Quantitative Data Analysis

SOCIOL 212B
Winter 2025

Lecture 1. Asking a Research Question and a $\hat{Y}$ View of Regression

How computing looked **in the 1950s**

Source: NASA

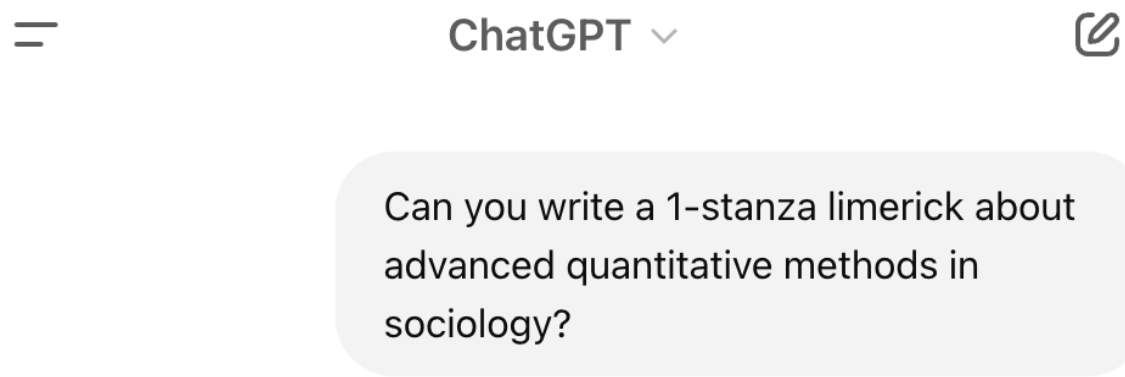
How computing looked **in the 1980s**

Source: Wikimedia

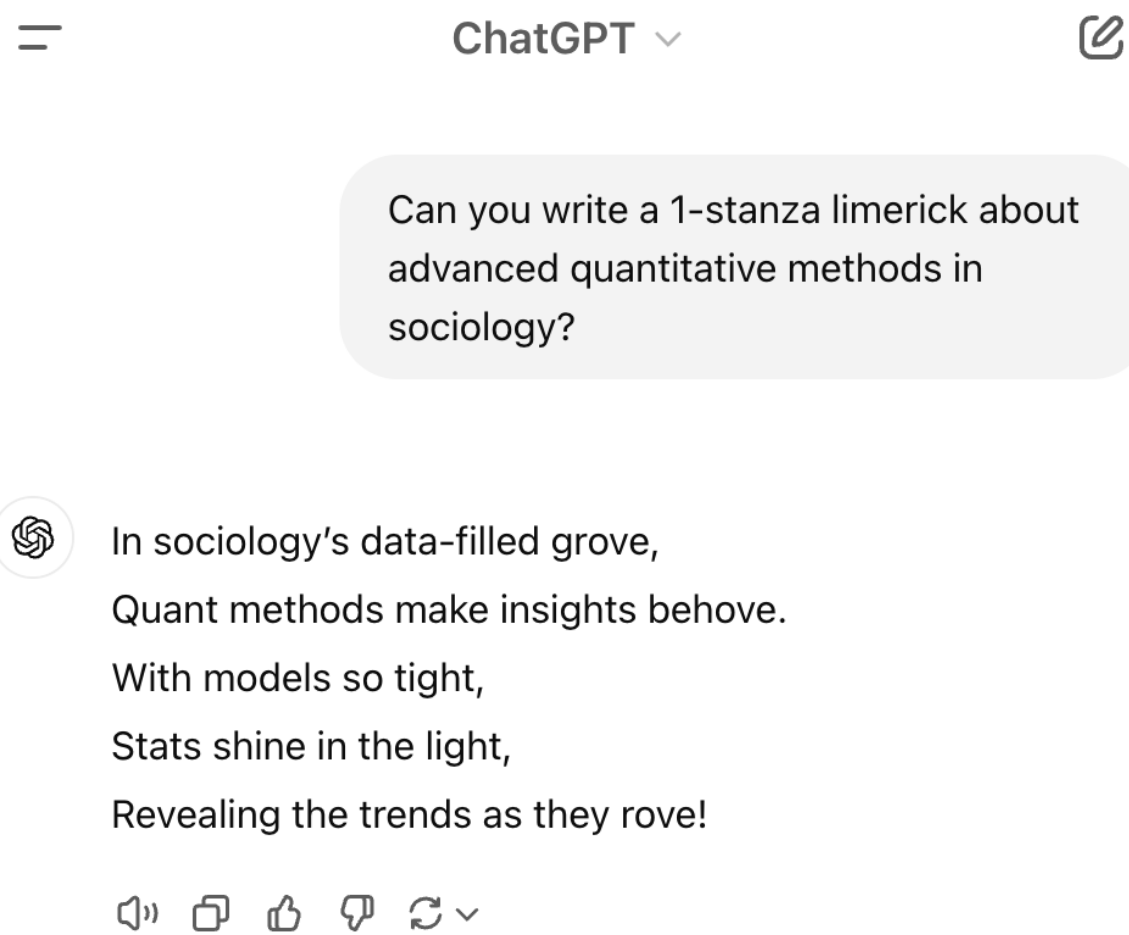
How computing looks **today**

Source: Apple

How computing looks **today**



How computing looks **today**



The SOCIOL 212 sequence

- new computational tools
- applied to write a social science paper

Course Plan

ilundberg.github.io/soc212b

Learning goals for today

By the end of class, you will be able to

- define an estimand in your project
  - unit-specific quantity
  - target population
- motivate regression from a $\hat{Y}$ view
  - as a tool to estimate despite sparse data
  - with the risk of various modeling errors
- make predictions to describe population subgroups
- organize your code in directories

Asking research questions

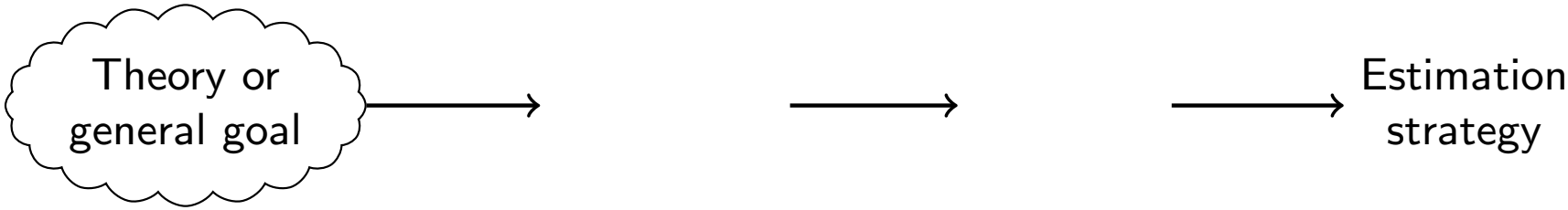
American Sociological Review
1–34

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

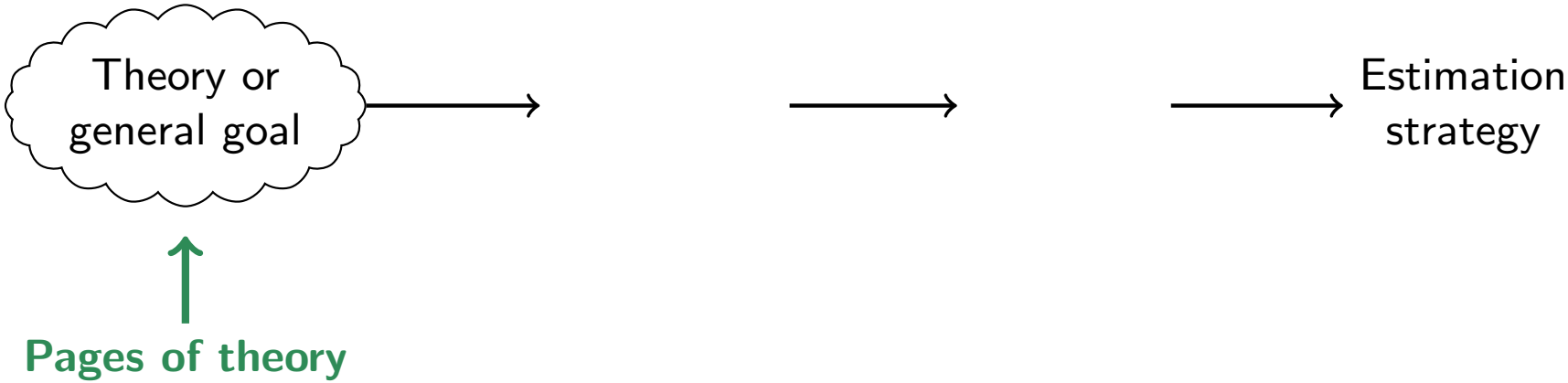
What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory

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

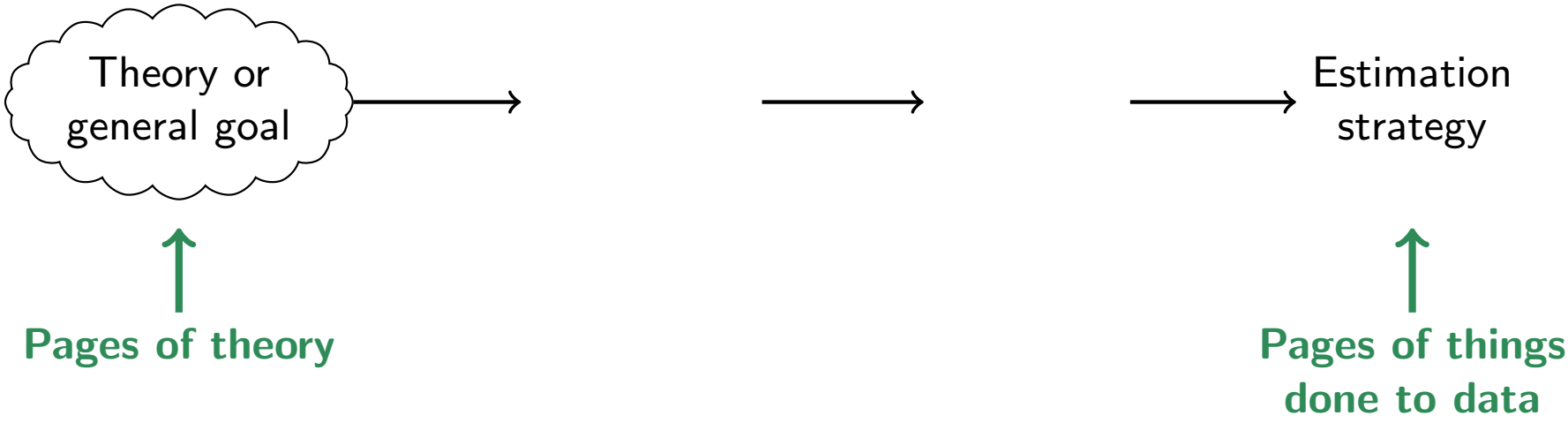
Research framework: Estimands connect theory to evidence



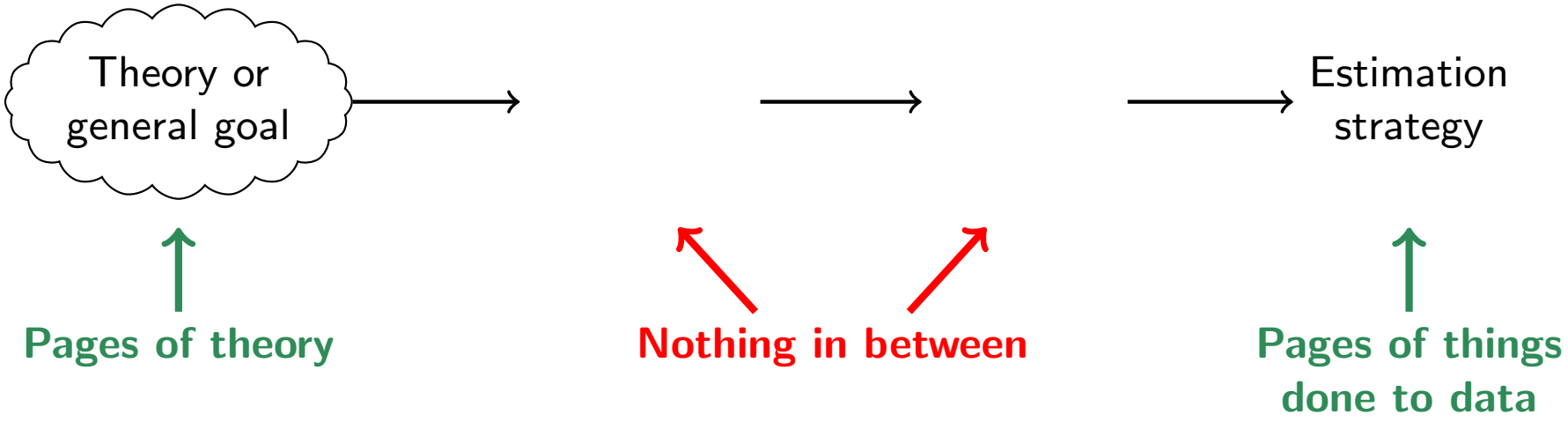
Research framework: Estimands connect theory to evidence



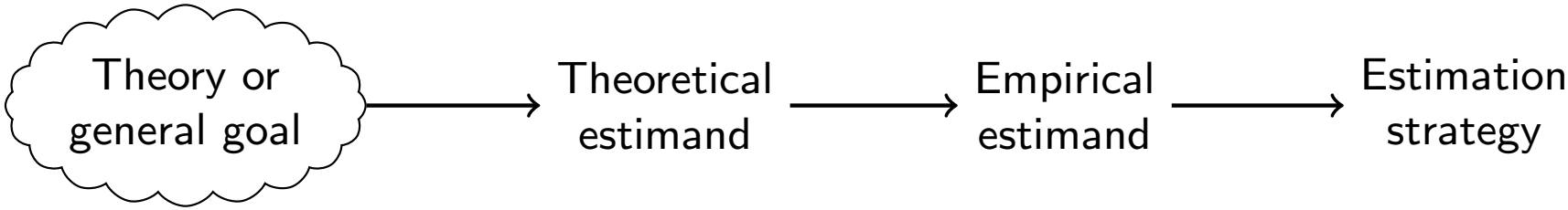
Research framework: Estimands connect theory to evidence



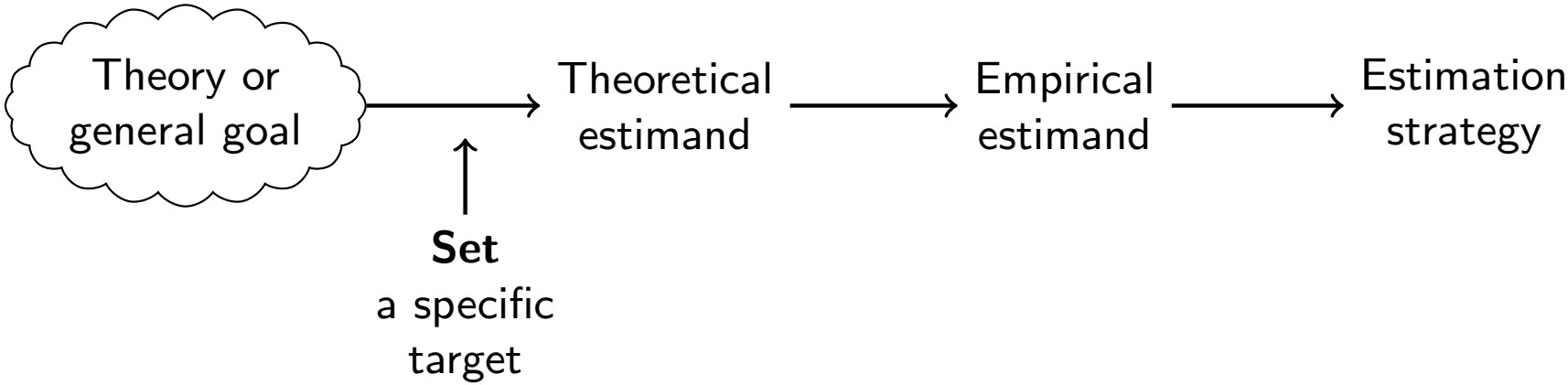
Research framework: Estimands connect theory to evidence



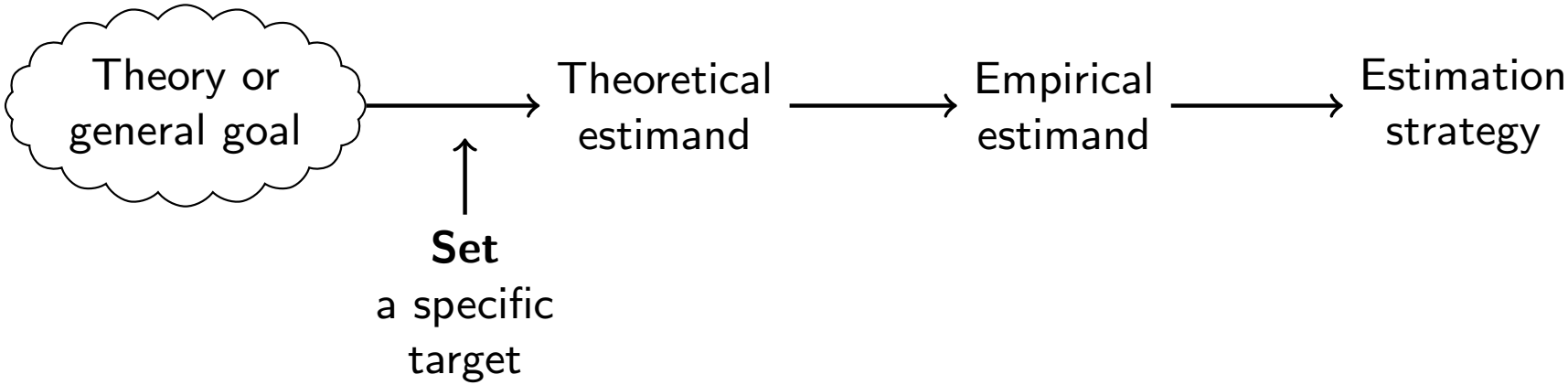
Research framework: Estimands connect theory to evidence



Research framework: Estimands connect theory to evidence



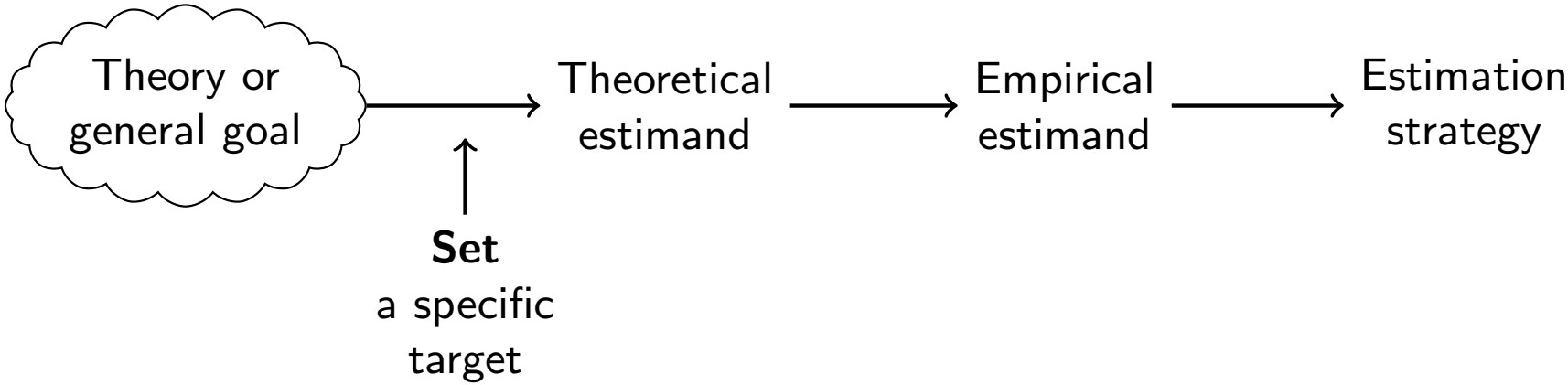
Research framework: Estimands connect theory to evidence



Definition

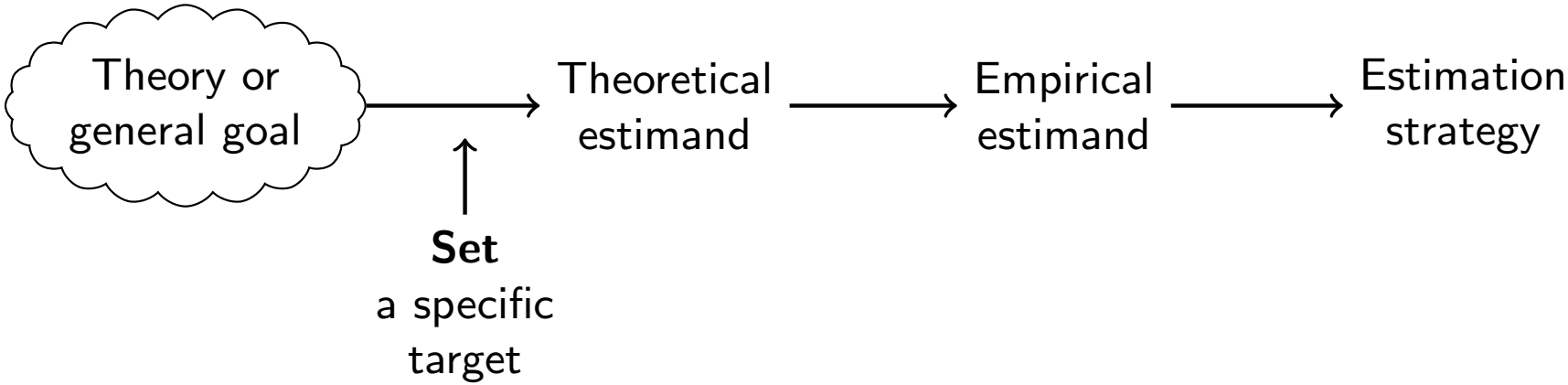
A **unit-specific quantity** aggregated over a **target population**

Research framework: Estimands connect theory to evidence



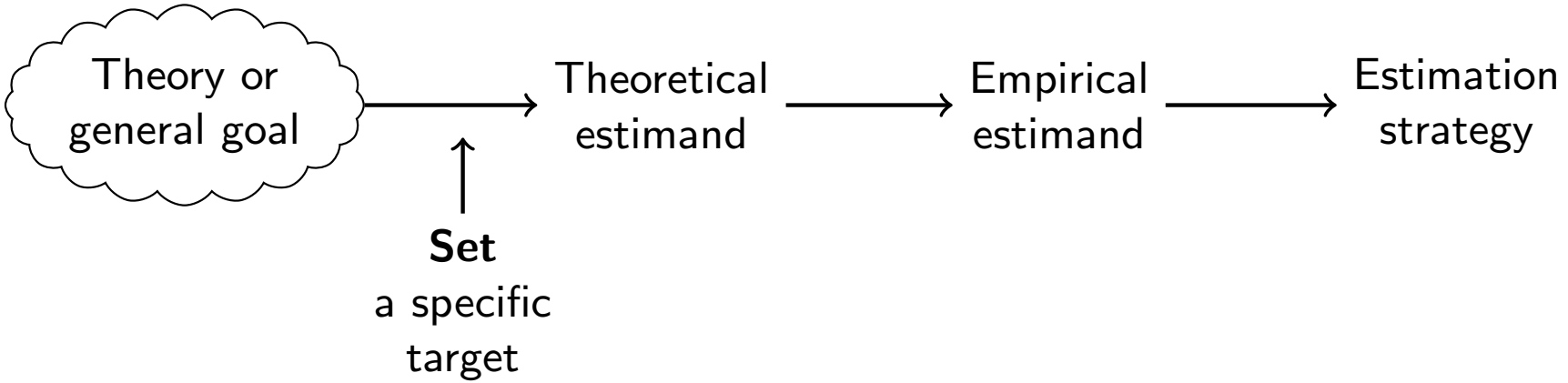
Definition	Example
A **unit-specific quantity** aggregated over a **target population**	$\frac{1}{\text{Size of U.S. adult population}} \sum_{i \text{ in U.S. adult population}} \left(\text{Employed}_i \right)$

Research framework: Estimands connect theory to evidence



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

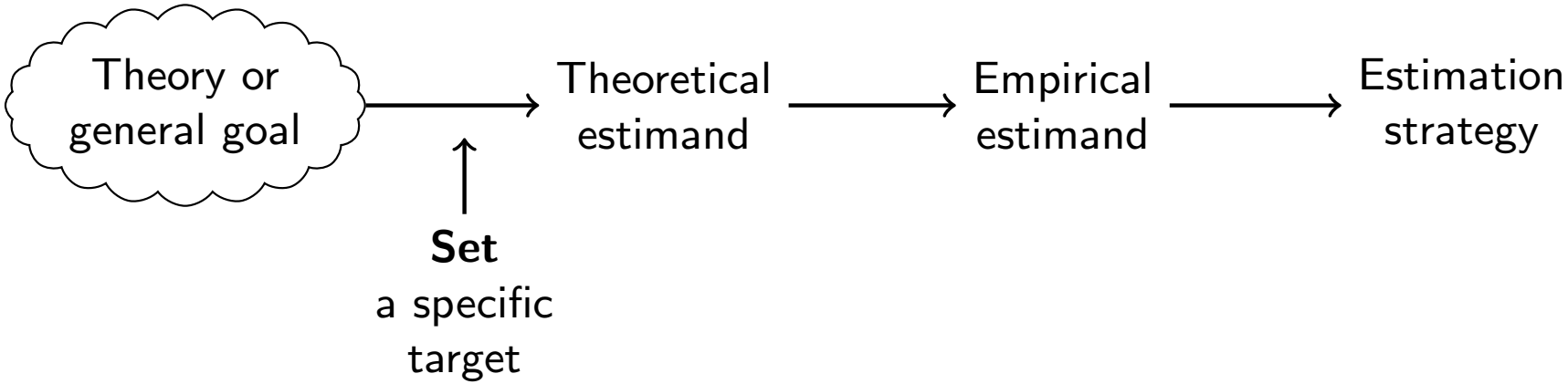
Research framework: Estimands connect theory to evidence



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

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

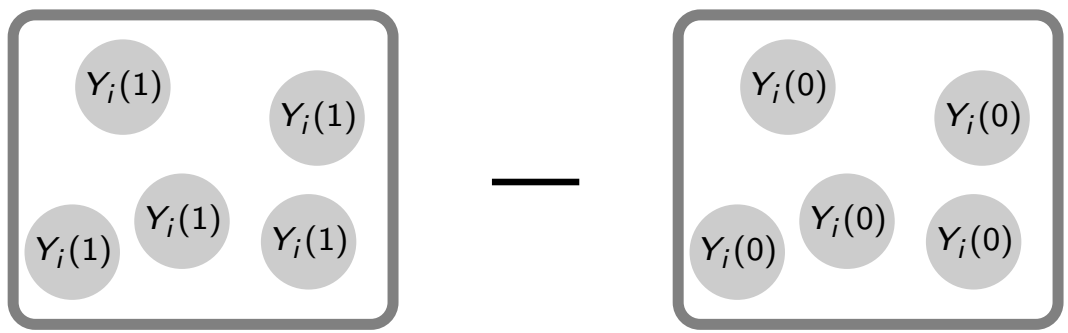
Research framework: Estimands connect theory to evidence



Definition

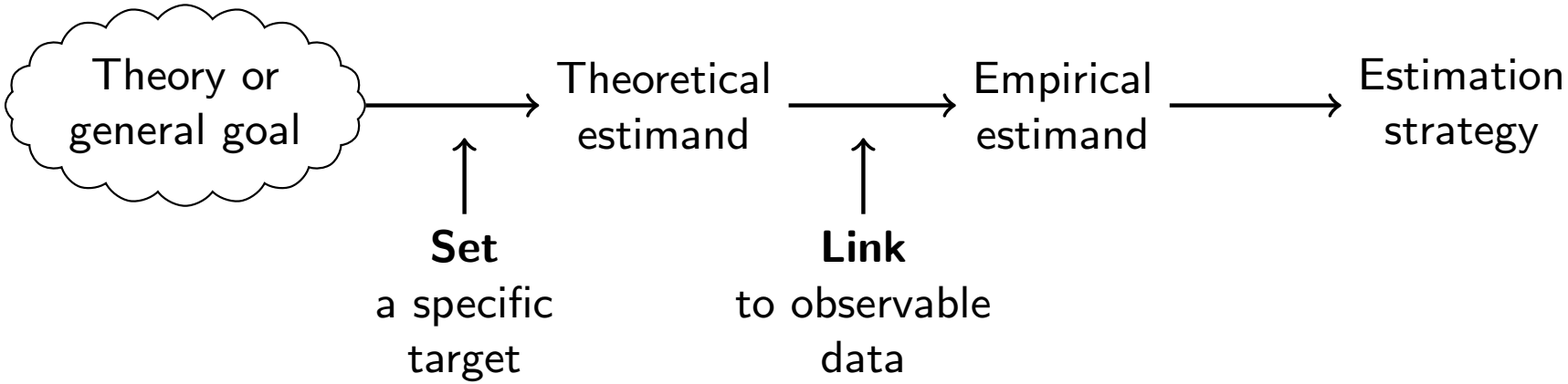
A **unit-specific quantity** aggregated over a **target population**

Example



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

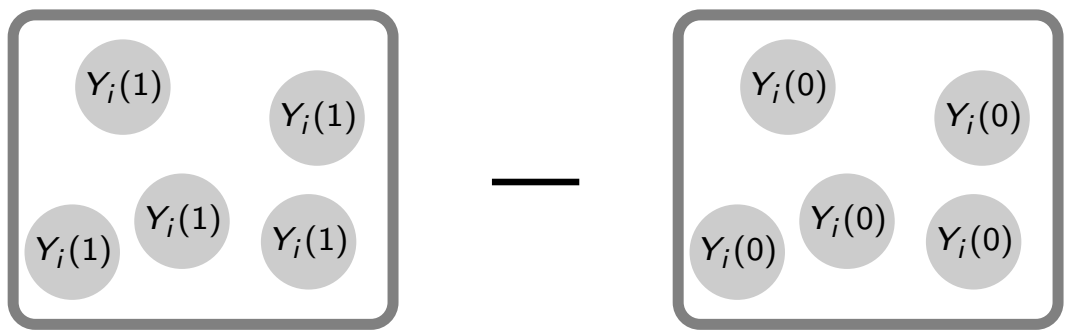
Research framework: Estimands connect theory to evidence



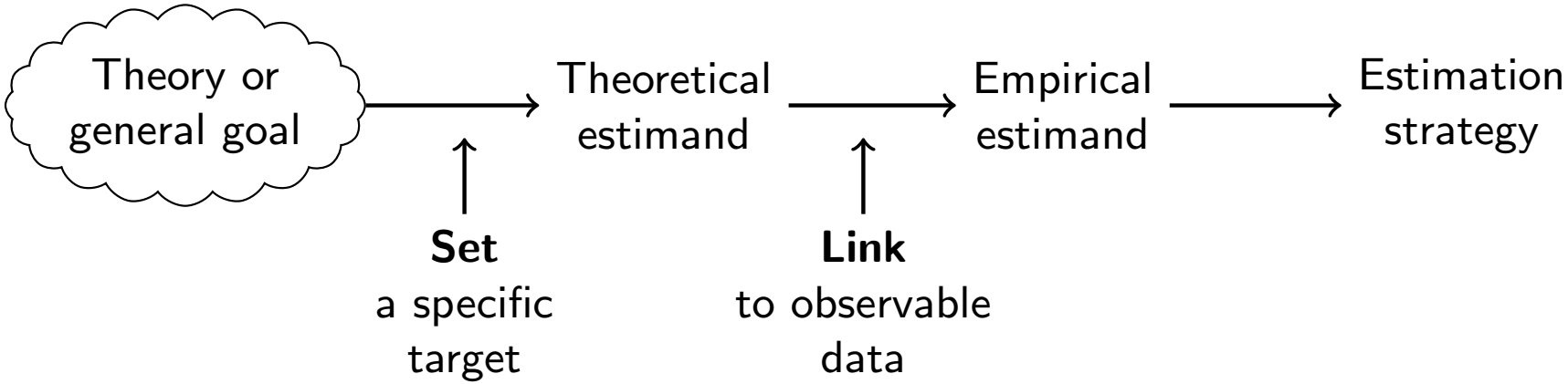
Definition

A quantity involving **observable data**

Example



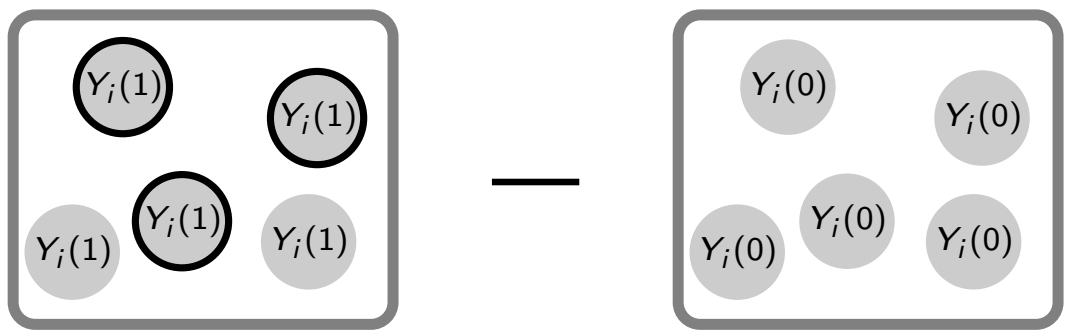
Research framework: Estimands connect theory to evidence



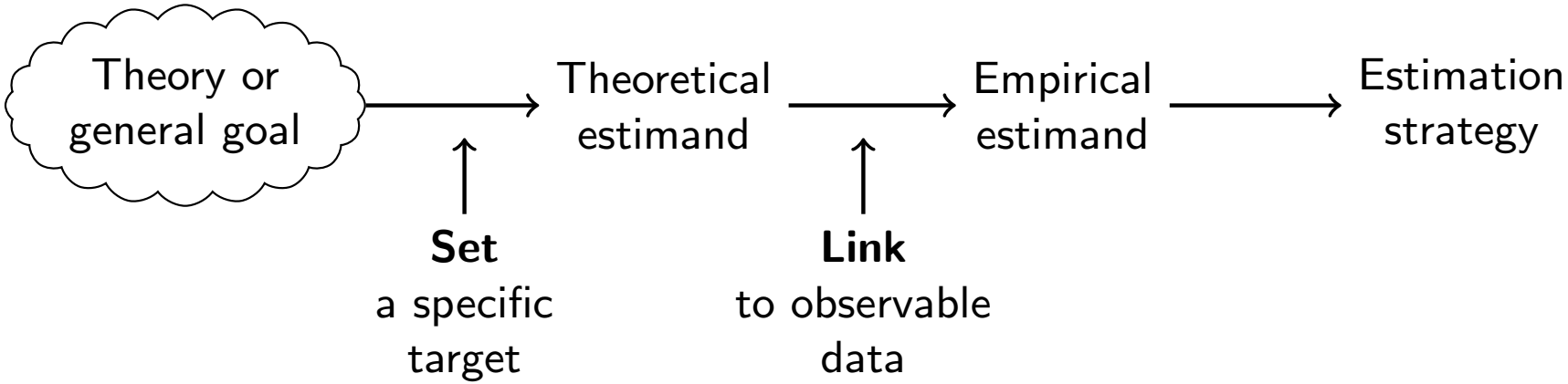
Definition

A quantity involving **observable data**

Example



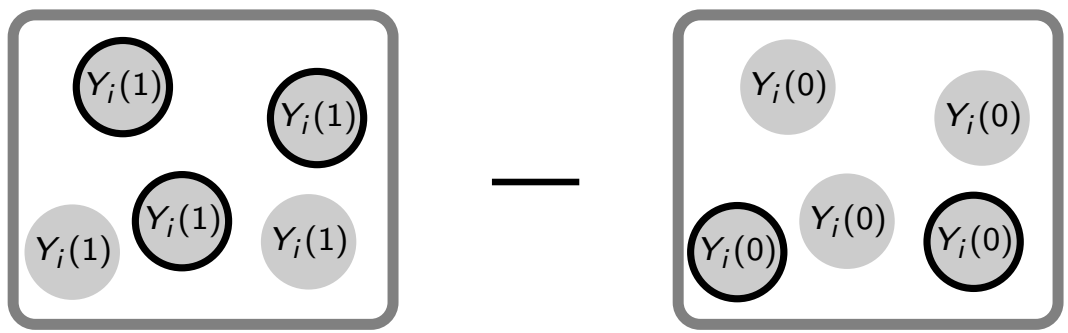
Research framework: Estimands connect theory to evidence



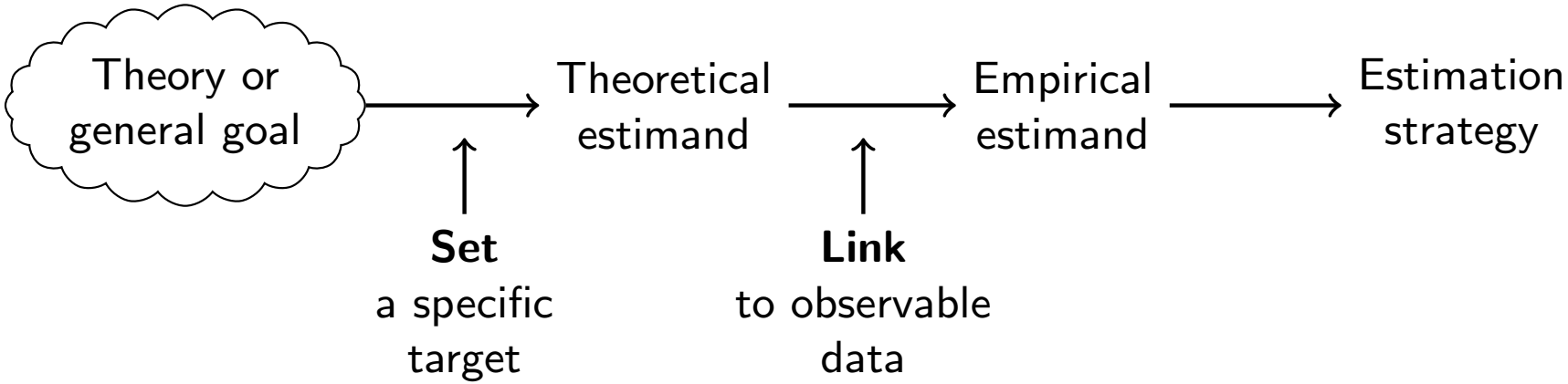
Definition

A quantity involving **observable data**

Example



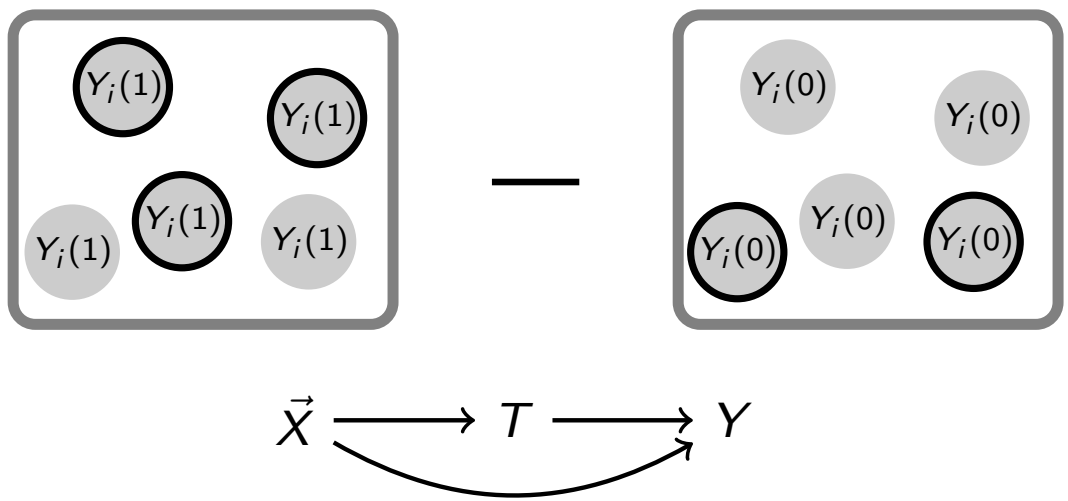
Research framework: Estimands connect theory to evidence



Definition

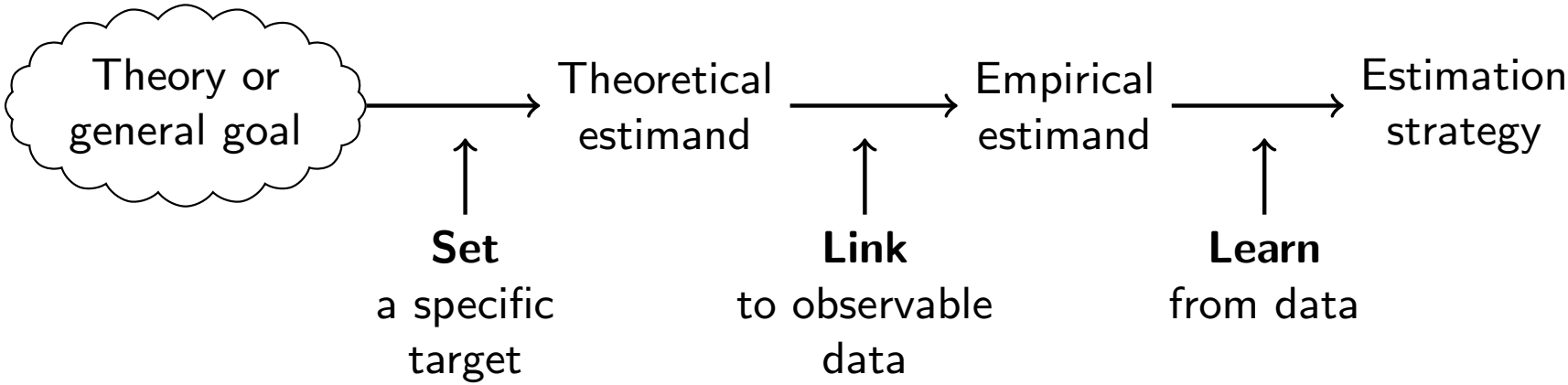
A quantity involving **observable data**

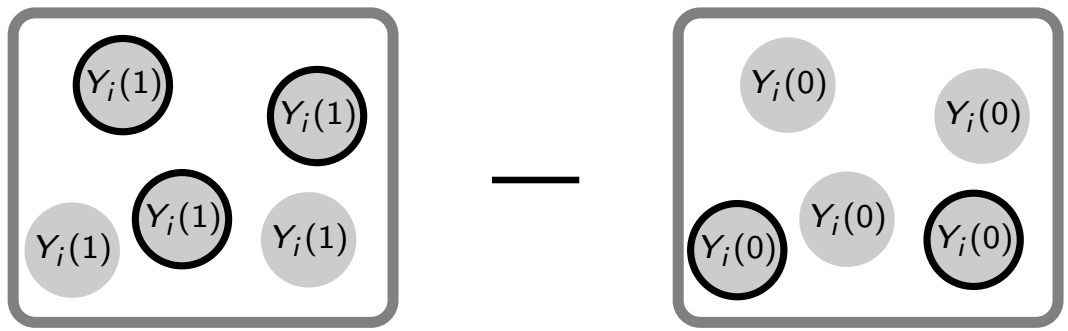
Example



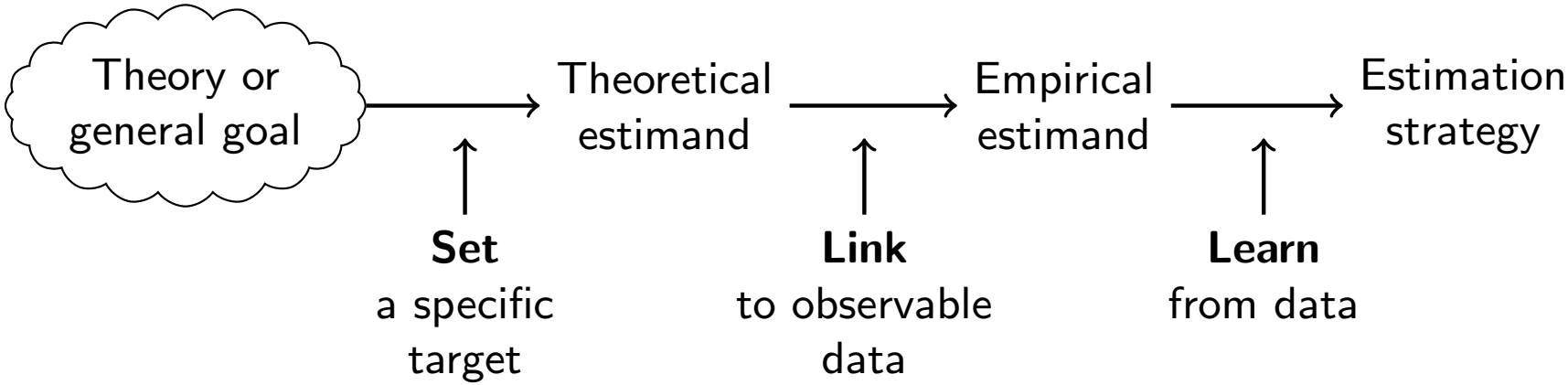
Pearl 2009, Imbens and Rubin 2015, Morgan and Winship 2015, Elwert and Winship 2014

Research framework: Estimands connect theory to evidence



Definition	Example
An algorithm applied to data	

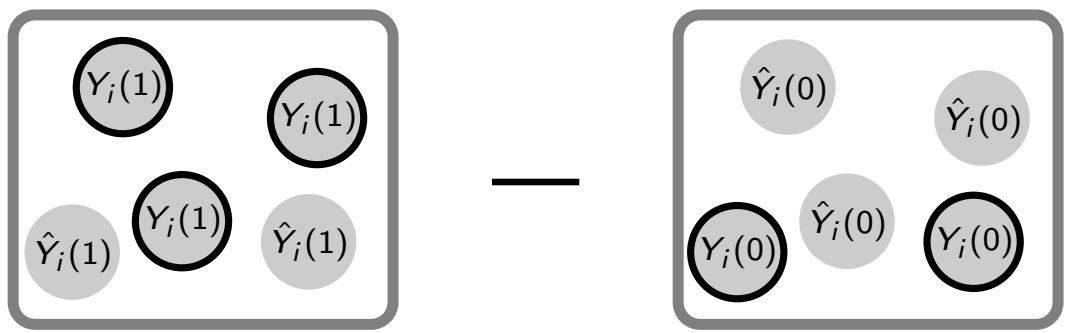
Research framework: Estimands connect theory to evidence



Definition

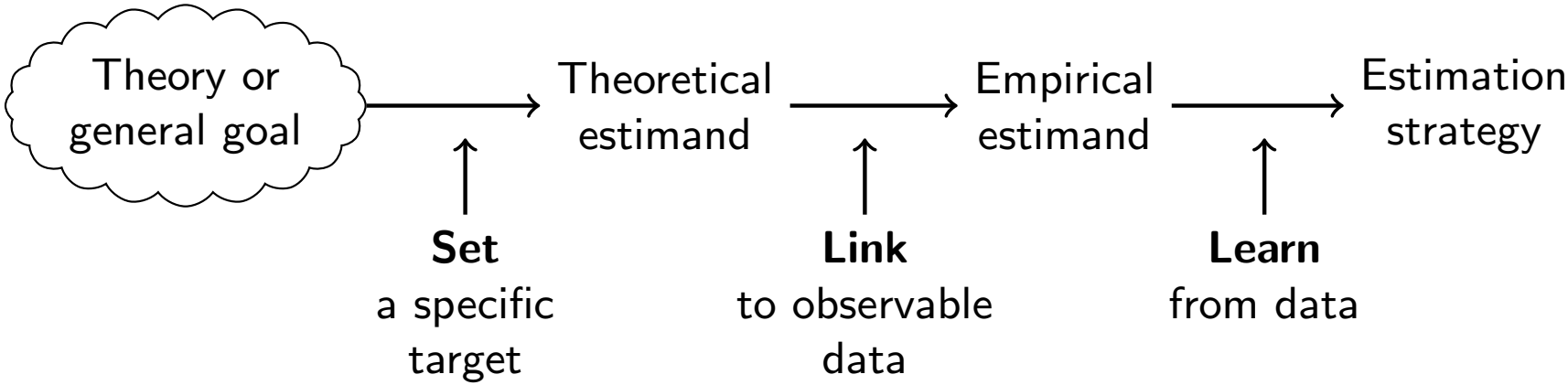
An algorithm applied to data

Example

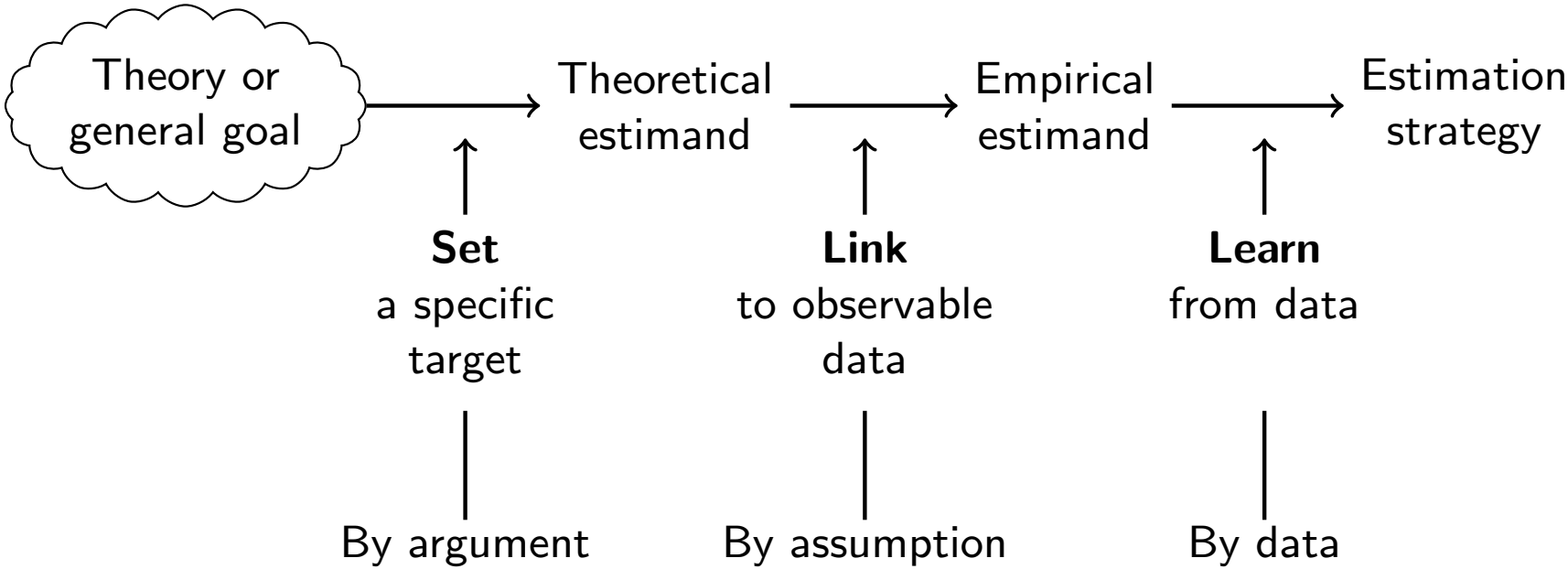


Young 2009, Watts 2014, Berk et al. 2019, Molina and Garip 2019

Research framework: Estimands connect theory to evidence



Research framework: Estimands connect theory to evidence



Defining an estimand

An estimand involves a

- unit-specific quantity
- target population

We will practice with

- simple guiding examples
- then with your projects

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?

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

Defining an estimand: Your project

Form small groups. In your projects,

- What is the unit-specific quantity (or quantities)?
- What is the target population(s)?

Course Intro

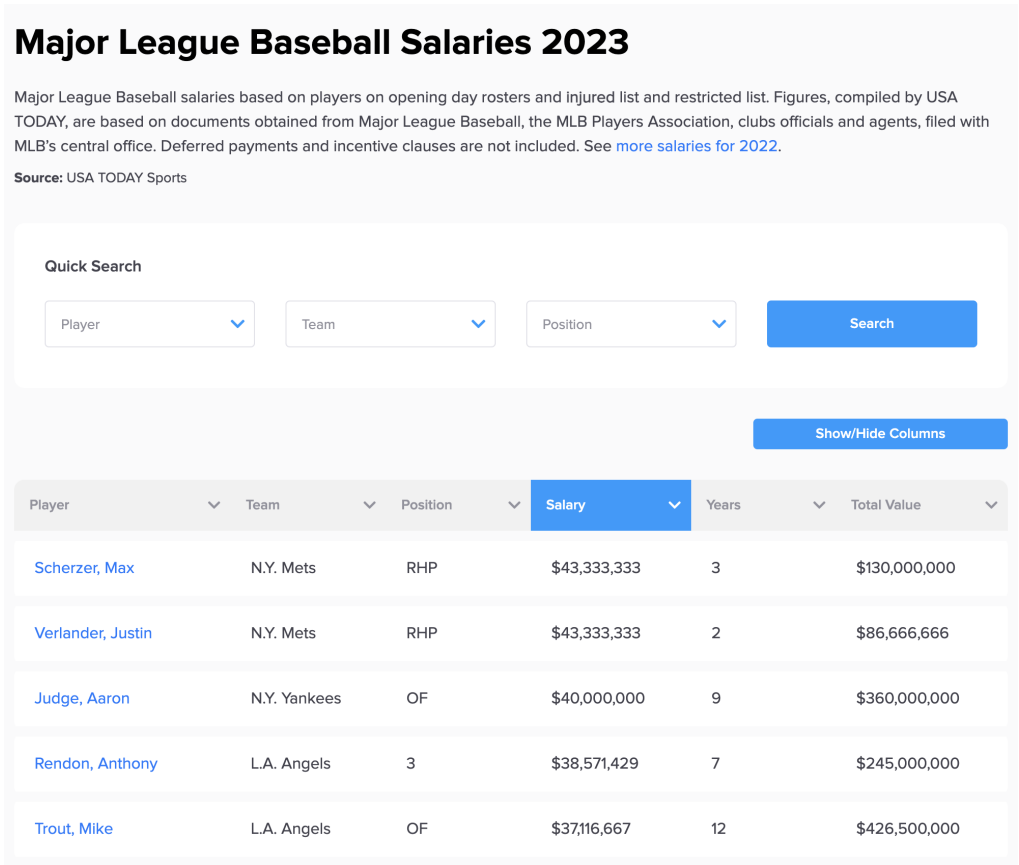
Define an Estimand

$\hat{Y}$ View of Regression

Computer Tutorial

Organizing Your Workflow

Baseball salaries



Major League Baseball Salaries 2023

Major League Baseball salaries based on players on opening day rosters and injured list and restricted list. Figures, compiled by USA TODAY, are based on documents obtained from Major League Baseball, the MLB Players Association, clubs officials and agents, filed with MLB's central office. Deferred payments and incentive clauses are not included. See more salaries for 2022.

Source: USA TODAY Sports

Quick Search

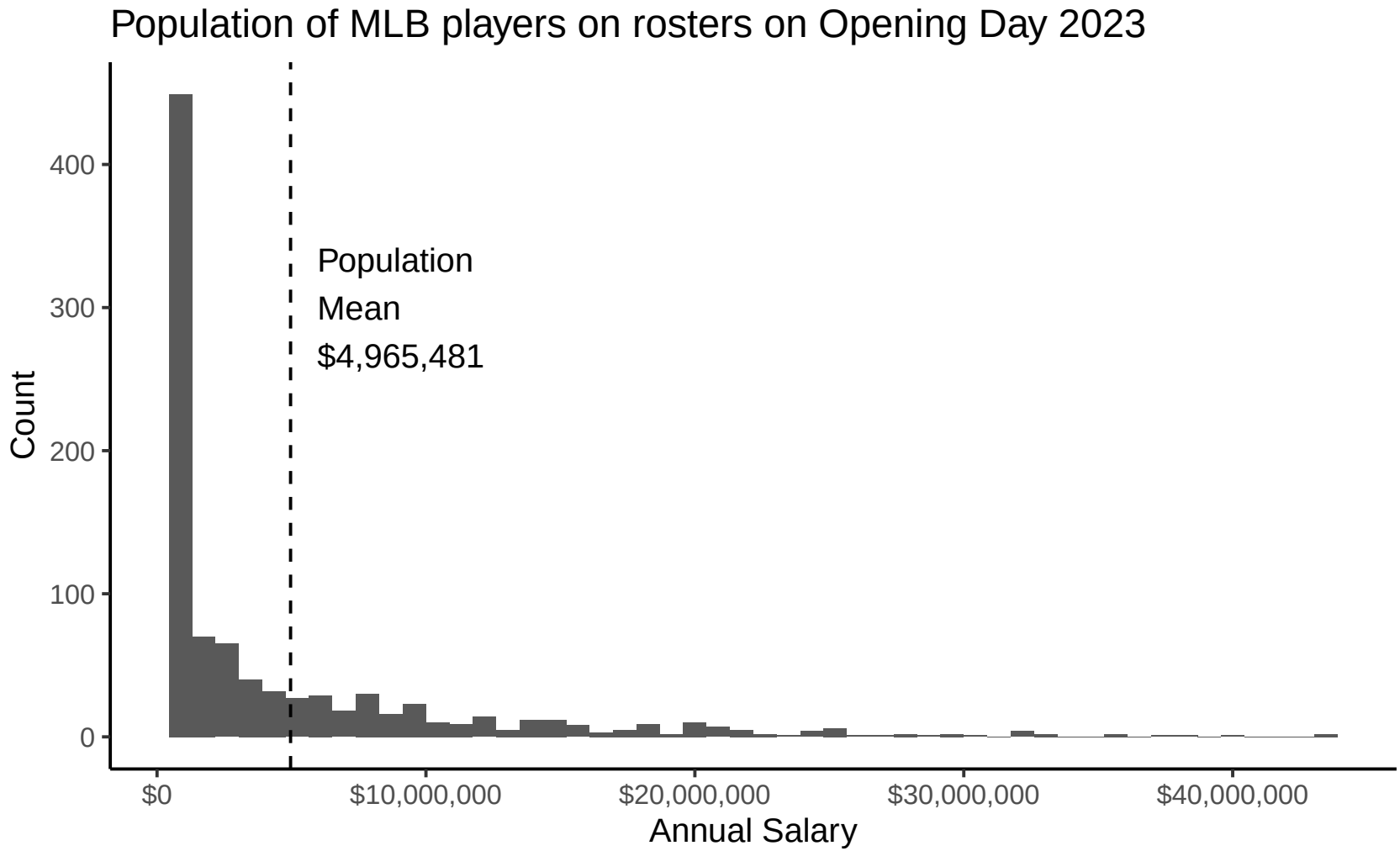
Player | Team | Position | Search

Show/Hide Columns

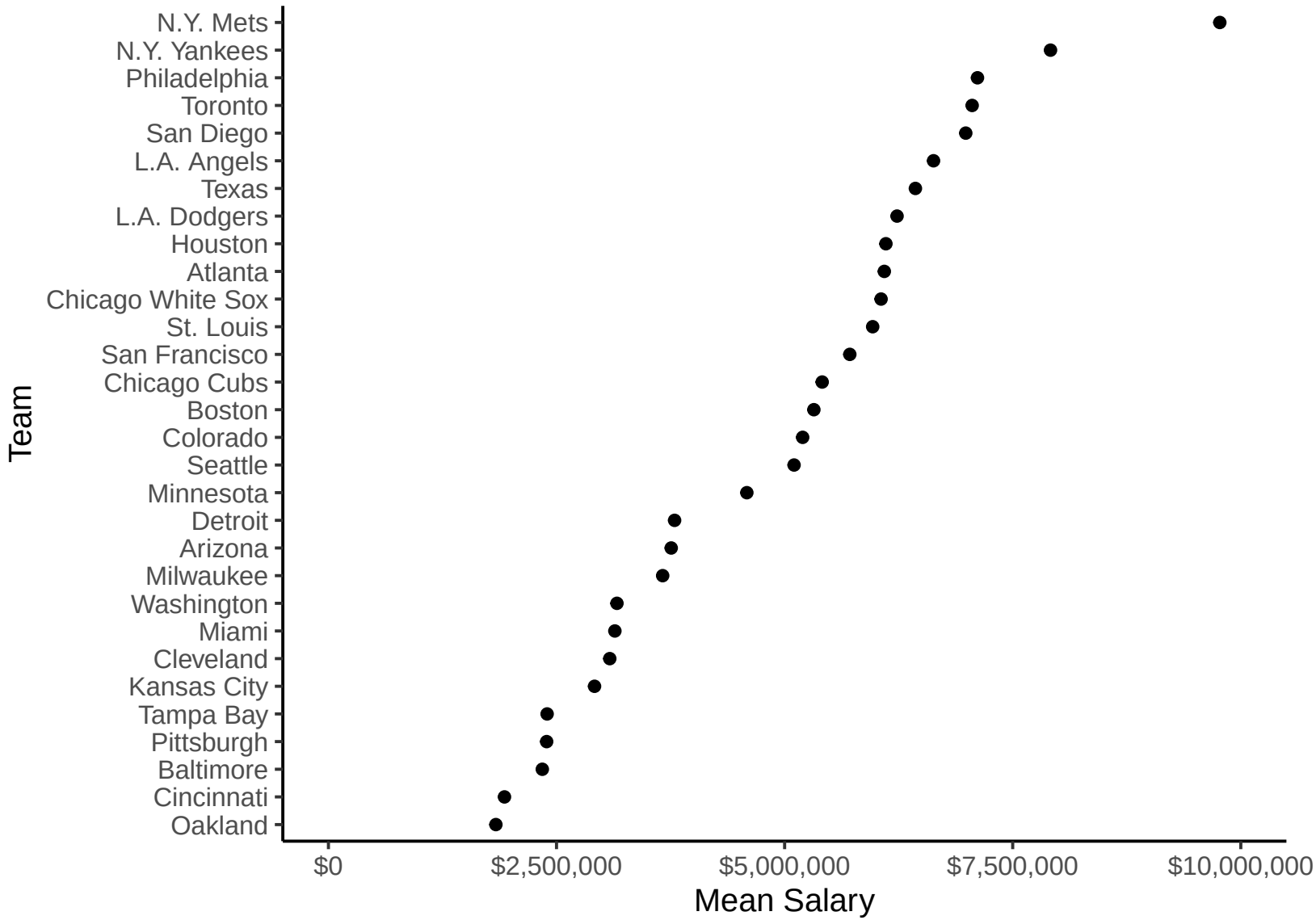
Player	Team	Position	Salary	Years	Total Value
Scherzer, Max	N.Y. Mets	RHP	$43,333,333	3	$130,000,000
Verlander, Justin	N.Y. Mets	RHP	$43,333,333	2	$86,666,666
Judge, Aaron	N.Y. Yankees	OF	$40,000,000	9	$360,000,000
Rendon, Anthony	L.A. Angels	3	$38,571,429	7	$245,000,000
Trout, Mike	L.A. Angels	OF	$37,116,667	12	$426,500,000

databases.usatoday.com/major-league-baseball-salaries-2023/

Baseball salaries

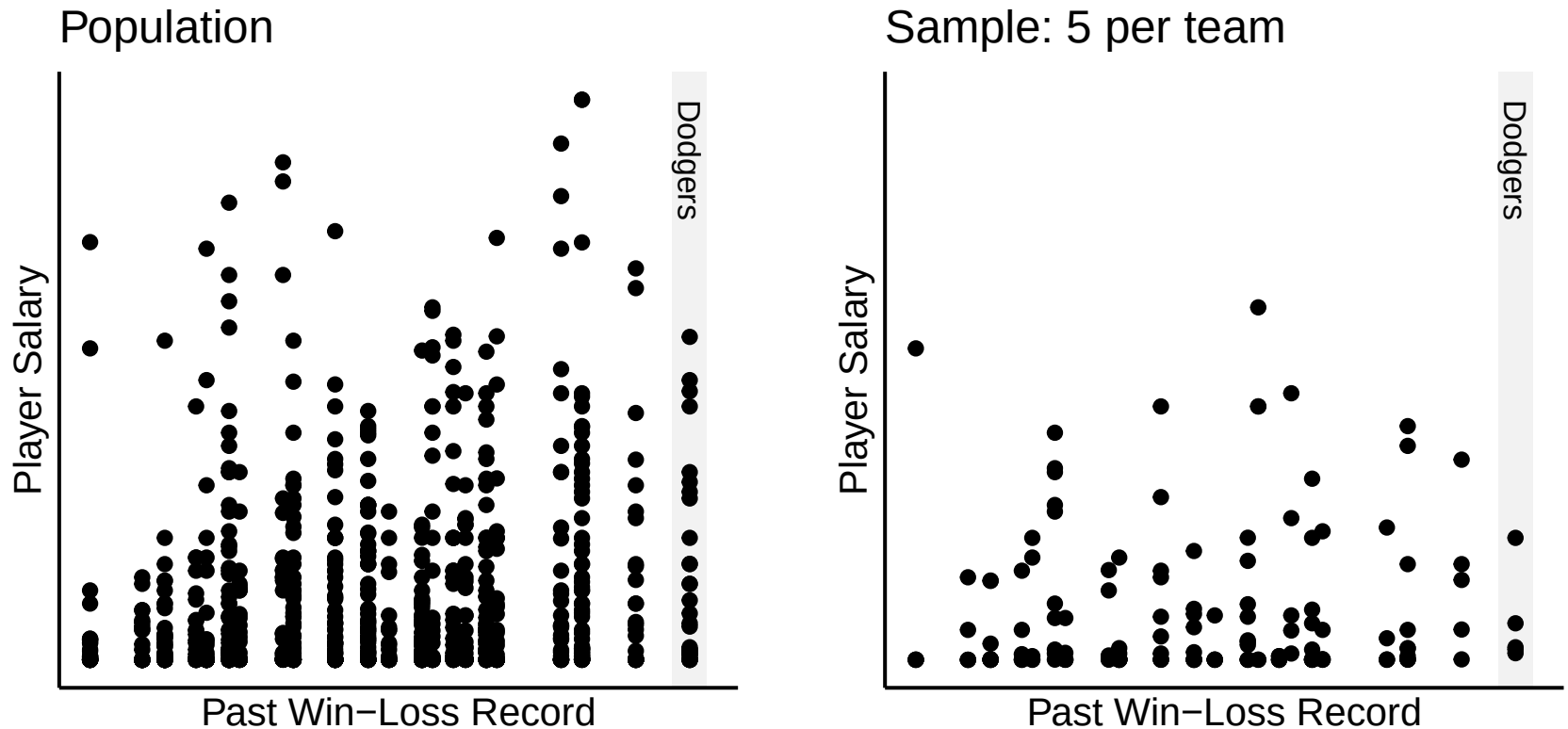


Baseball salaries

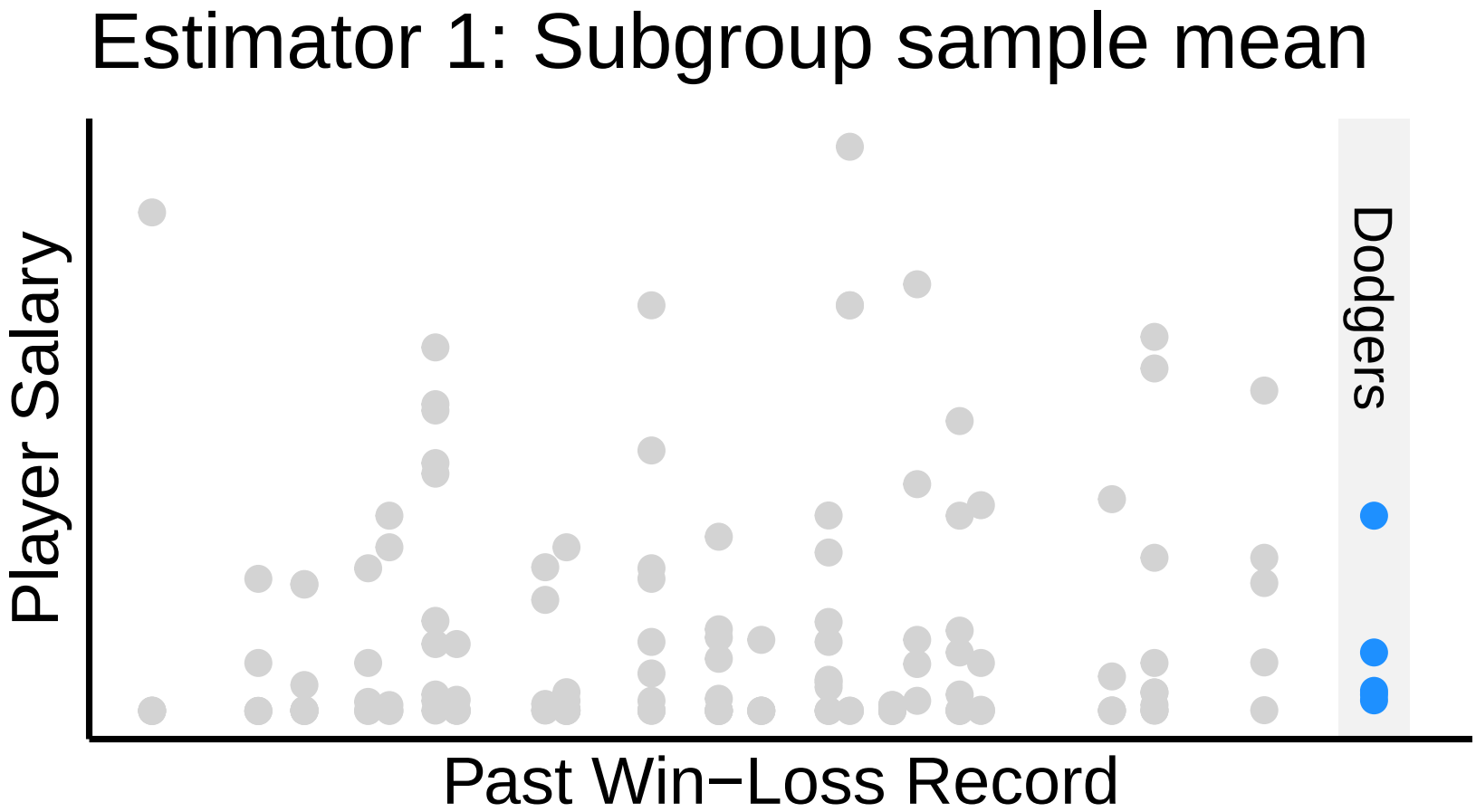


Statistical learning from samples

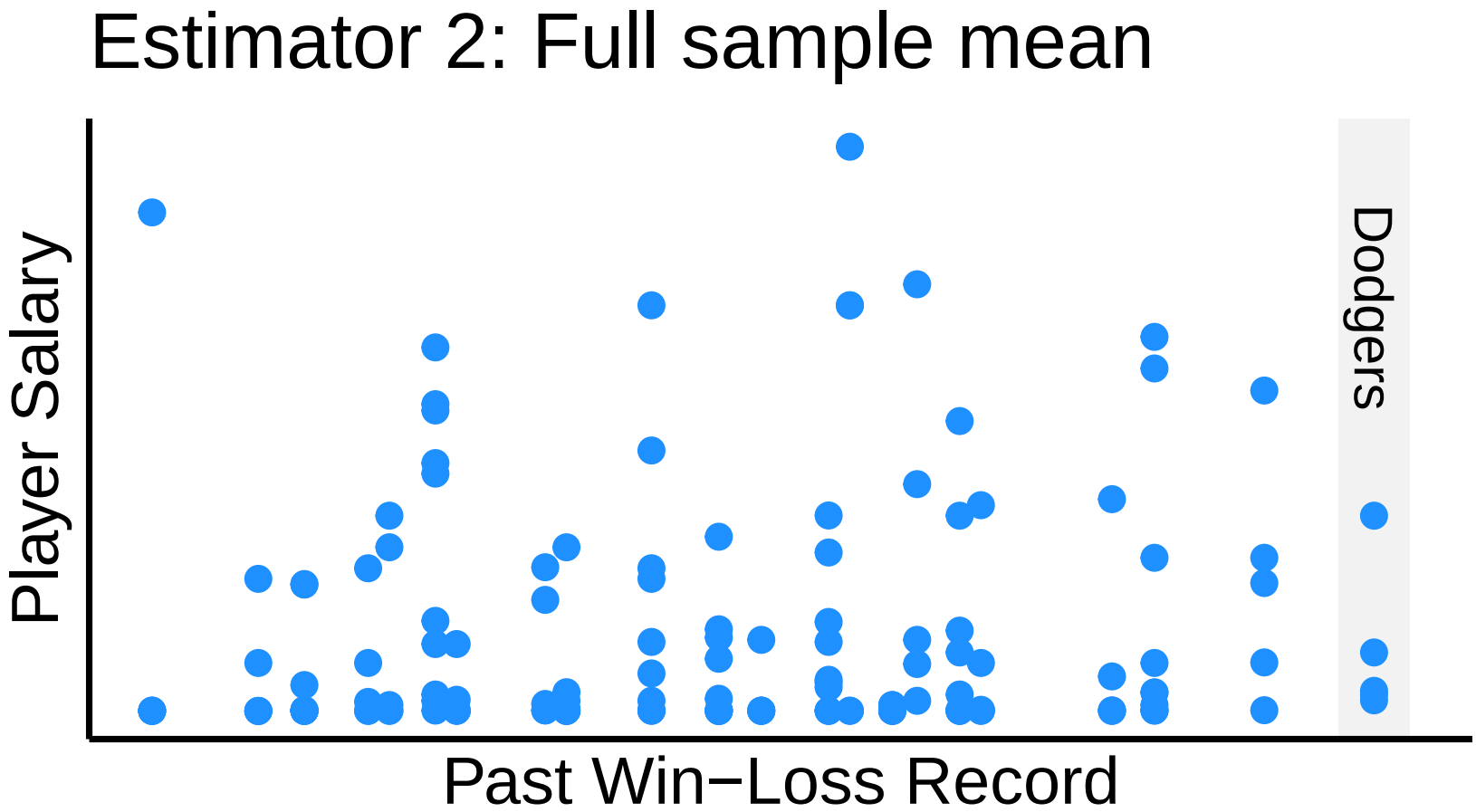
With only the sample, how would you estimate the mean salary of all the Dodgers?



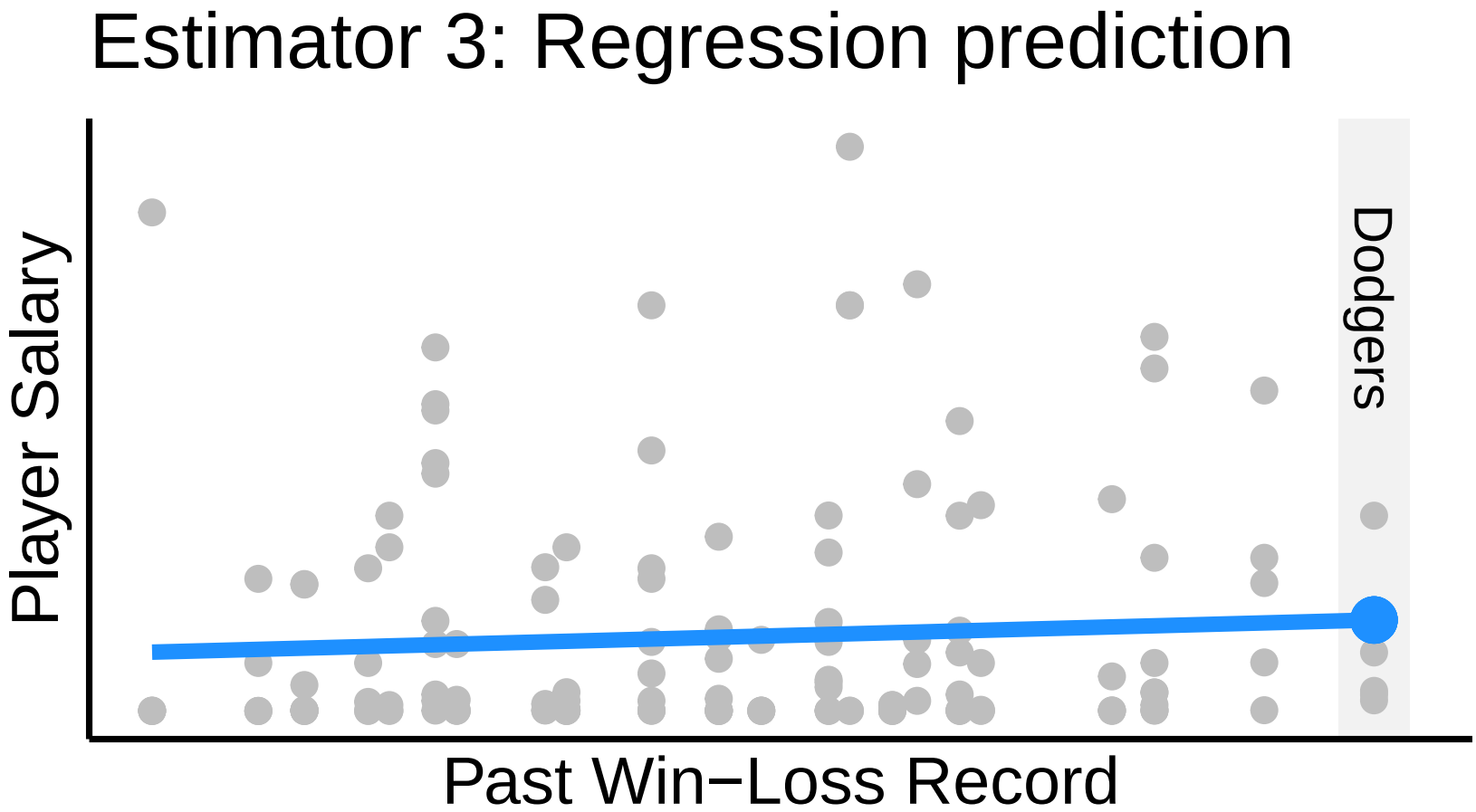
Three estimators for the Dodgers' mean salary



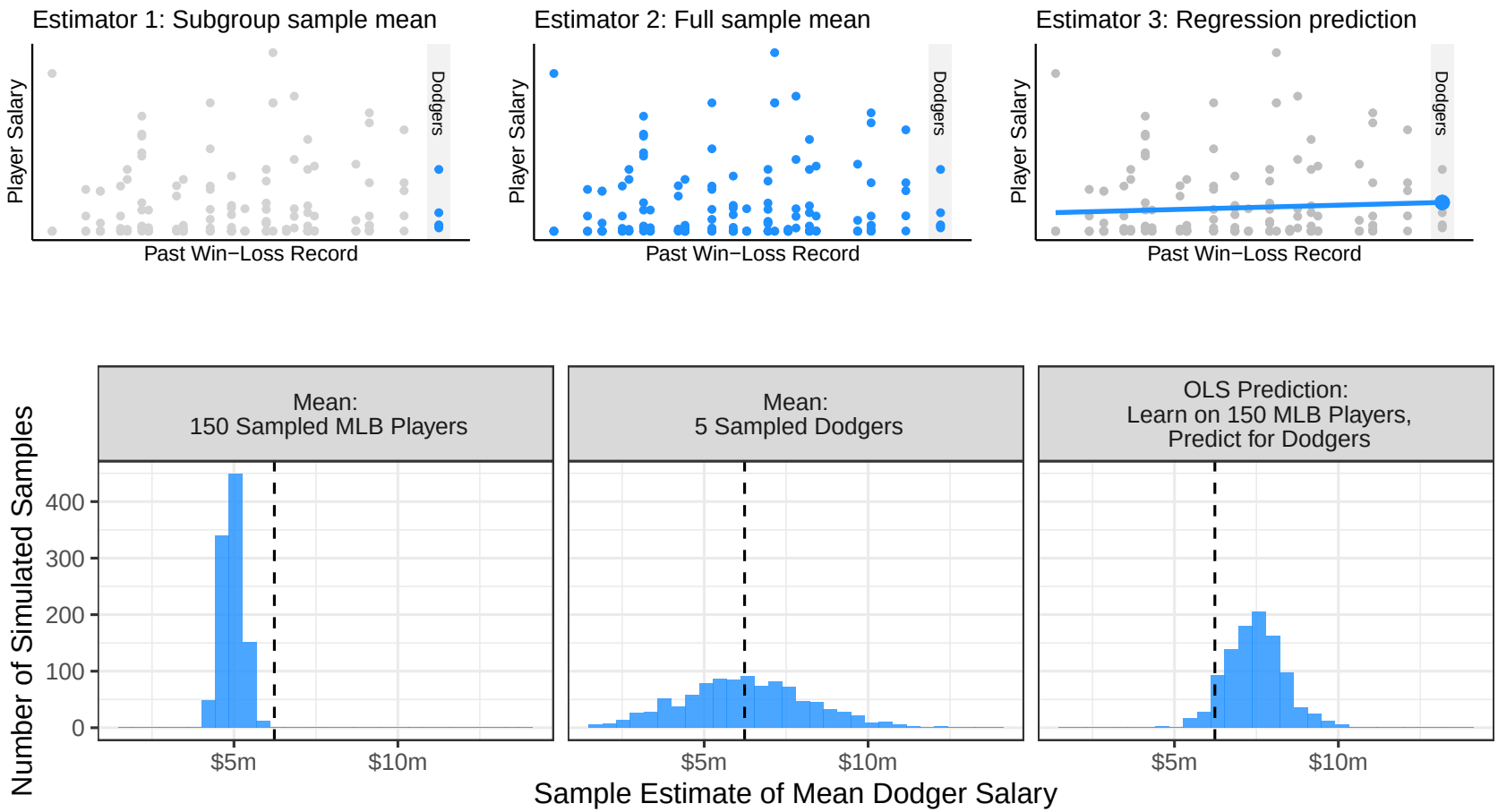
Three estimators for the Dodgers' mean salary



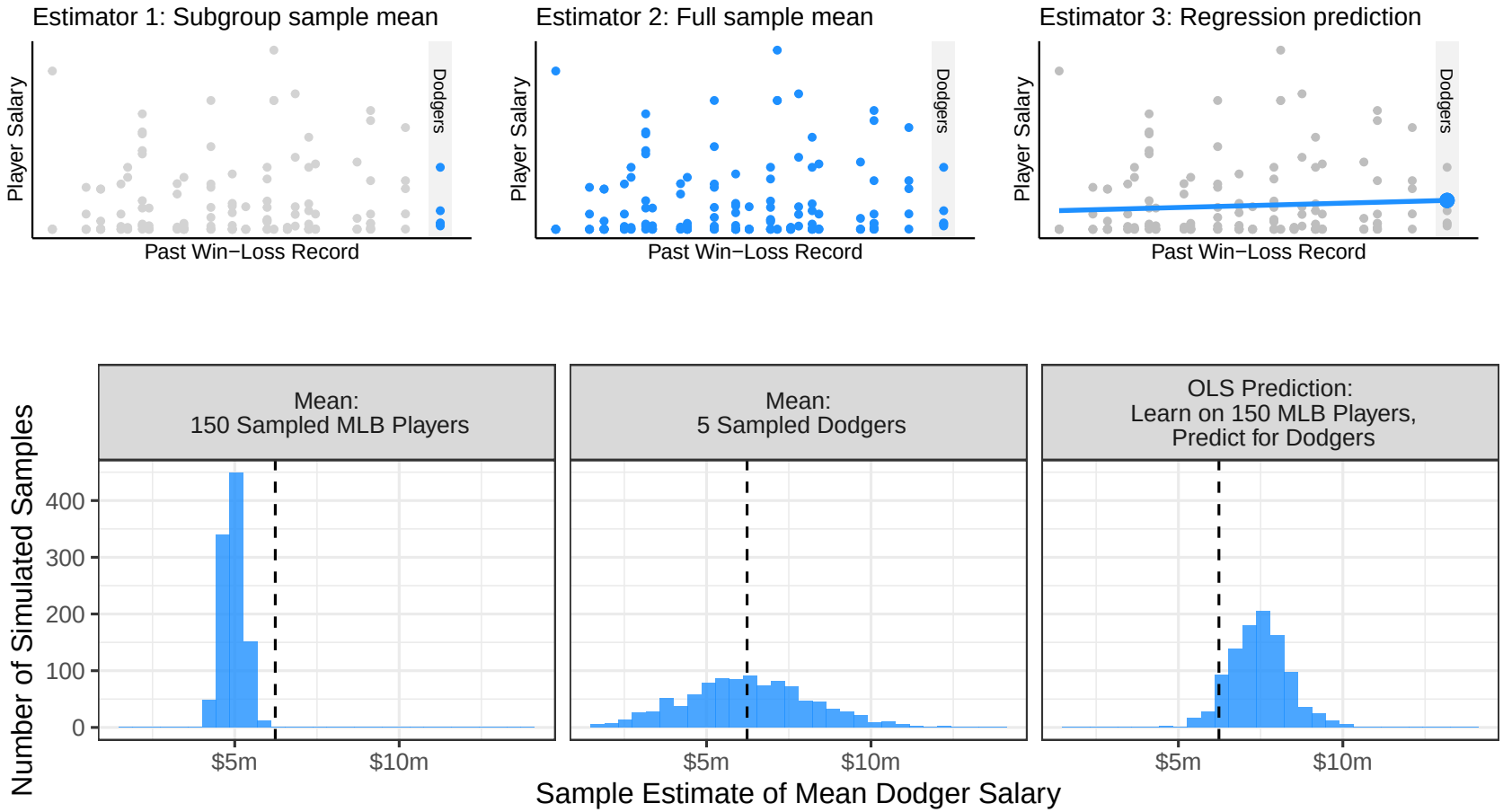
Three estimators for the Dodgers' mean salary



Three estimators for the Dodgers' mean salary



Three estimators for the Dodgers' mean salary



Which do you prefer? Why?

Statistical learning: A somewhat unusual view

Statistical learning: A somewhat unusual view

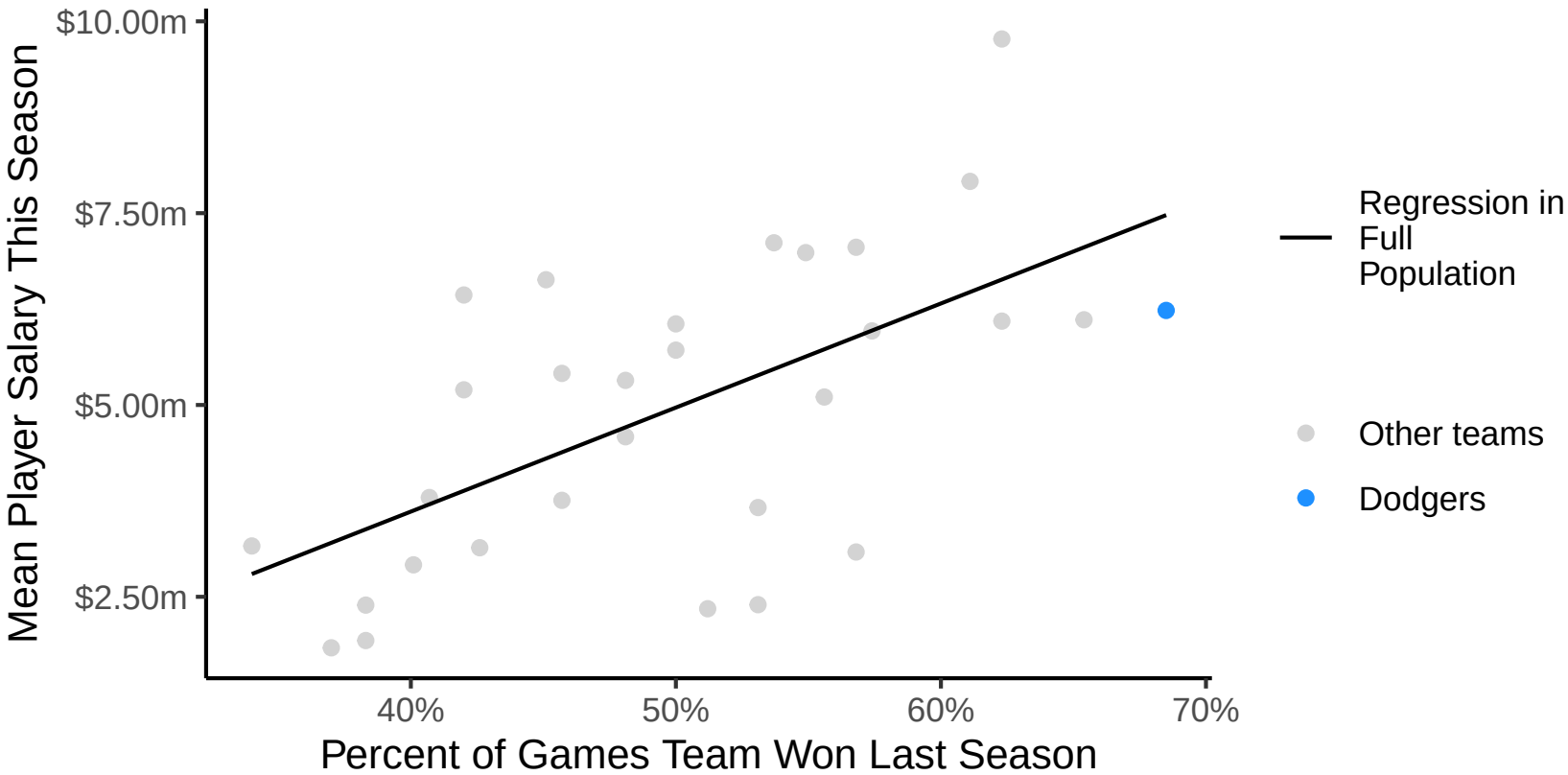
1. the entire goal of modeling is to solve sparse data
   - we sample very few Dodgers,
     so we use non-Dodgers to help our estimate

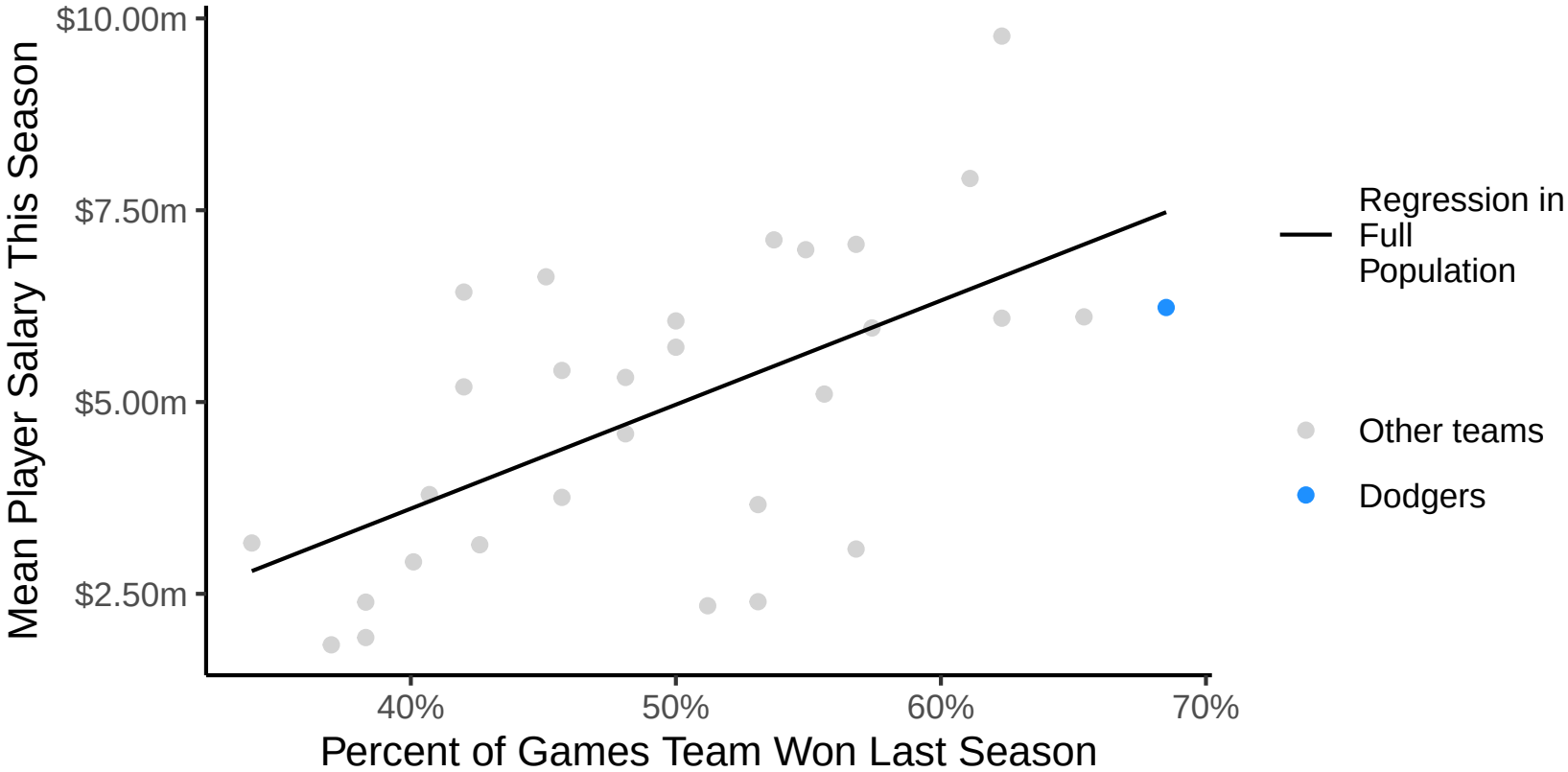
Statistical learning: A somewhat unusual view

1. the entire goal of modeling is to solve sparse data
   - we sample very few Dodgers,
     so we use non-Dodgers to help our estimate
2. in a huge sample, a model is unnecessary
   - estimate Dodger population mean
     by the Dodger sample mean

Statistical learning: A somewhat unusual view

1. the entire goal of modeling is to solve sparse data
   - we sample very few Dodgers,
     so we use non-Dodgers to help our estimate
2. in a huge sample, a model is unnecessary
   - estimate Dodger population mean
     by the Dodger sample mean
3. in a tiny sample, models may perform poorly
   - might even better to estimate a subgroup mean (Dodgers)
     by taking the mean of the whole sample (all MLB)

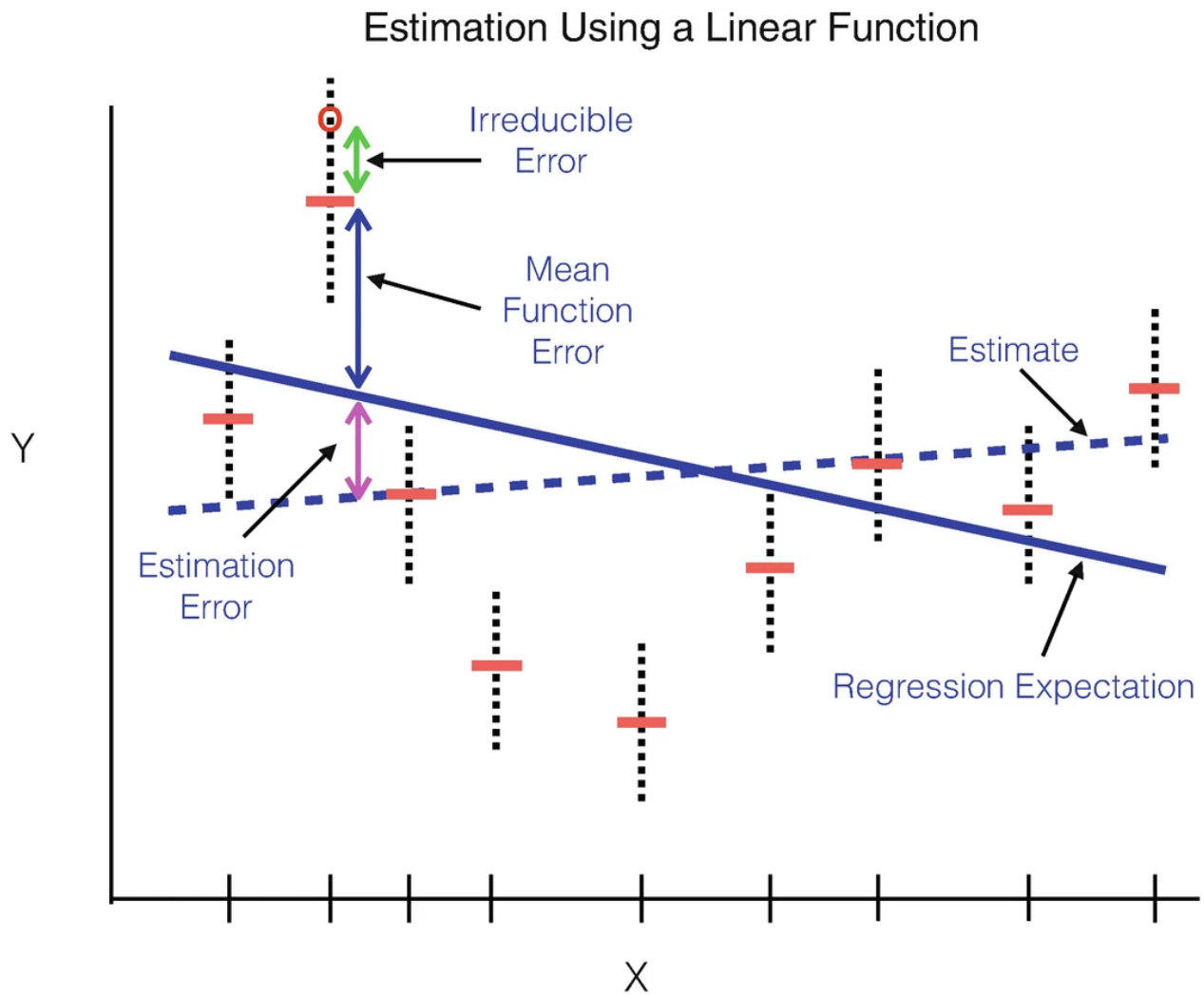

$10.00m
$7.50m
$5.00m
$2.50m
Mean Player Salary This Season
40%
50%
60%
70%
Percent of Games Team Won Last Season
Regression in Full Population
Other teams
Dodgers



The model is wrong. Why might we still use it?

$\hat{Y}$ view of regression: Modeling errors

Berk Fig 1.6



Course Intro

Define an Estimand

$\hat{Y}$ View of Regression

Computer Tutorial

Organizing Your Workflow

Course Intro

Define an Estimand

$\hat{Y}$ View of Regression

Computer Tutorial

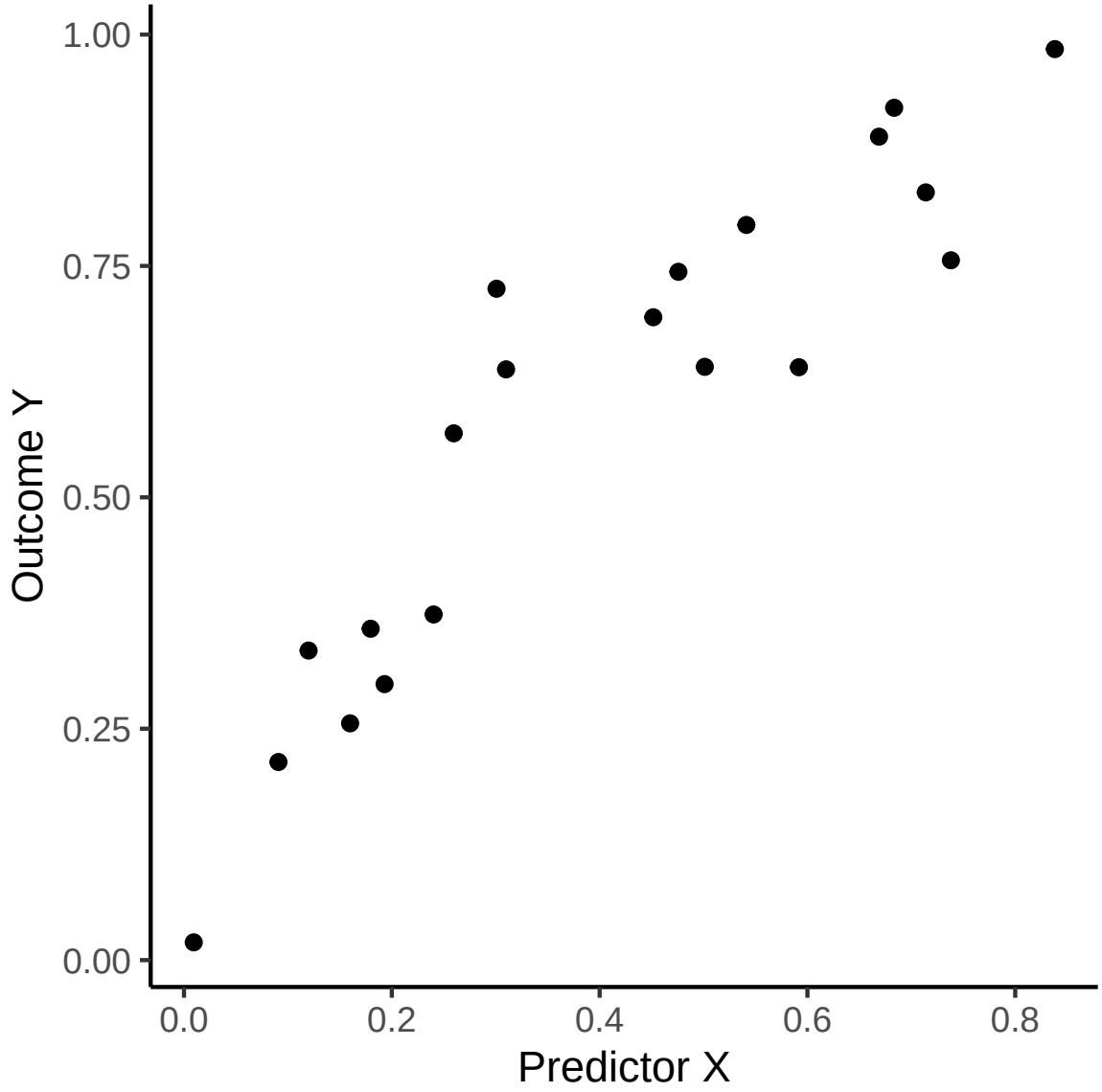
Organizing Your Workflow

A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA

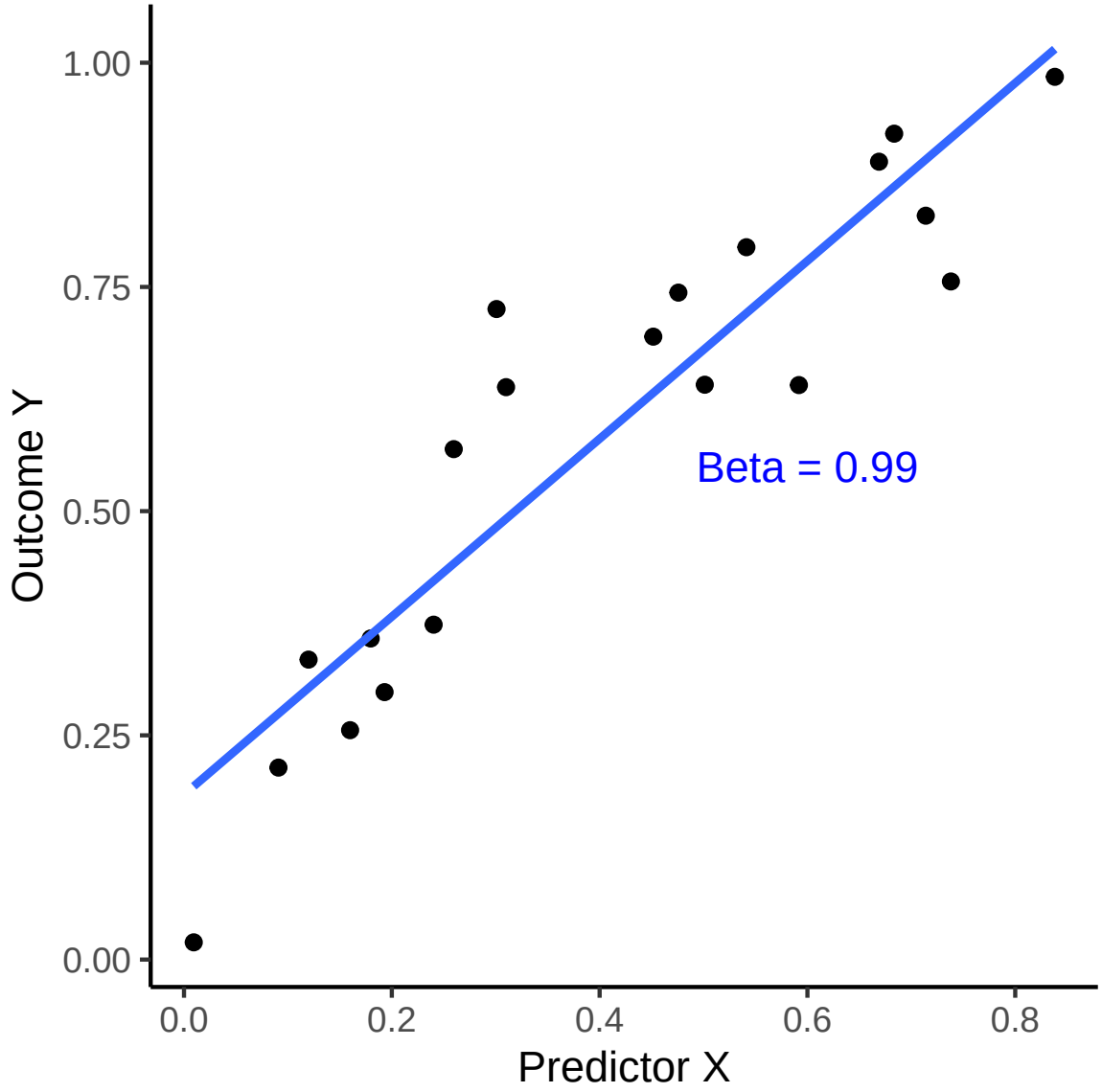
A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA



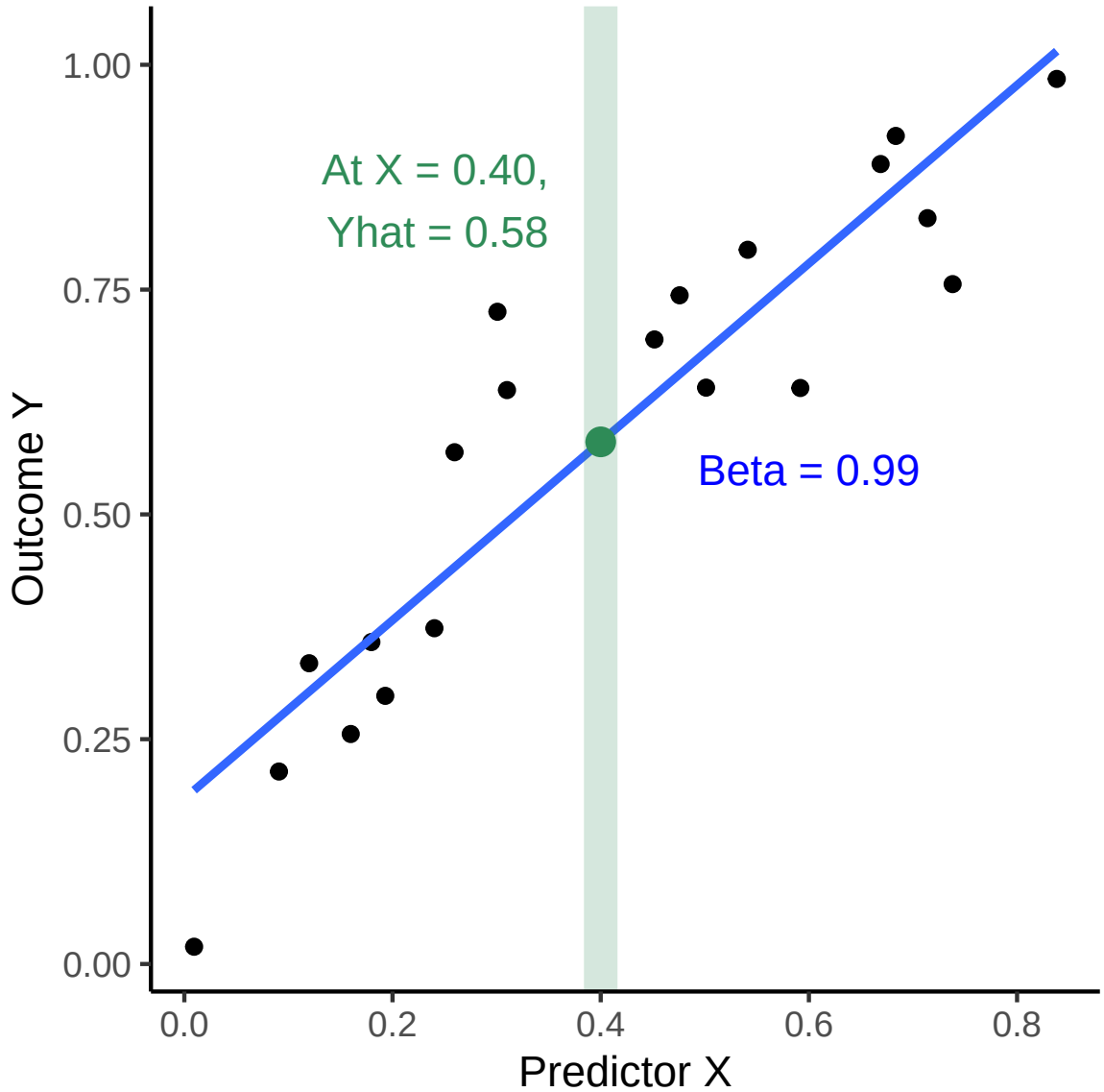
A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA



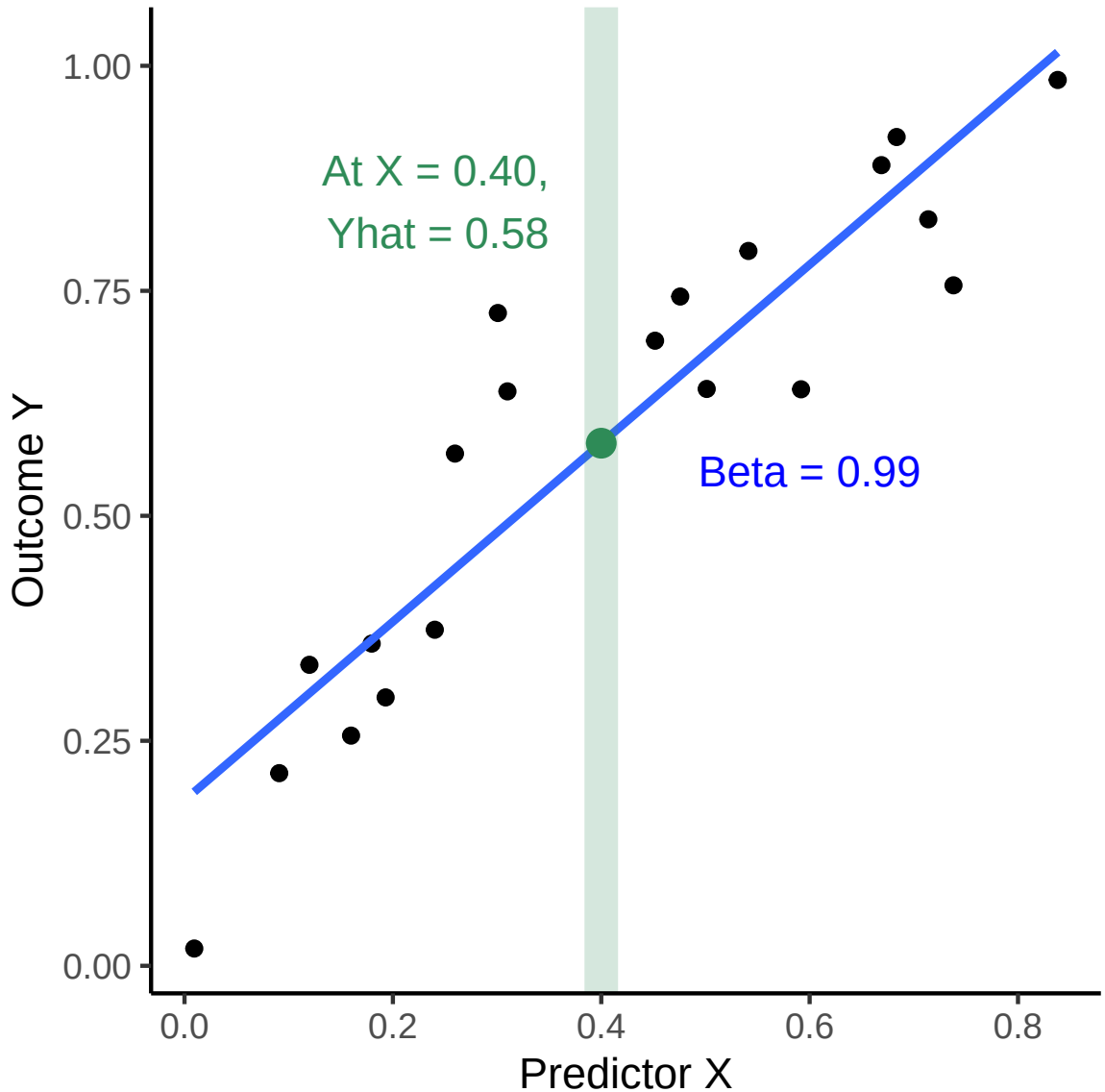
A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA



A $\hat{Y}$ view of description: Predict a subgroup mean

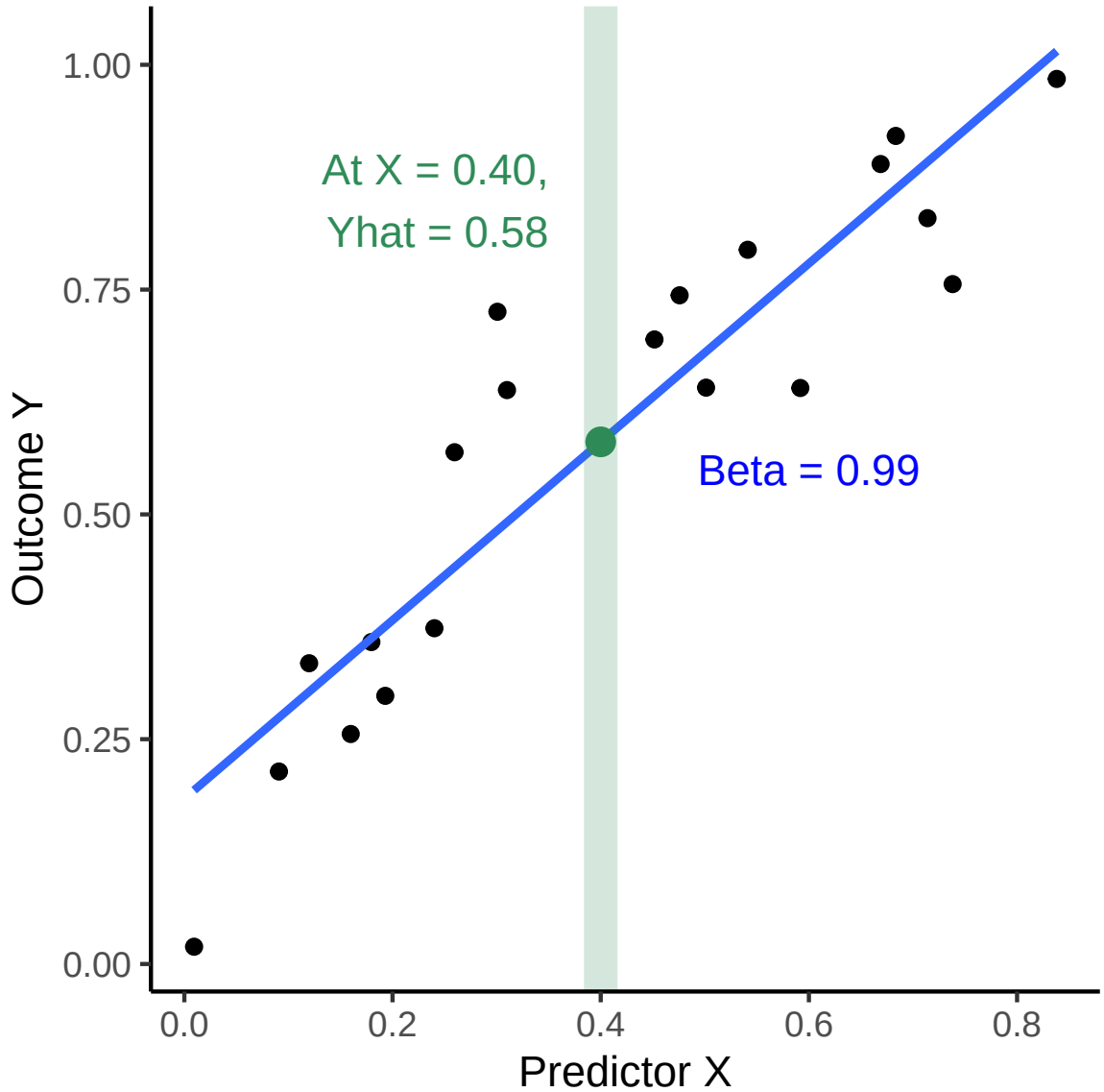
With Kristin Liao, UCLA



Concrete steps

A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA

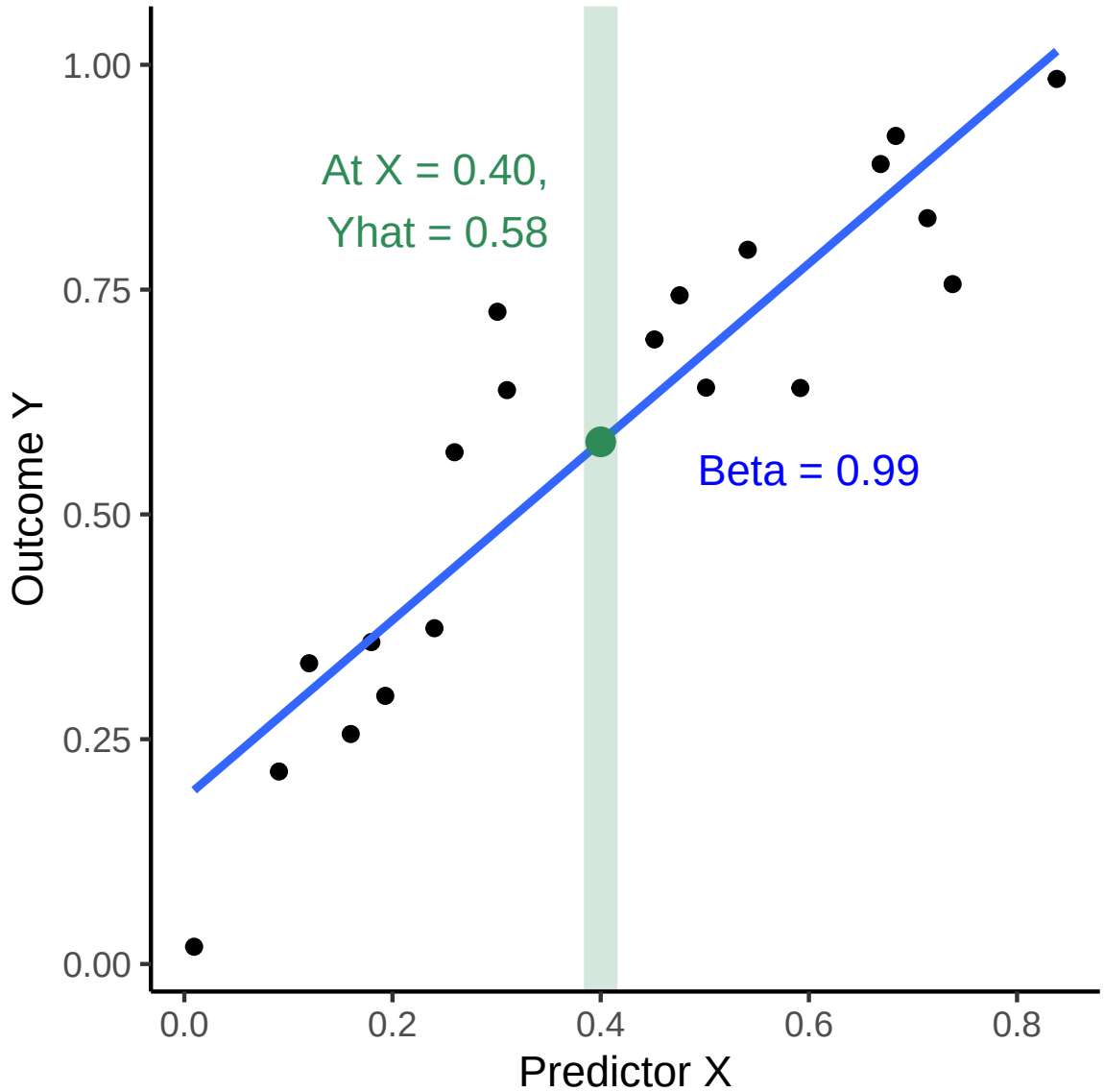


Concrete steps

In full data, learn a model

A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA



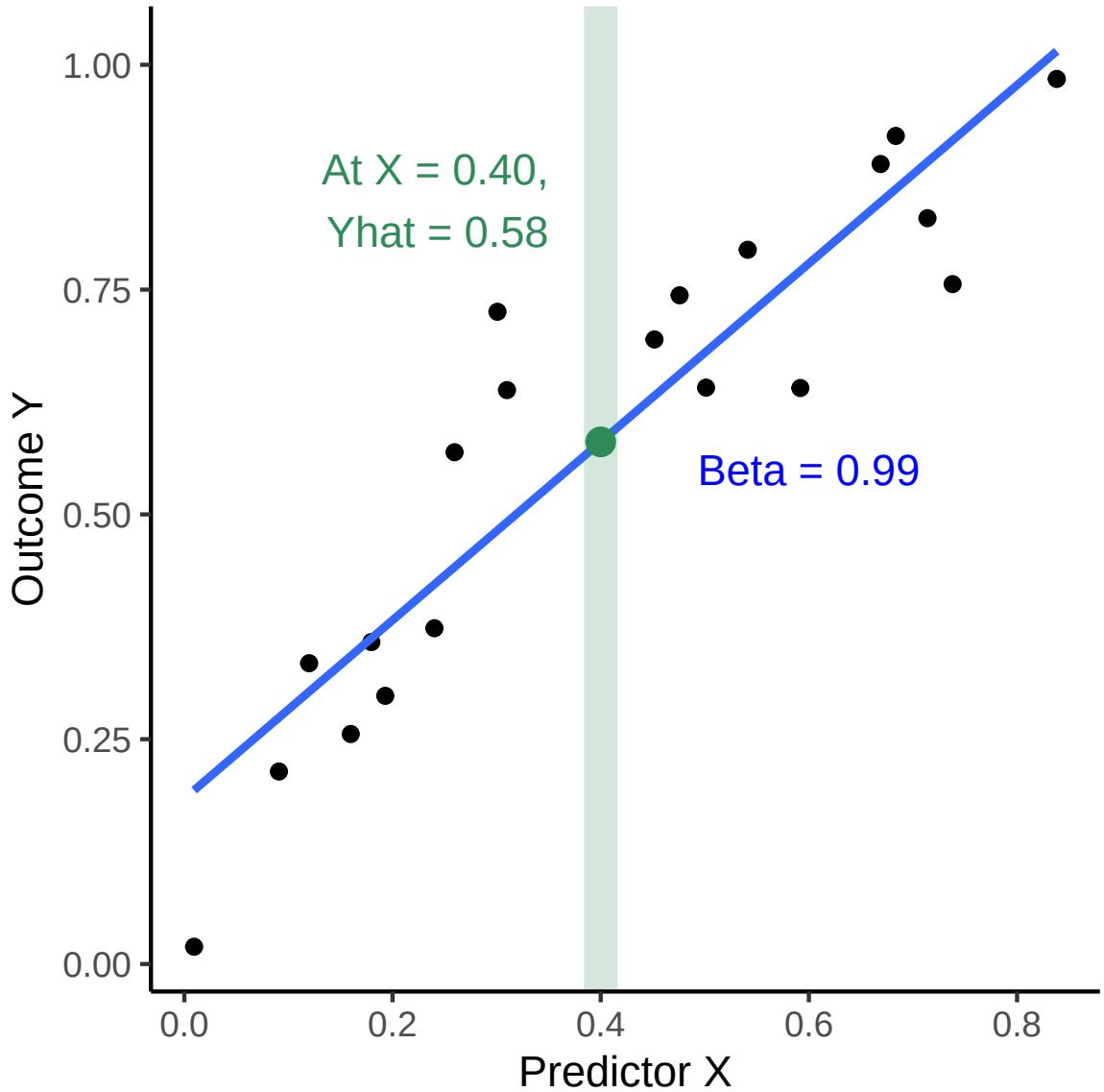
Concrete steps

In full data, learn a model

Define new data in which to predict

A $\hat{Y}$ view of description: Predict a subgroup mean

With Kristin Liao, UCLA



Concrete steps

In full data, learn a model

Define new data in which to predict

Report prediction

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

Data for learning

- 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

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.

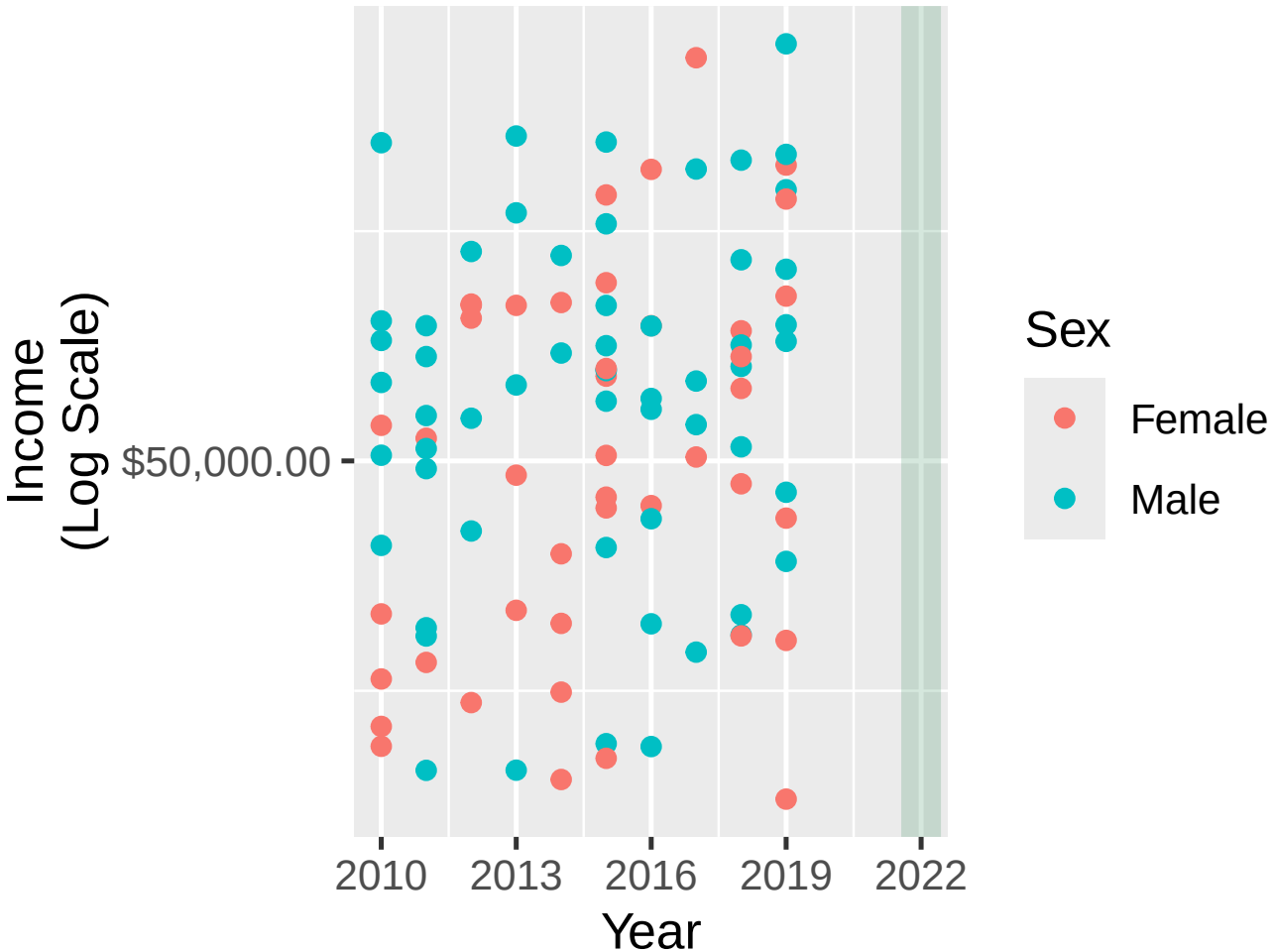
```
simulate <- function(n = 100) {
  read_csv("https://ilundberg.github.io/description/assets/truth.csv") |>
    slice_sample(n = n, weight_by = weight, replace = T) |>
    mutate(income = exp(rnorm(n(), meanlog, sdlog))) |>
    select(year, age, sex, income)
}
```

We can see how it works below.

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

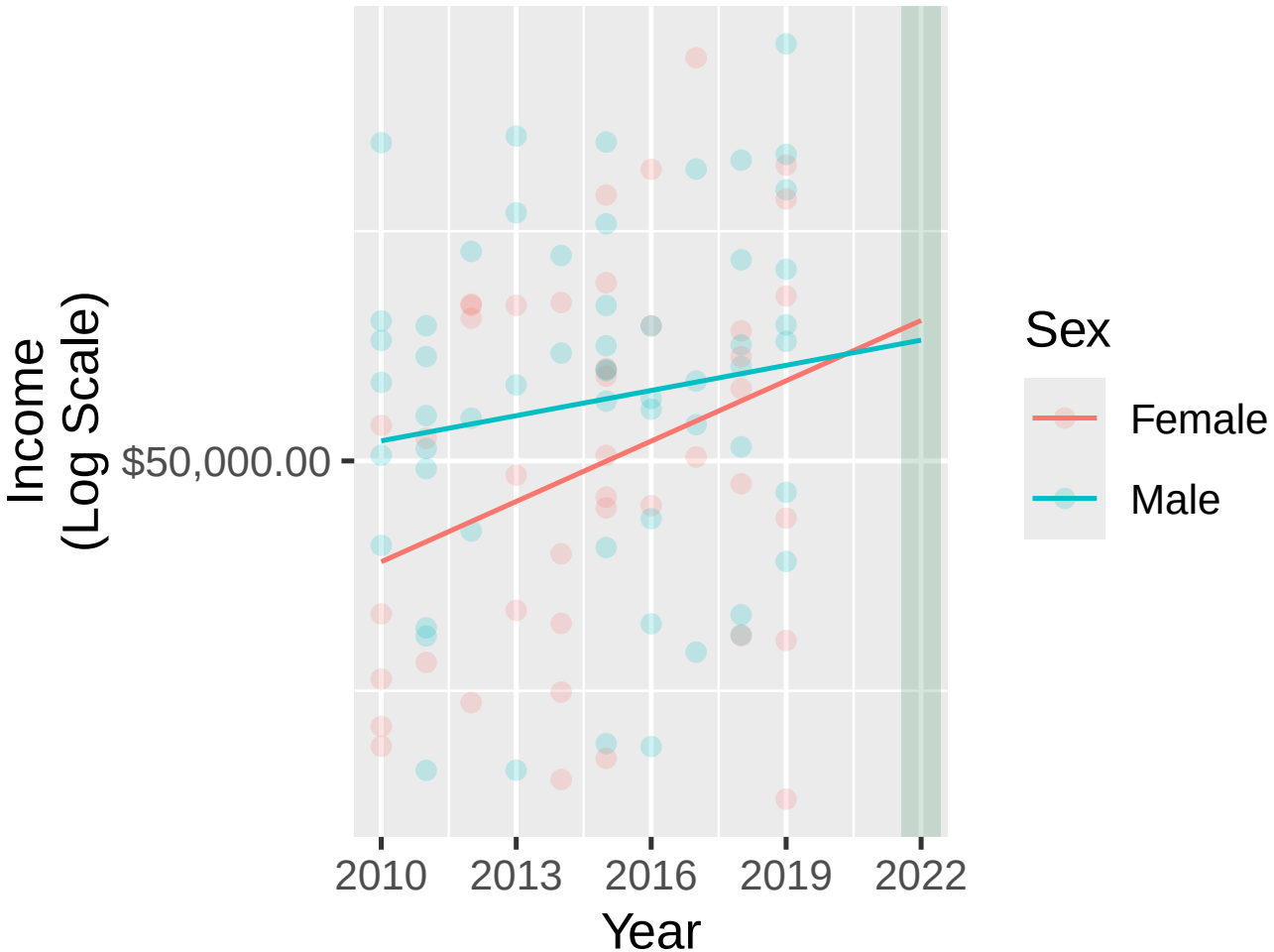
Computer tutorial: Introduction

ilundberg.github.io/description



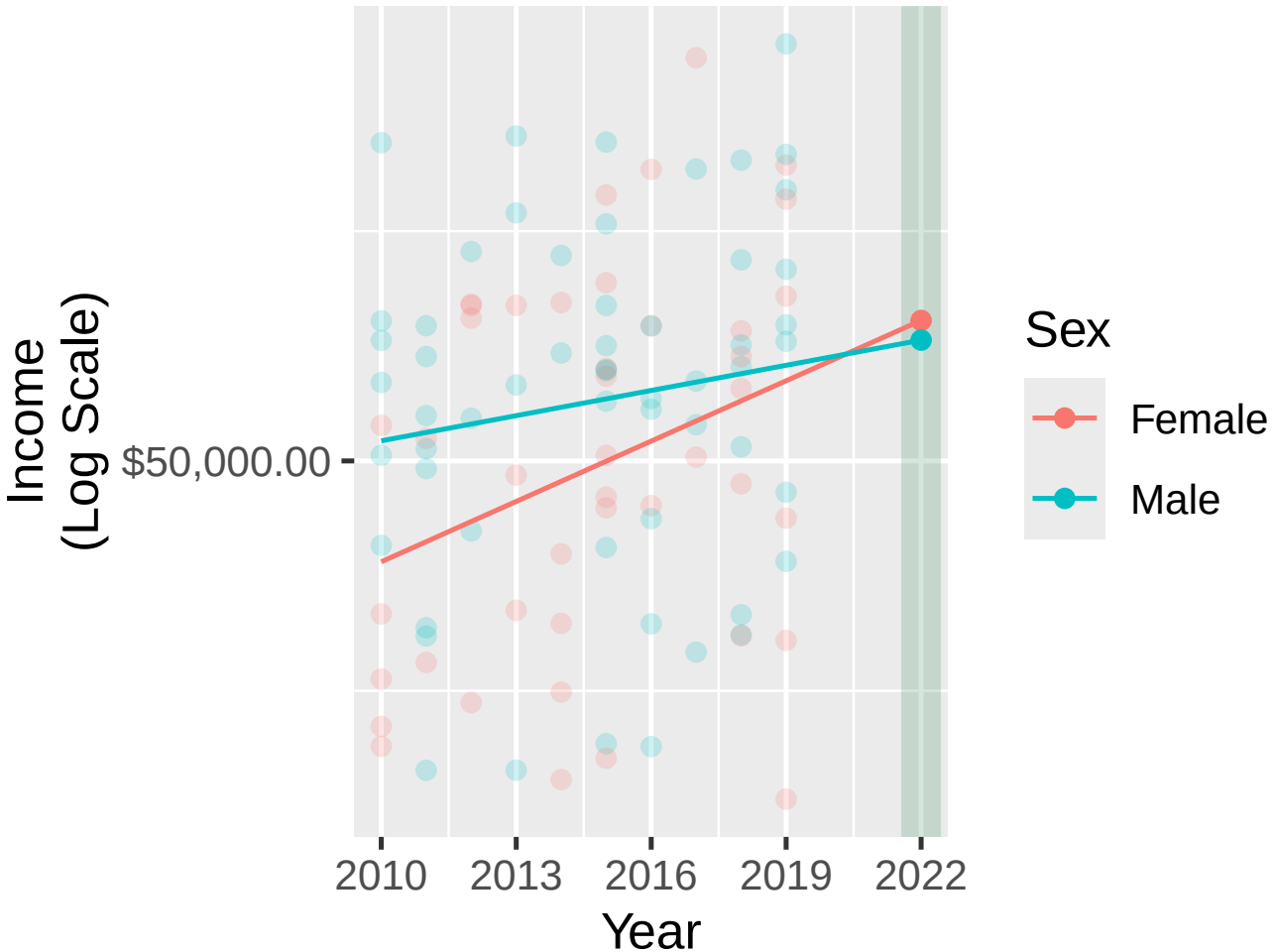
Computer tutorial: Introduction

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

When you finish:

- How could you use regression to estimate a subgroup mean in your own project?

Course Intro

Define an Estimand

$\hat{Y}$ View of Regression

Computer Tutorial

Organizing Your Workflow

Course Intro

Define an Estimand

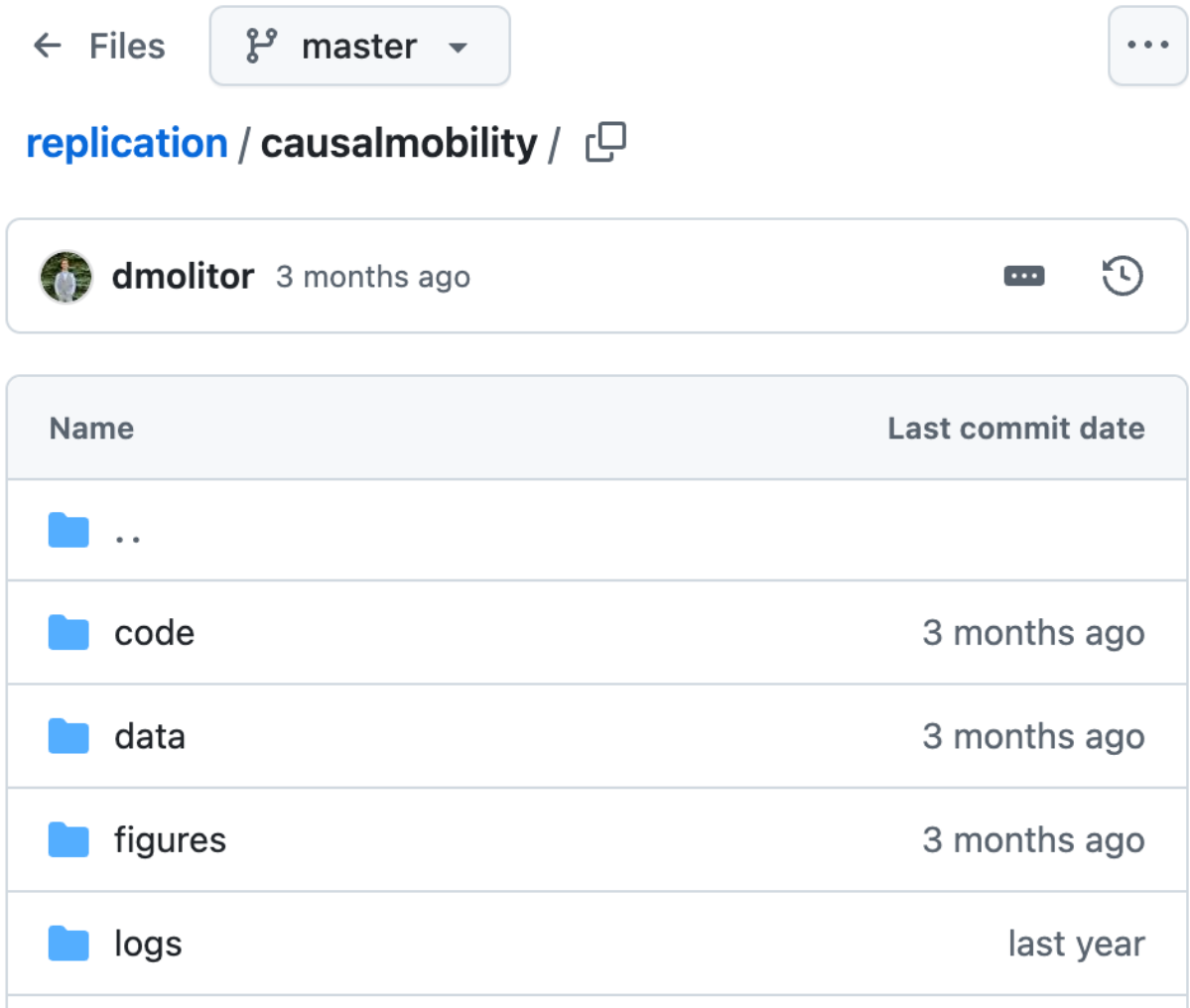
$\hat{Y}$ View of Regression

Computer Tutorial

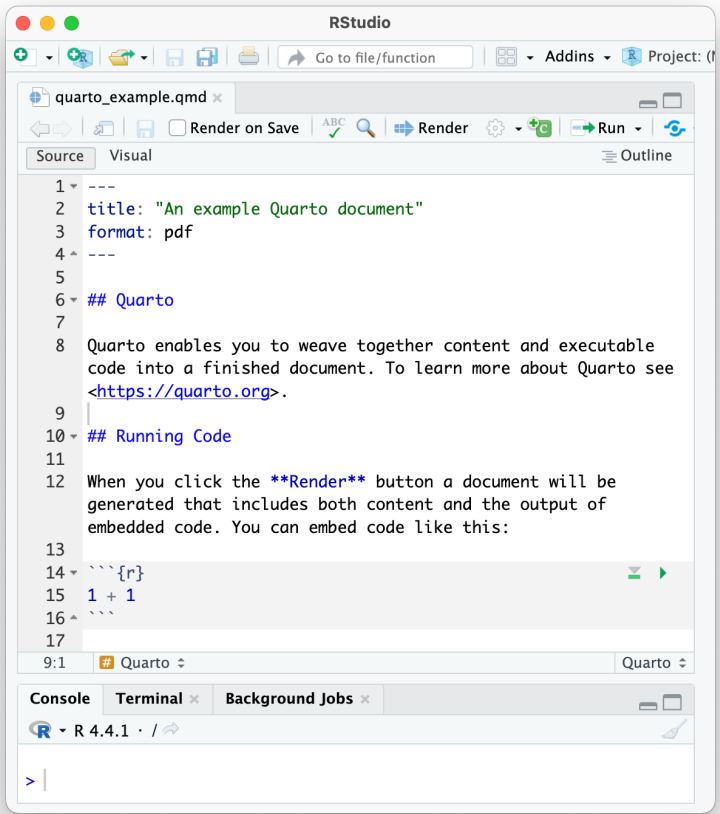
Organizing Your Workflow

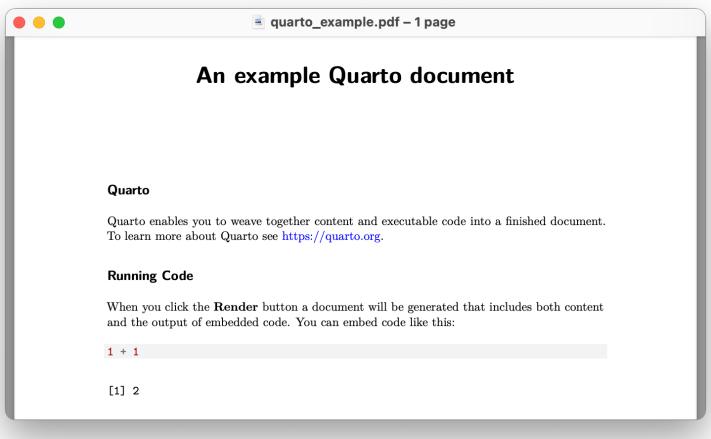
Organizing your workflow: Code scripts

I organize a project folder like this:



Organizing your workflow: Quarto documents





see the RStudio Quarto tutorial

Learning goals for today

By the end of class, you will be able to

- define an estimand in your project
  - unit-specific quantity
  - target population
- motivate regression from a $\hat{Y}$ view
  - as a tool to estimate despite sparse data
  - with the risk of various modeling errors
- make predictions to describe population subgroups
- organize your code in directories