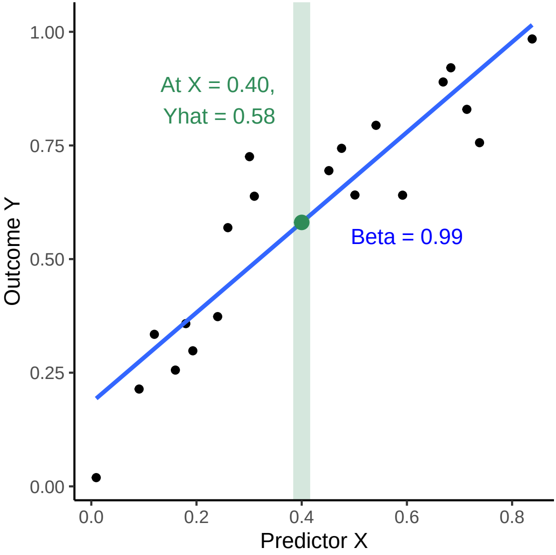
Model pools information across subgroups

Report $\hat{Y}$, not $\hat{\beta}$

Benefits

A $\hat{Y}$ view of description

With Kristin Liao, UCLA



Why model?

A subgroups may have few units

Model pools information across subgroups

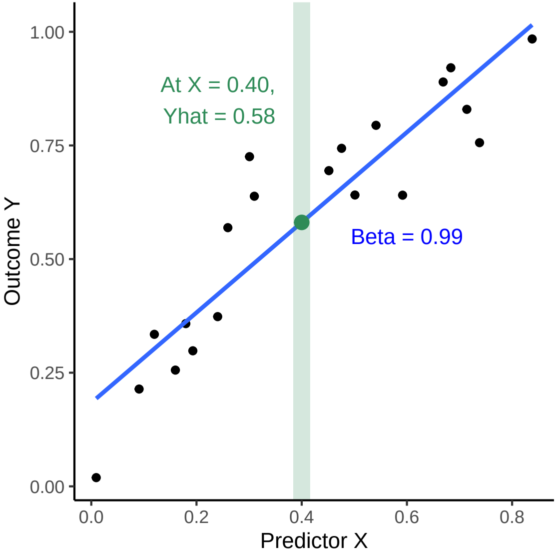
Report $\hat{Y}$, not $\hat{\beta}$

Benefits

Jargon-free results

A $\hat{Y}$ view of description

With Kristin Liao, UCLA



Why model?

A subgroups may have few units

Model pools information across subgroups

Report $\hat{Y}$, not $\hat{\beta}$

Benefits

Jargon-free results

Plug in machine learning

Concrete exercise: Sex gap in pay

ilundberg.github.io/description

Sample of 5 million cases	(true nonparametric estimates)
Simulate a sample of 100	(evaluate sample-based estimators)

Concrete exercise: Sex gap in pay

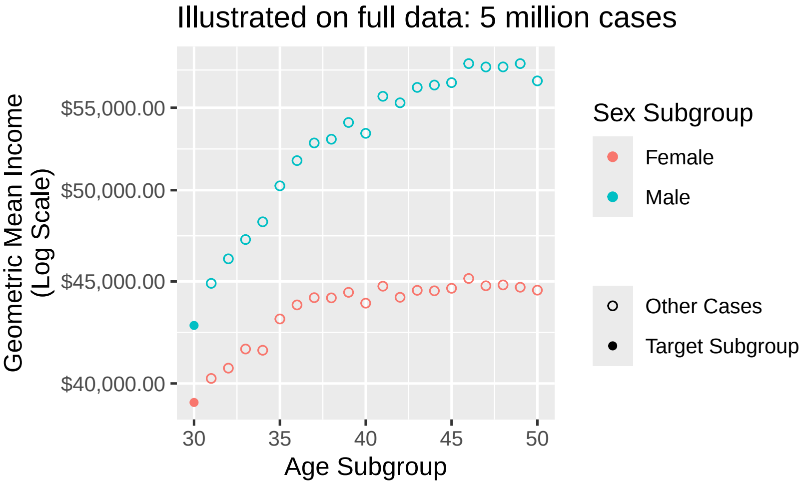
ilundberg.github.io/description

Source of 5 million cases

- American Community Survey (ACS) 2010–2019
- Adults age 30–50
- Worked 35+ hours per week in 50+ weeks last year
- Outcome: Annual wage and salary income

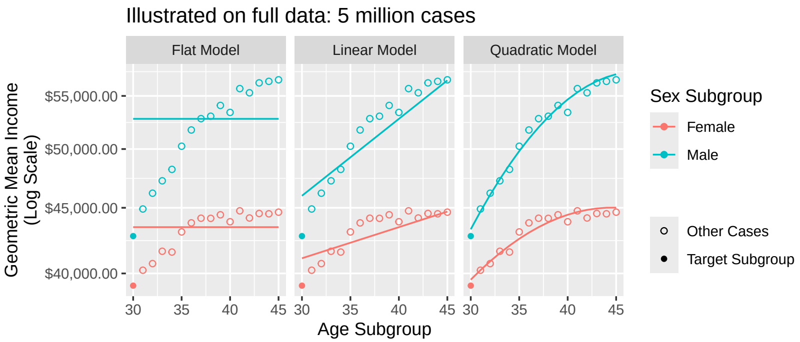
Concrete exercise: Sex gap in pay

ilundberg.github.io/description



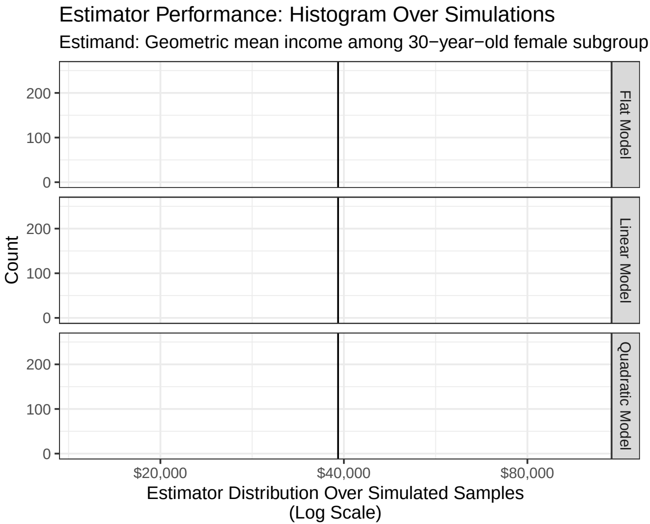
Concrete exercise: Sex gap in pay

ilundberg.github.io/description



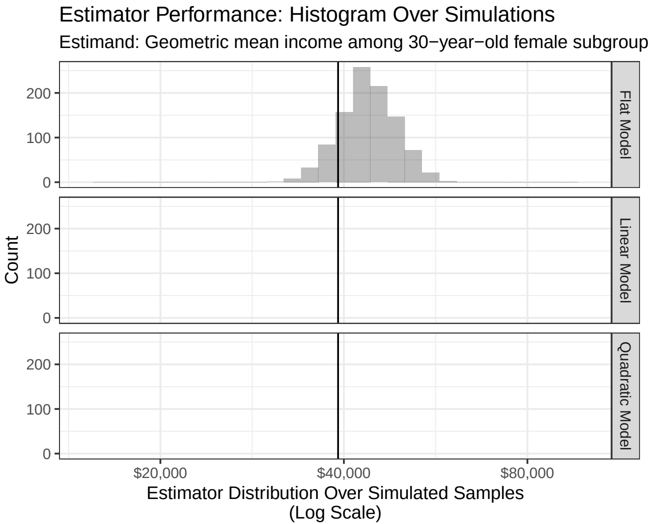
Evaluate models

ilundberg.github.io/description



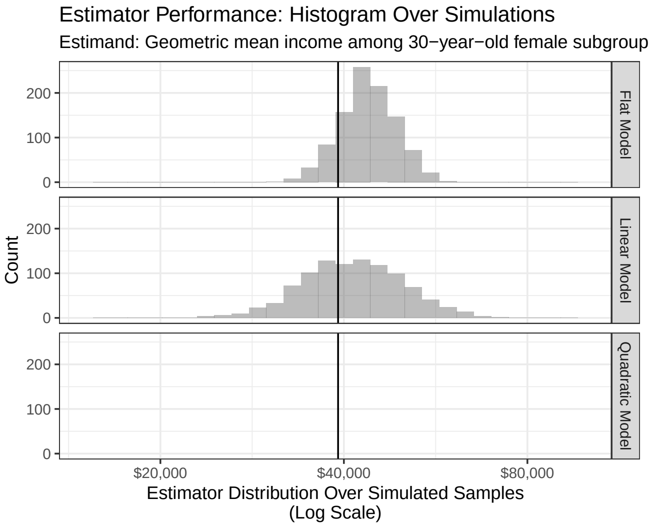
Evaluate models

ilundberg.github.io/description



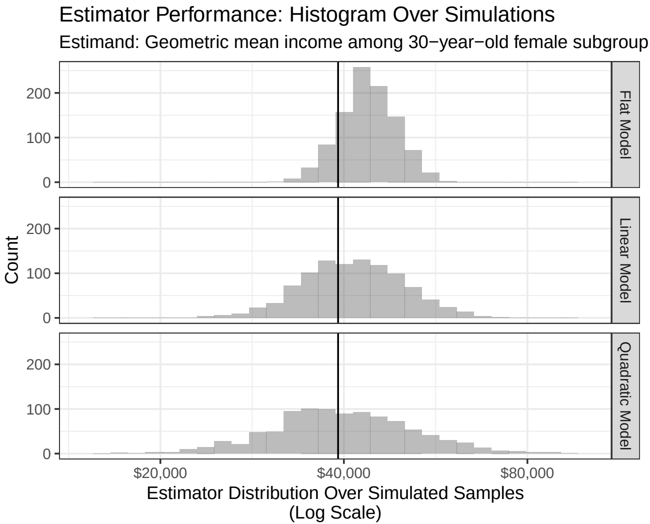
Evaluate models

ilundberg.github.io/description



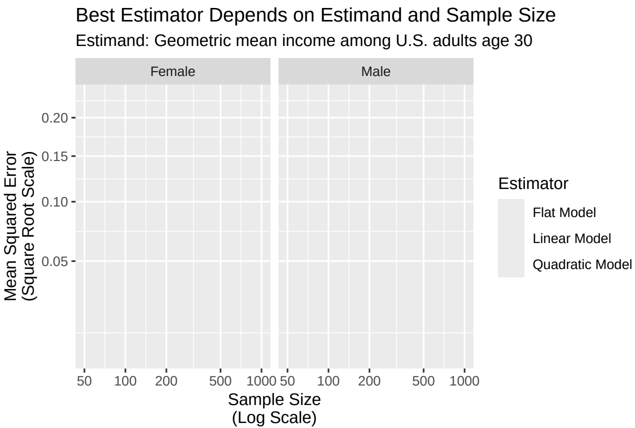
Evaluate models

ilundberg.github.io/description



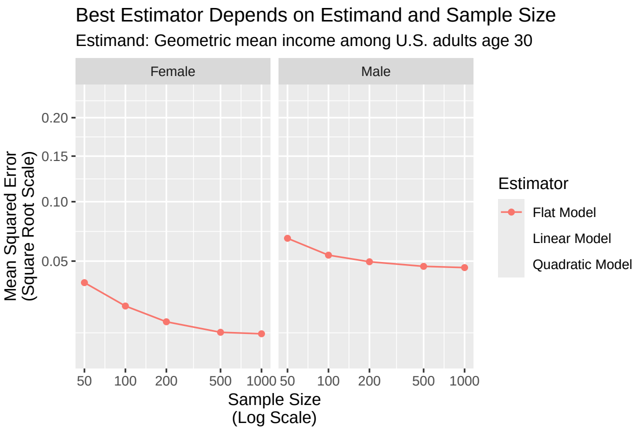
Evaluate models

ilundberg.github.io/description



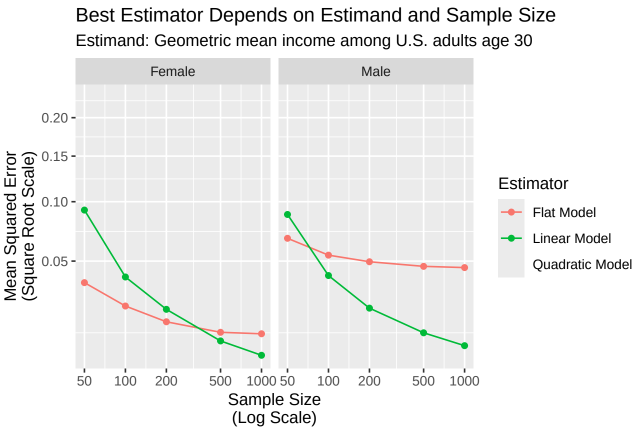
Evaluate models

ilundberg.github.io/description



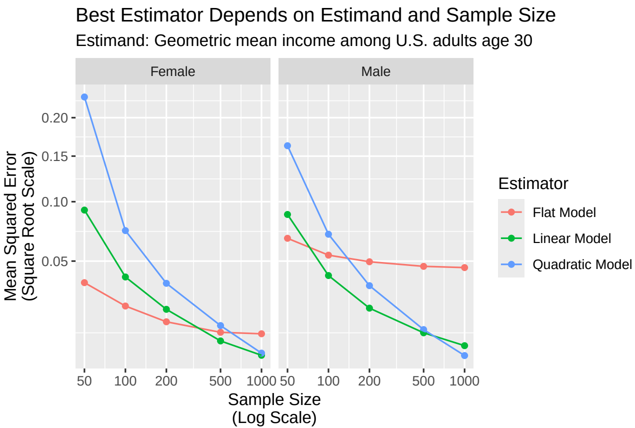
Evaluate models

ilundberg.github.io/description



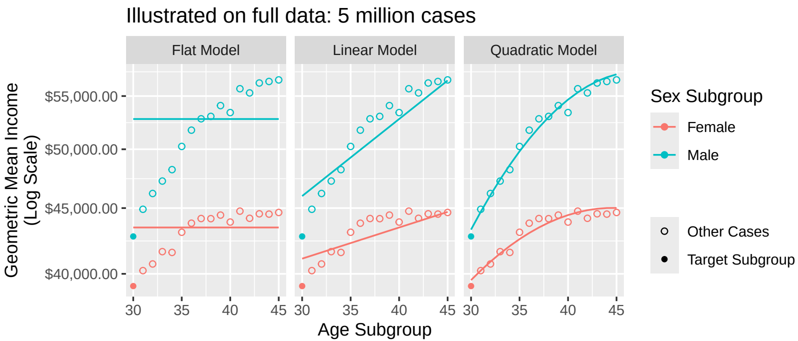
Evaluate models

ilundberg.github.io/description



Evaluate models

ilundberg.github.io/description



Implications of a $\hat{Y}$ view of description

ilundberg.github.io/description

Implications of a $\hat{Y}$ view of description

ilundberg.github.io/description

- a model as a means to an end
  - we would rather not model
  - model only when you lack data

Implications of a $\hat{Y}$ view of description

ilundberg.github.io/description

- a model as a means to an end
  - we would rather not model
  - model only when you lack data
- misspecified models are ok
  - flat model was wrong
  - flat model was best (lower variance)

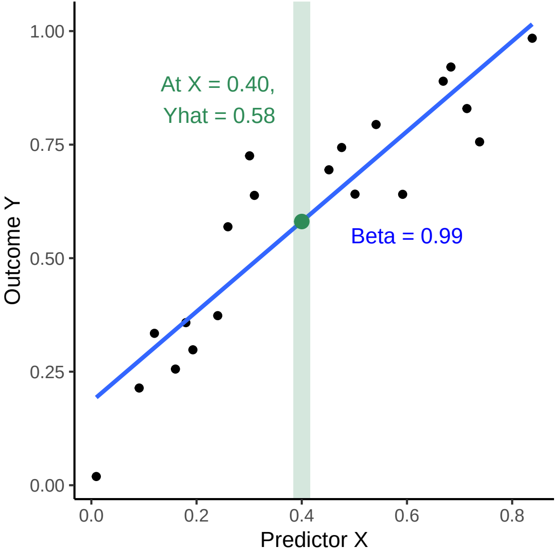
Implications of a $\hat{Y}$ view of description

ilundberg.github.io/description

- a model as a means to an end
  - we would rather not model
  - model only when you lack data
- misspecified models are ok
  - flat model was wrong
  - flat model was best (lower variance)
- machine learning becomes a plug-in

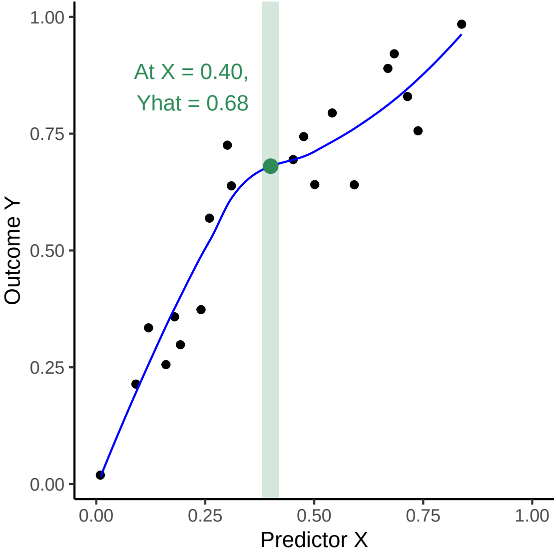
With $\hat{Y}$ description, machine learning becomes a plug-in

ilundberg.github.io/description



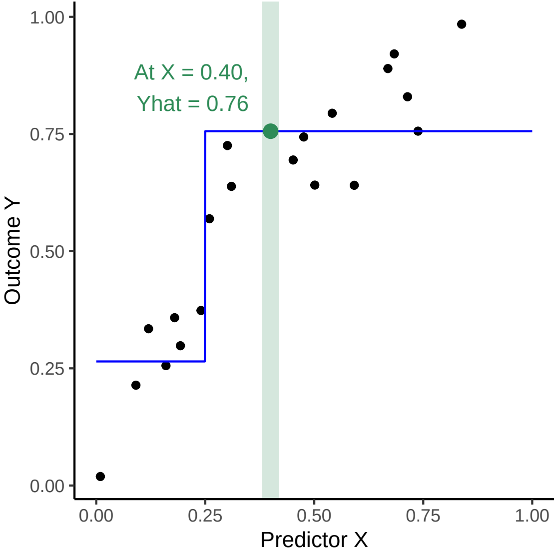
With $\hat{Y}$ description, machine learning becomes a plug-in

ilundberg.github.io/description



With $\hat{Y}$ description, machine learning becomes a plug-in

ilundberg.github.io/description



Computer tutorial: Introduction

ilundberg.github.io/description

Computer tutorial: Introduction

ilundberg.github.io/description

We will give you data:

- male and female incomes at age 30–50 in 2010–2019

You will make a forecast:

- male and female geometric mean income at age 30–50 in 2022

Computer tutorial: Introduction

ilundberg.github.io/description

Prepare the environment by loading the `tidyverse` package.

```
library(tidyverse)
```

The function below simulates a sample of 100 cases.

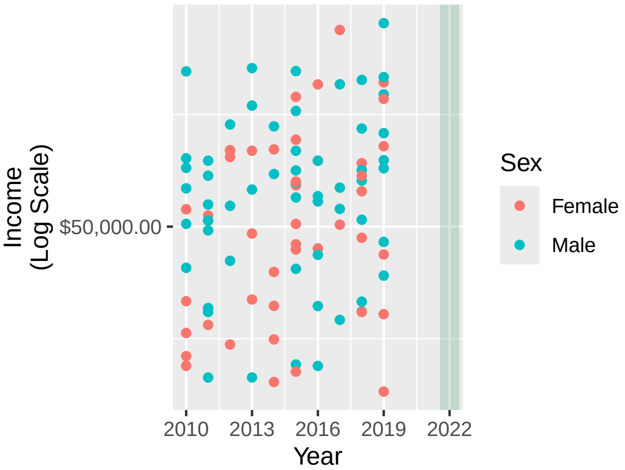
```
simulate <- function(n = 100) {
  read_csv("https://ilundberg.github.io/description/assets/truth.csv") |>
    slice_sample(n = n, weight_by = weight, replace = T) |>
    mutate(income = exp(rnorm(n(), meanlog, sdlog))) |>
    select(year, age, sex, income)
}
```

We can see how it works below.

```
simulated <- simulate(n = 100)
```

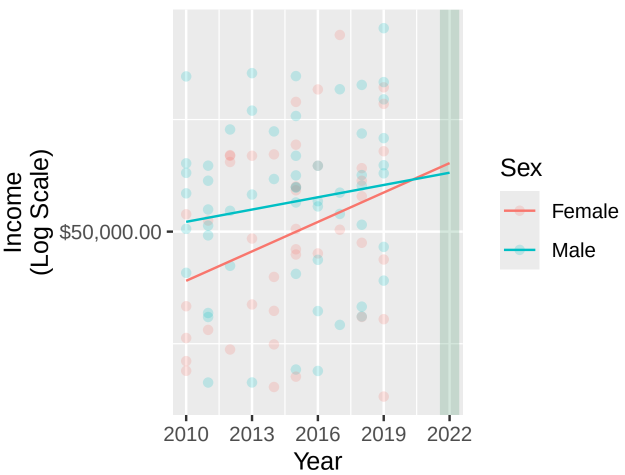
Computer tutorial: Introduction

ilundberg.github.io/description



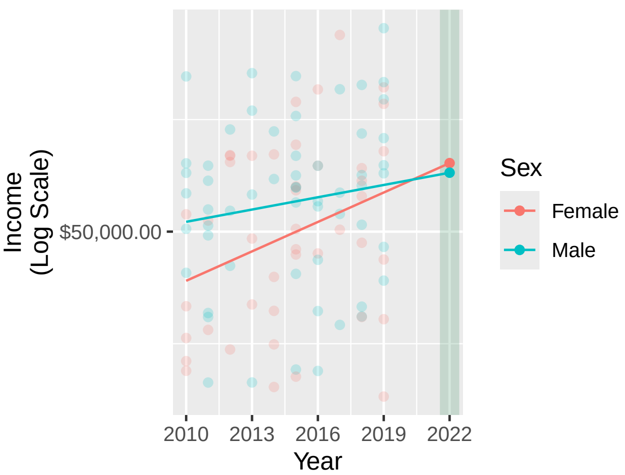
Computer tutorial: Introduction

ilundberg.github.io/description



Computer tutorial: Introduction

ilundberg.github.io/description



Computer tutorial: Introduction

ilundberg.github.io/description

We will give you data:

- male and female incomes at age 30–50 in 2010–2019

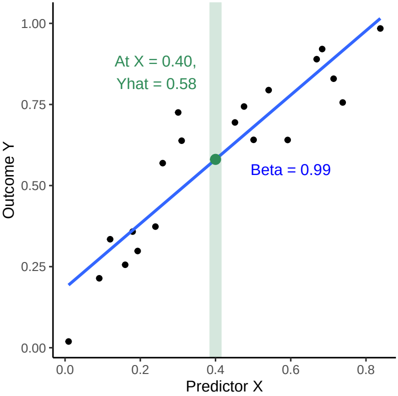
You will make a forecast:

- male and female geometric mean income at age 30–50 in 2022

We will see who comes closest

- to gold-standard truth from ACS 2022

Thanks!



Ian Lundberg
ianlundberg.org
ianlundberg@ucla.edu

Kristin Liao
kristinliao.com
ktliao@ucla.edu