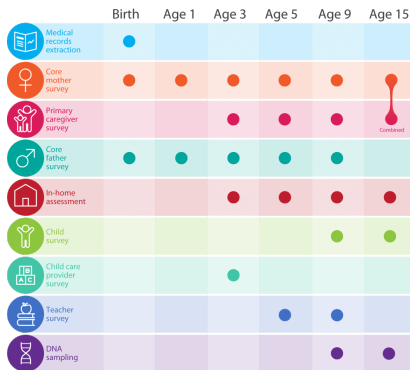# Making the most of the Fragile Families and Child Wellbeing Study

Ian Lundberg

Department of Sociology
Office of Population Research
Princeton University

ilundberg at princeton dot edu

10 April 2019
Annual Meeting of the Population Association of America
Slides posted: https://github.com/ilundberg/slides

A **temptation**

Study **all causes** of some outcome by throwing them all in a model.

A **temptation**

Study **all causes** of some outcome by throwing them all in a model.

← There are **so many** variables!

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

⟵ There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

There are
so many
variables!

←

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

Suppose you study the process that leads to **material hardship**.

Suppose you study the process that leads to **material hardship**.

*Coefficients made up for example.

Black	.05
Hispanic	.07
Mother < HS	.02
Past material hardship	.56
Income	-.43
Multigenerational household	-.43
Married parents	-.43
Mother depressed	.12
Father depressed	.14
Mother unemployed	.67
Father unemployed	.74

Table 1. Predictors of material hardship.

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

*Coefficients made up for example.

Black	.05
Hispanic	.07
Mother < HS	.02
Past material hardship	.56
Income	-.43
Multigenerational household	-.43
Married parents	-.43
Mother depressed	.12
Father depressed	.14
Mother unemployed	.67
Father unemployed	.74

You may be tempted to interpret **all** the coefficients.

Table 1. Predictors of material hardship.

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

*Coefficients made up for example.

Black	.05
Hispanic	.07
Mother < HS	.02
Past material hardship	.56
Income	-.43
Multigenerational household	-.43
Married parents	-.43
Mother depressed	.12
Father depressed	.14
Mother unemployed	.67
Father unemployed	.74

Table 1. Predictors of material hardship.

You may be tempted to interpret **all** the coefficients.

That is unwise (causal inference)

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

*Coefficients made up for example.

Black	.05
Hispanic	.07
Mother < HS	.02
Past material hardship	.56
Income	-.43
Multigenerational household	-.43
Married parents	-.43
Mother depressed	.12
Father depressed	.14
Mother unemployed	.67
Father unemployed	.74

You may be tempted to interpret **all** the coefficients.

That is unwise (causal inference)

It is also futile. A structural model of the outcome is a **losing battle**.

Table 1. Predictors of material hardship.

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

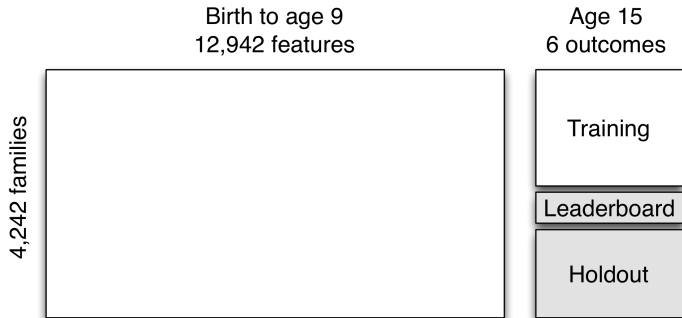— We invited hundreds of social and data scientists to try

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

— We invited hundreds of social and data scientists to try

Birth to age 9
12,942 features
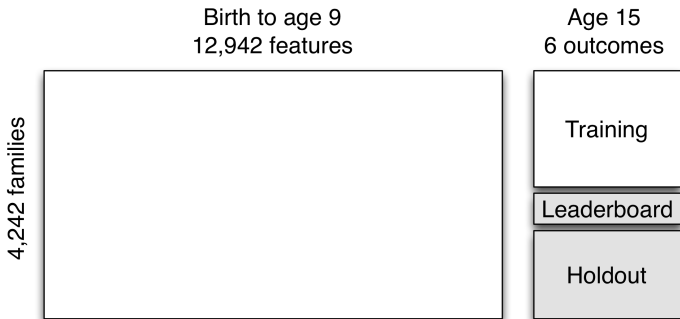
Age 15
6 outcomes

4,242 families

Training

Leaderboard

Holdout

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

— We invited hundreds of social and data scientists to try

— They used many of the variables and many strategies

This slide is about the Fragile Families Challenge. Results are not yet published.
You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

— We invited hundreds of social and data scientists to try

— They used many of the variables and many strategies
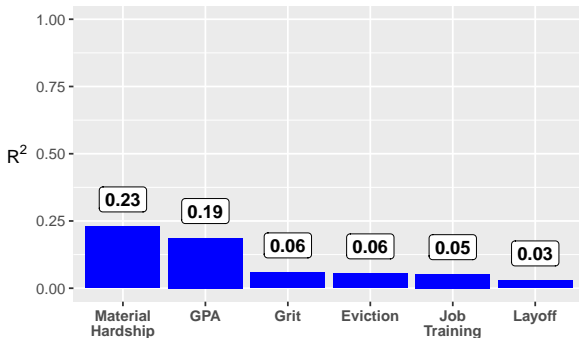
— Predictive performance was poor

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

— We invited hundreds of social and data scientists to try

— They used many of the variables and many strategies

— Predictive performance was poor

This slide is about the Fragile Families Challenge. Results are not yet published. You can read more here or watch a video overview here.

Suppose you study the process that leads to **material hardship**.

Problem: We cannot even predict material hardship

— We invited hundreds of social and data scientists to try

— They used many of the variables and many strategies

— Predictive performance was poor

> If you aim to
>
> — **predict** an outcome well (our goal) or
> — understand **all causes** of an outcome (harder)
>
> you are likely to be **disappointed** with these data.

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

⟵

There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
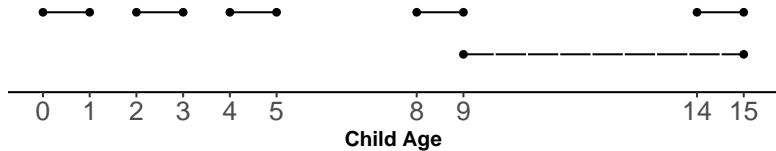that can be stated
without a regression model

— A specific question can be **descriptive**
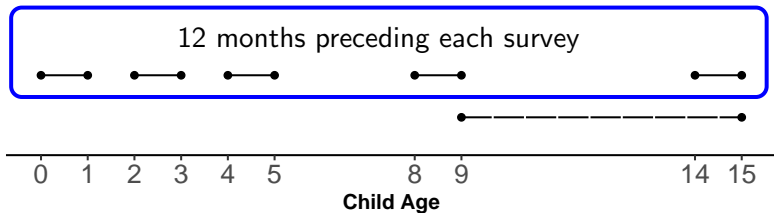
— A specific question can be **causal**

In the past 12 months, were you evicted from your home or apartment for not paying the rent or mortgage?
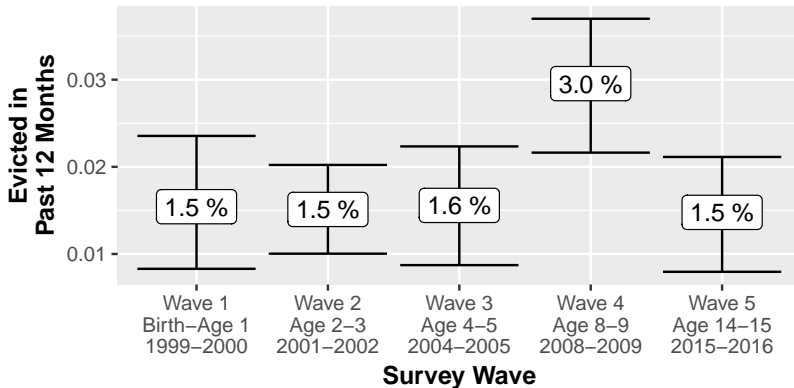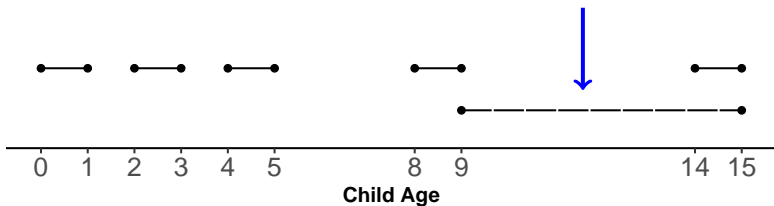
Child ages covered by eviction reports
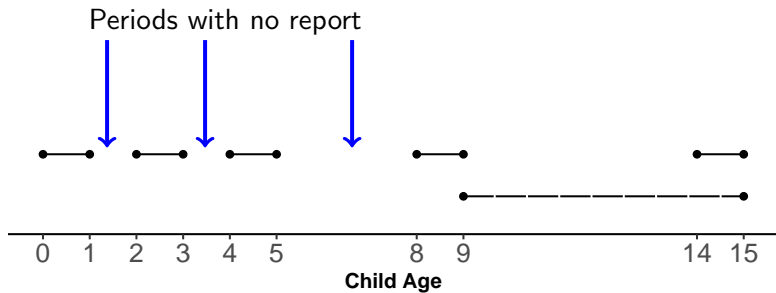
Child ages covered by eviction reports

Child ages covered by eviction reports
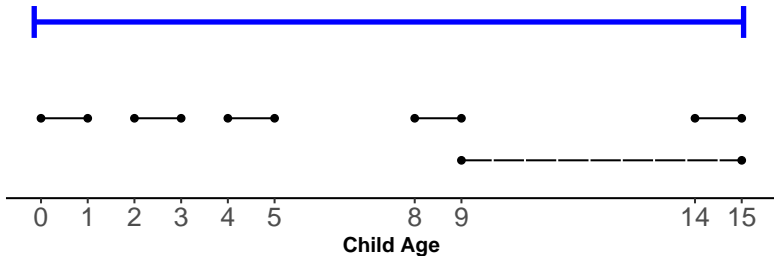
Retrospective:
Any eviction between
ages 9 and 15
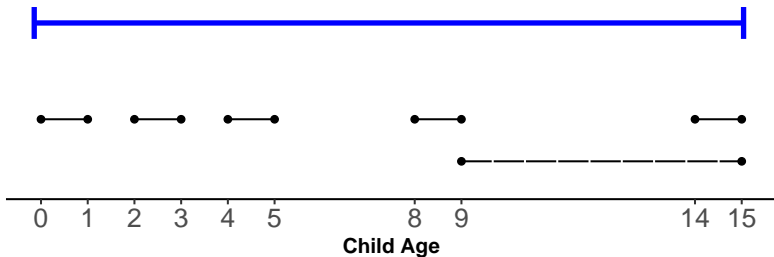
Child Age

Child ages covered by eviction reports

Child ages covered by eviction reports

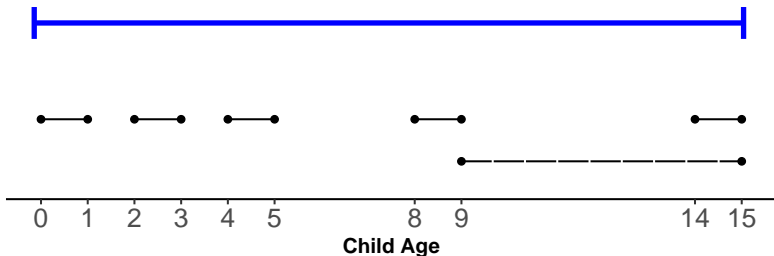Child ages covered by eviction reports

Goal: Proportion evicted at any point

Child Age

Lower bound: **7.9 %** reported an eviction

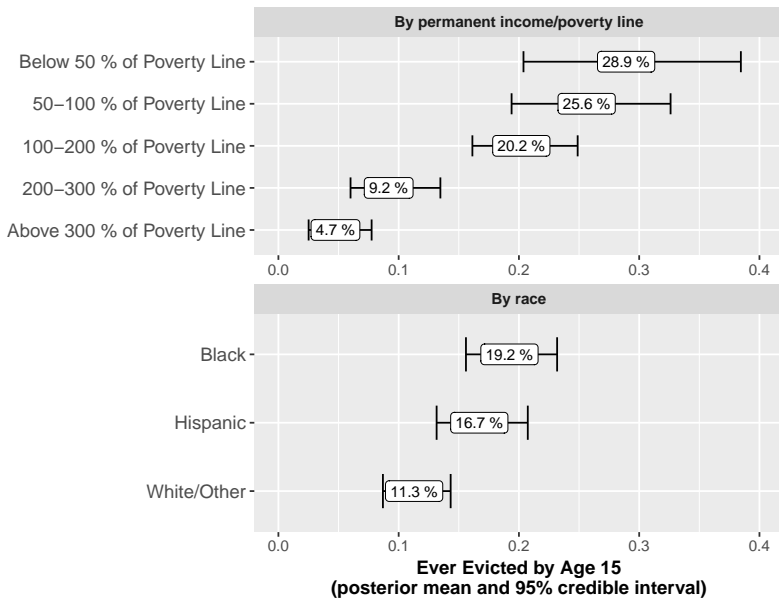Child ages covered by eviction reports

Goal: Proportion evicted at any point

Child Age

Lower bound: **7.9 %** reported an eviction

Model-based: **14.8 %** evicted at some point

Imputes over periods with no report
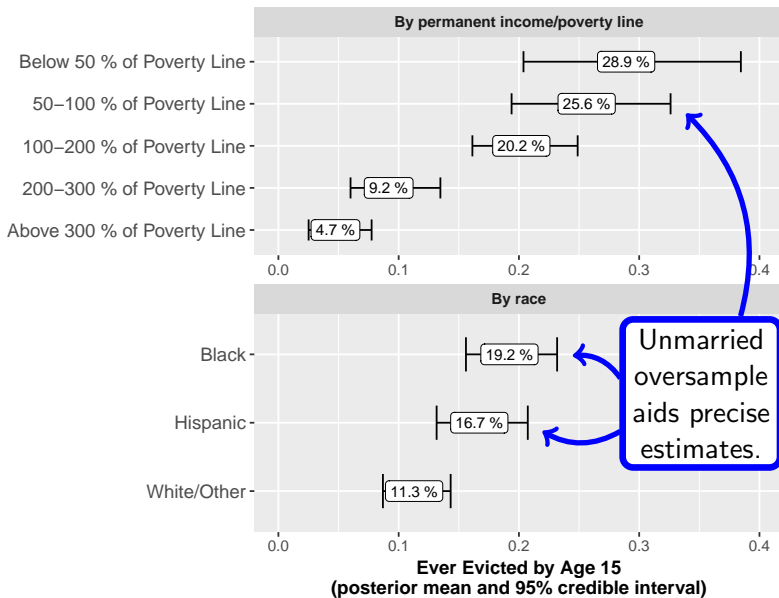
By permanent income/poverty line

Below 50 % of Poverty Line — 28.9 %
50–100 % of Poverty Line — 25.6 %
100–200 % of Poverty Line — 20.2 %
200–300 % of Poverty Line — 9.2 %
Above 300 % of Poverty Line — 4.7 %

By race

Black — 19.2 %
Hispanic — 16.7 %
White/Other — 11.3 %

Ever Evicted by Age 15
(posterior mean and 95% credible interval)

By permanent income/poverty line

Below 50 % of Poverty Line	28.9 %
50–100 % of Poverty Line	25.6 %
100–200 % of Poverty Line	20.2 %
200–300 % of Poverty Line	9.2 %
Above 300 % of Poverty Line	4.7 %

By race

Black	19.2 %
Hispanic	16.7 %
White/Other	11.3 %

Ever Evicted by Age 15
(posterior mean and 95% credible interval)

Unmarried oversample aids precise estimates.

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

There are
so many
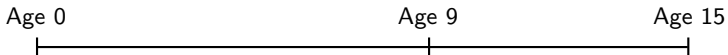variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

⟵

There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

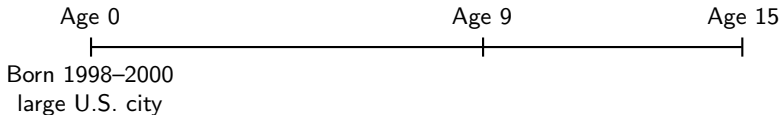— A specific question can be **causal** ⟵

Does **public housing** protect
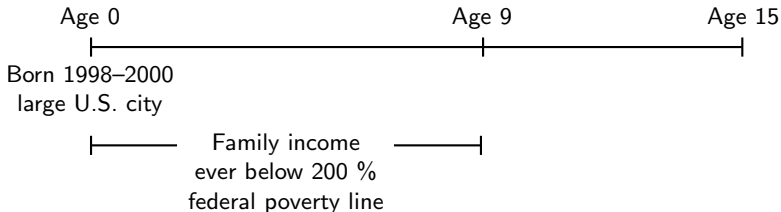low-income families from **eviction**?

Does **public housing** protect
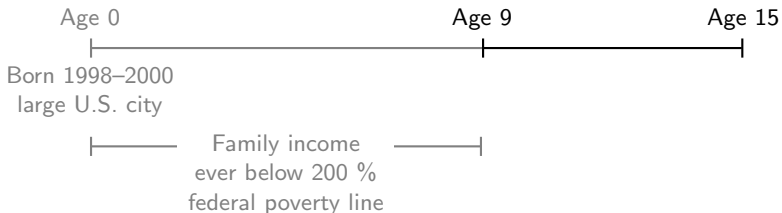low-income families from **eviction**?

Age 0   Age 9   Age 15
|——————————————————————|————————————————————|

Does **public housing** protect low-income families from **eviction**?

Age 0   Age 9   Age 15

Born 1998–2000
large U.S. city

Does **public housing** protect low-income families from **eviction**?

Age 0   Age 9   Age 15

Born 1998–2000
large U.S. city

Family income
ever below 200 %
federal poverty line

Does **public housing** protect low-income families from **eviction**?

Age 0   Age 9   Age 15

Born 1998–2000
large U.S. city

Family income
ever below 200 %
federal poverty line

Does **public housing** protect low-income families from **eviction**?

Age 0

Age 9

Age 15

Born 1998–2000
large U.S. city

Resides in
public housing

Family income
ever below 200 %
federal poverty line

Does **public housing** protect low-income families from **eviction**?

Age 0	Age 9	Age 15
Born 1998–2000 large U.S. city	Resides in **public housing**	
Family income ever below 200 % federal poverty line	Ever **evicted**?	

Effect of **public housing** on **eviction**

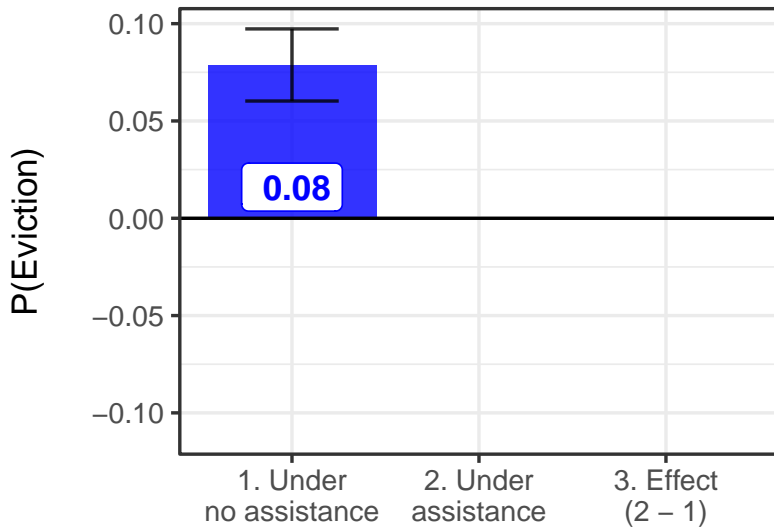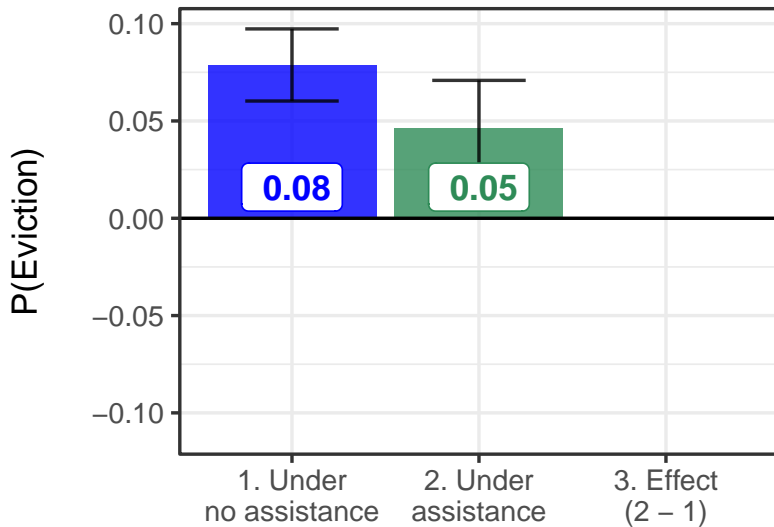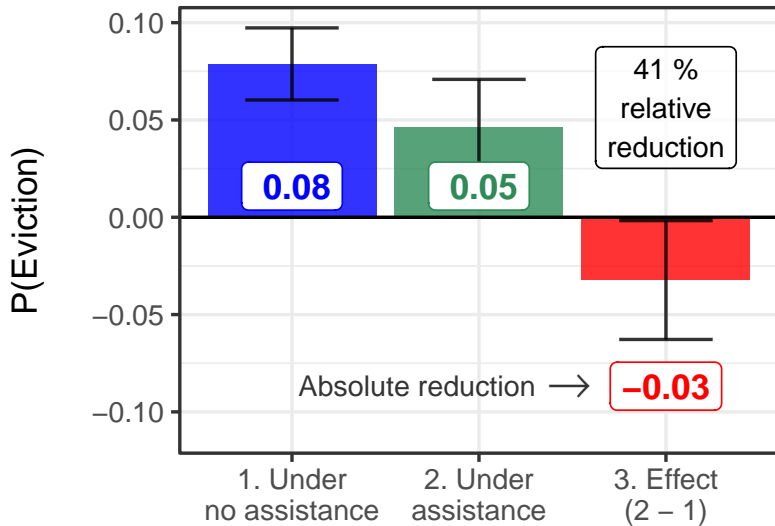# Effect of **public housing** on **eviction**

Effect of **public housing** on **eviction**

Effect of **public housing** on **eviction**

A **temptation**

Study **all causes** of some outcome
by throwing them all in a model.

⟵ There are
so many
variables!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

Cautionary notes

Target population: Describe carefully

— **Not** Nationally representative

— **Not** Residents of cities

— **Yes** Probability sample of births in U.S. cities
with populations over 200,000 in 1998–2000

Cautionary notes

Target population: Describe carefully

- — **Not** Nationally representative
- — **Not** Residents of cities
- — **Yes** Probability sample of births in U.S. cities
  with populations over 200,000 in 1998–2000

Oversample: Describe carefully

- — **Not** Oversample of disadvantaged families
- — **Yes** Oversample (3:1) of nonmarital births
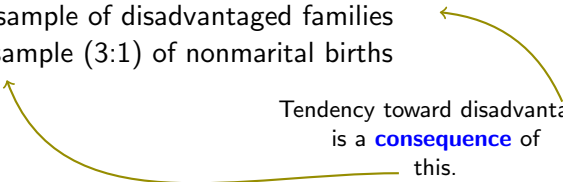
Cautionary notes

Target population: Describe carefully

— **Not** Nationally representative
— **Not** Residents of cities
— **Yes** Probability sample of births in U.S. cities
with populations over 200,000 in 1998–2000

Oversample: Describe carefully

— **Not** Oversample of disadvantaged families
— **Yes** Oversample (3:1) of nonmarital births

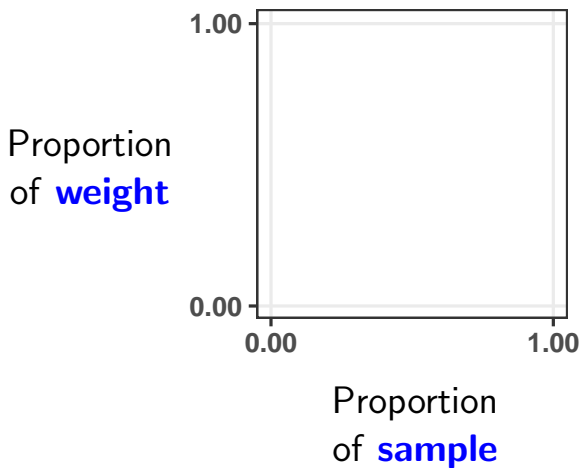Tendency toward disadvantage
is a **consequence** of
this.

Cautionary notes

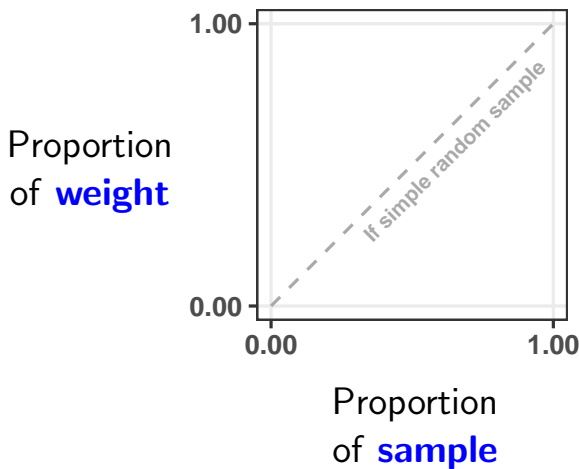**Weights** are highly unequal.

Cautionary notes

Weights are highly unequal.

National mother weights at birth

Proportion
of **weight**

Proportion
of **sample**

Cautionary notes

Weights are highly unequal.

National mother weights at birth

Proportion of **weight**

If simple random sample

Proportion of **sample**

Cautionary notes

Weights are highly unequal.

National mother weights at birth

Proportion of **weight**

If simple random sample

Proportion of **sample**

Cautionary notes

Weights are highly unequal.

National mother weights at birth

Proportion of **weight** (vertical axis), Proportion of **sample** (horizontal axis)

If simple random sample

Cautionary notes

Weights are highly unequal.

National mother weights at birth

Proportion of **weight**

84% of weight rests on **25 %** of sample

Proportion of **sample**

Cautionary notes

National mother weights at birth

Proportion of **weight**

84% of weight rests on 25 % of sample

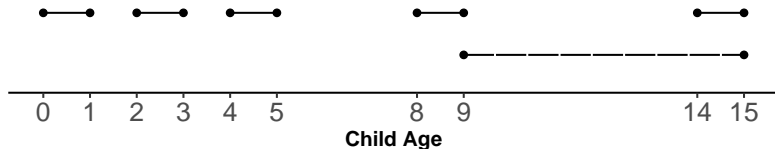Proportion of **sample**
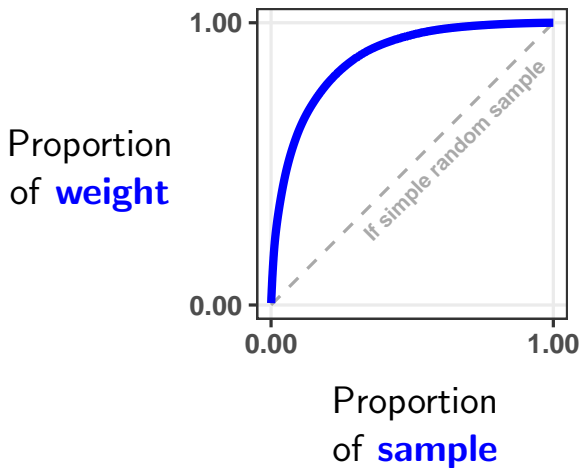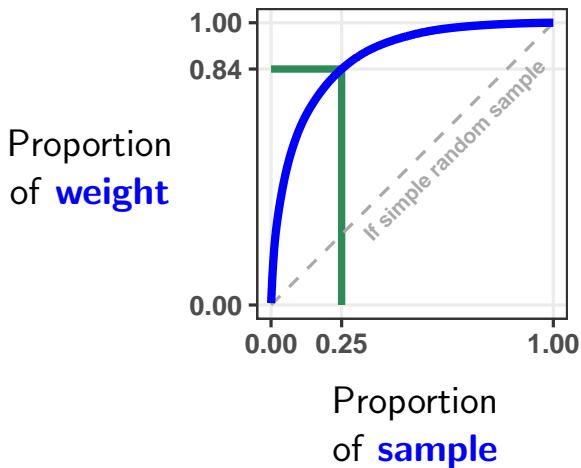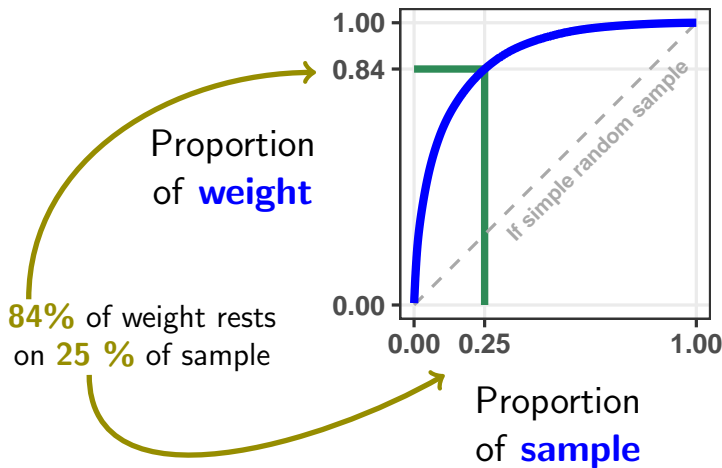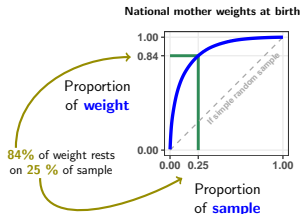
Weights are not bad.

They are just high-variance.

This is a consequence of the oversample
that enables precise estimates for non-marital births.

Advice: Think carefully about whether to weight.
Weighted and unweighted samples are very different.

Comparison to other datasets

Gaps between surveys: Not annual

Target population is specific.   (but so is NLSY-CYA and PSID-CDS)

Many **likert-type** scales influenced by psychology

Opportunities

There are lots of variables no one is studying.
Pick a variable and run with it!

Opportunities

There are lots of variables no one is studying.
Pick a variable and run with it!

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

Opportunities

There are lots of variables no one is studying.
Pick a variable and run with it!

Words of **caution**

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

Opportunities

There are lots of variables no one is studying.
Pick a variable and run with it!

Words of **caution**

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

Target population:
births in cities

Nonmarital oversample

Unequal weights

Opportunities

There are lots of variables no one is studying.
Pick a variable and run with it!

Words of **caution**

A **recommendation**

Aim to answer a **specific question**
that can be stated
without a regression model

— A specific question can be **descriptive**

— A specific question can be **causal**

Target population:
births in cities

Nonmarital oversample

Unequal weights

For the right question,
these are **advantages**

References

Lundberg, Ian, and Louis Donnelly. 2019. "A research note on the prevalence of housing eviction among children born in U.S. cities." *Demography* 56(1):391-404. [accepted manuscript][replication package].

Lundberg, Ian, Sarah L. Gold, Louis Donnelly, Jeanne Brooks-Gunn, and Sara S. McLanahan. "Government assistance protects low-income families from eviction and rent non-payment." To be presented at the 2019 Annual Meeting of the Population Association of America.

Salganik, Matthew J., Ian Lundberg, Alex Kindel, Sara S. McLanahan, and over 100 additional authors. "Measuring the predictability of life outcomes with a scientific mass collaboration." Manuscript in progress. You can read more here or watch a video overview here.