The Nonlinear and Heterogeneous Effects of Parental Income on Children's Educational Attainment

Ian Lundberg
Cornell
ilundberg@cornell.edu

Jennie E. Brand
UCLA
brand@soc.ucla.edu

2 March 2023
Institute for Analytical Sociology
Linkoping University

Research reported in this presentation was supported by the National Science Foundation under Award Number 2104607 and by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD041022.

How computing looked **in the 1950s**

Source: NASA

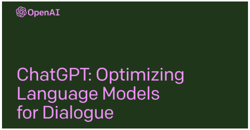
How computing looked **in the 1980s**

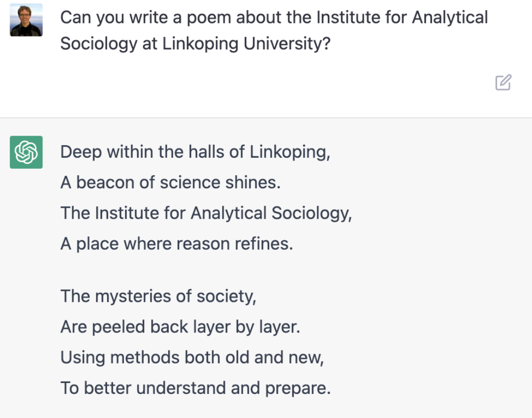
Source: Wikimedia

How computing looks **today**

Source: Apple

How computing looks **today**





Source: OpenAI

Computing has advanced rapidly since the 1960s

Computing has advanced rapidly since the 1960s

How has quantitative social science changed?

How stratification research looked **in the 1960s**

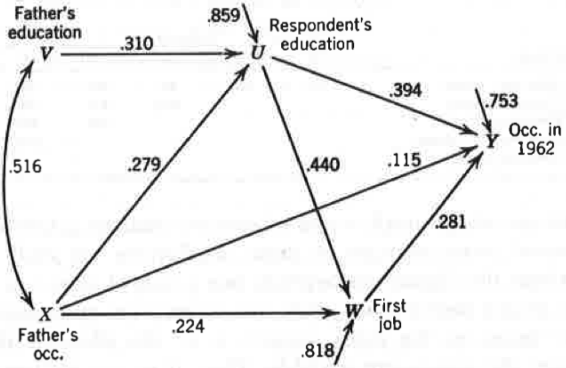


Figure 5.1. Path coefficients in basic model of the process of stratification.

Source: Blau & Duncan 1967

How stratification research looks **today**

	Logistic Regression	
	Enrollment in College by Age 20	
	(1)	(2)
log(Family Income in Adolescence)	0.961***	0.451***
	(0.045)	(0.056)
Race (white / other omitted)		
Hispanic		0.010
		(0.105)
Non-Hispanic Black		0.205*
		(0.099)
Parents' education		
One parent finished college		0.944***
		(0.087)
Both parents finished college		1.374***
		(0.129)
log(Family Wealth in Adolescence)		0.211***
		(0.023)
Constant	−11.049***	−8.028***
	(0.501)	(0.549)
Observations	4,777	4,777
Log Likelihood	−2,542.000	−2,391.006
Akaike Inf. Crit.	5,088.000	4,800.012
Data: NLSY97	*p<0.05; **p<0.01; ***p<0.001	

How stratification research looks **today**

	Logistic Regression	
	Enrollment in College by Age 20	
	(1)	(2)
log(Family Income in Adolescence)	0.961^{***}	0.451^{***}
	(0.045)	(0.056)
Race (white / other omitted)		
Hispanic		0.010
		(0.105)
Non-Hispanic Black		0.205^{*}
		(0.099)
Parents' education		
One parent finished college		0.944^{***}
		(0.087)
Both parents finished college		1.374^{***}
		(0.129)
log(Family Wealth in Adolescence)		0.211^{***}
		(0.023)
Constant	-11.049^{***}	-8.028^{***}
	(0.501)	(0.549)
Observations	4,777	4,777
Log Likelihood	−2,542.000	−2,391.006
Akaike Inf. Crit.	5,088.000	4,800.012
Data: NLSY97	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$	

Why have regressions stuck around?

Why have regressions stuck around?

- We've learned a ton that way

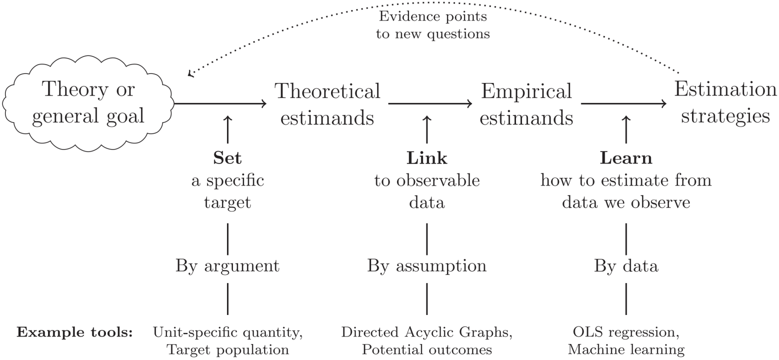
Why have regressions stuck around?

- We've learned a ton that way
- We can interpret regressions

Why have regressions stuck around?

- We've learned a ton that way
- We can interpret regressions
- We have small sample sizes

How might we move on from regression?



Lundberg, Johnson, & Stewart. 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. American Sociological Review 86(3): 532-565.

Road map for today

Lundberg & Brand. The Nonlinear and Heterogeneous Effects of Parental Income on Children's Educational Attainment

- Data
- Theory and hypotheses
- Causal inference
- Estimation + Results
- Discussion

Data

National Longitudinal Survey of Youth, 1997 Cohort

- Ages 12–17 in 1997
- Treatment: Family income in 1997
- Outcome: College enrollment by age 20
- Confounders
  - Family wealth
  - Parents' education
    - Neither finished college
    - One finished college
    - Both finished college
  - Race
    - Hispanic
    - Non-Hispanic Black
    - Non-Hispanic White / Other
- Raw sample $n = 8{,}984$. Analytical sample $n = 4{,}777$

Theory: We believe in a heterogeneous world

Theory: We believe in a heterogeneous world

For whom is college enrollment most responsive to income?

Theory: We believe in a heterogeneous world

For whom is college enrollment most responsive to income?

H1: Those with low income or wealth

- Theory: Financial constraints

Theory: We believe in a heterogeneous world

For whom is college enrollment most responsive to income?

H1: Those with low income or wealth

- Theory: Financial constraints

H2a: Those whose parents do not hold BAs

- Theory: Financial capital overcomes limited cultural capital

Theory: We believe in a heterogeneous world

For whom is college enrollment most responsive to income?

H1: Those with low income or wealth

- Theory: Financial constraints

H2a: Those whose parents do not hold BAs

- Theory: Financial capital overcomes limited cultural capital

H2b: Those whose parents hold BAs

- Theory: These parents know how to effectively convert financial capital into higher education opportunities

These theories invoke **causal effects**
that vary over subpopulations

Causal inference: A missing data problem

	Factual treatment	Outcome under treatment value	
		Placebo	Drug
Person 1	Placebo	•	◦
Person 2	Drug	◦	•
Person 3	Drug	◦	•
Person 4	Placebo	•	◦

Causal inference: A missing data problem

	Factual treatment	Outcome under treatment value					
		\$10k	\$20k	\$30k	\$40k	\$50k	...
Person 1	\$30k	○	○	●	○	○	○
Person 2	\$20k	○	●	○	○	○	○
Person 3	\$50k	○	○	○	○	●	○
Person 4	\$40k	○	○	○	●	○	○

Causal inference: A missing data problem

	Factual treatment	Outcome under treatment value
		$10k $20k $30k $40k $50k ...
Person 1	$30k	○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Person 2	$20k	○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Person 3	$50k	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○
Person 4	$40k	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○

Causal inference: A missing data problem

	Factual treatment	Outcome under treatment value
		$10k $20k $30k $40k $50k . . .
Person 1	$30k	○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Person 2	$20k	○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Person 3	$50k	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○
Person 4	$40k	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○

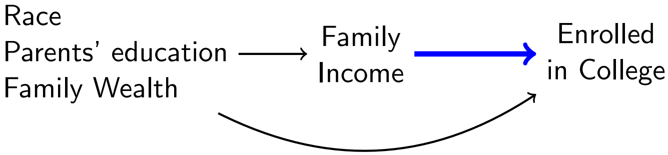
Consistency

$Y = Y^A$

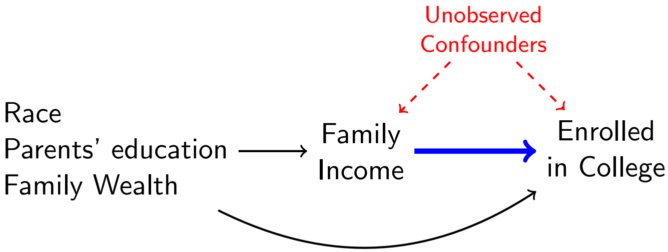
See Vanderweele 2009

Causal assumptions

Exchangeability
Greenland, Pearl,
& Robins 1999



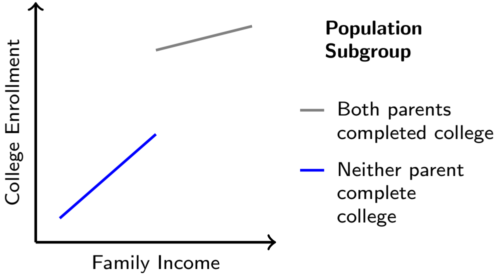
Causal assumptions



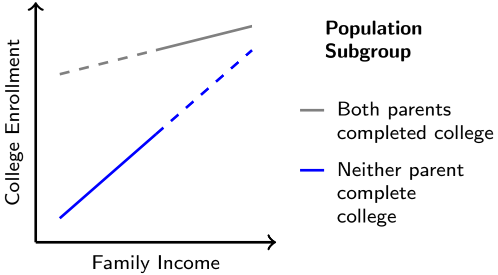
Exchangeability
Greenland, Pearl,
& Robins 1999

The **extrapolation problem**:
Within a population subgroup,
many treatments are unobserved

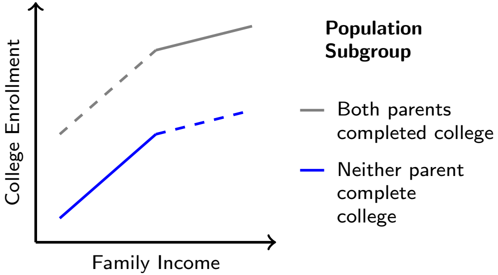
The **extrapolation problem**: Within a population subgroup, many treatments are unobserved



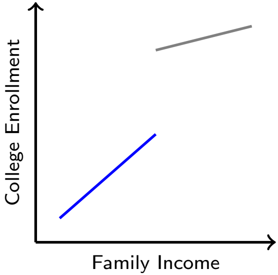
The **extrapolation problem**: Within a population subgroup, many treatments are unobserved



The **extrapolation problem**: Within a population subgroup, many treatments are unobserved



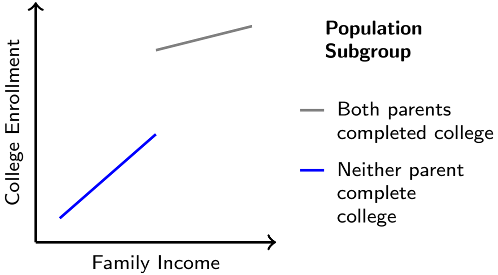
The **extrapolation problem**: Within a population subgroup, many treatments are unobserved



Population Subgroup

- Both parents completed college
- Neither parent complete college

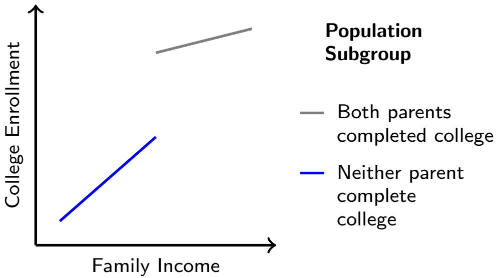
The **extrapolation problem**:
Within a population subgroup,
many treatments are unobserved

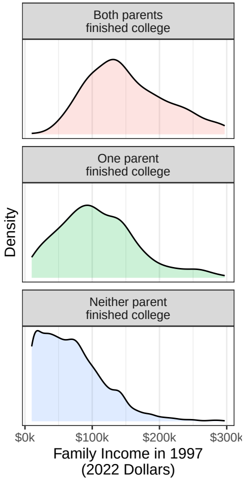


1) Only visualize the middle 90% of observed treatments

2) Only visualize if $n \geq 25$

The **extrapolation problem**:
Within a population subgroup,
many treatments are unobserved





1) Only visualize the middle 90% of observed treatments

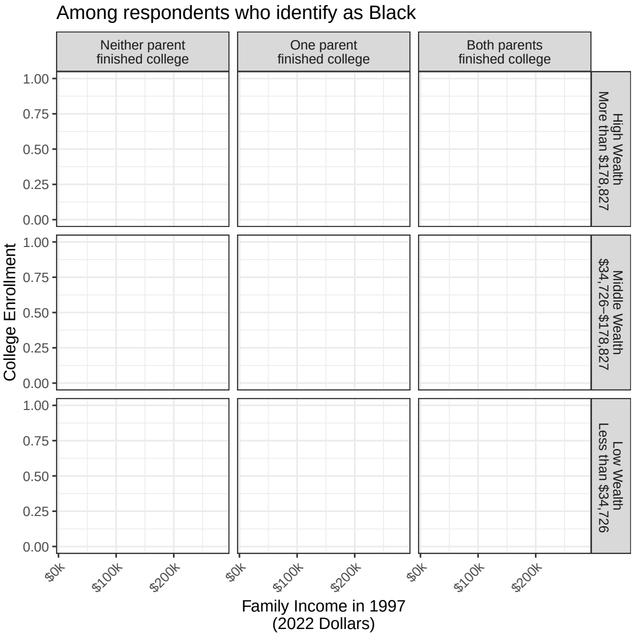
2) Only visualize if $n \geq 25$

Road map for today

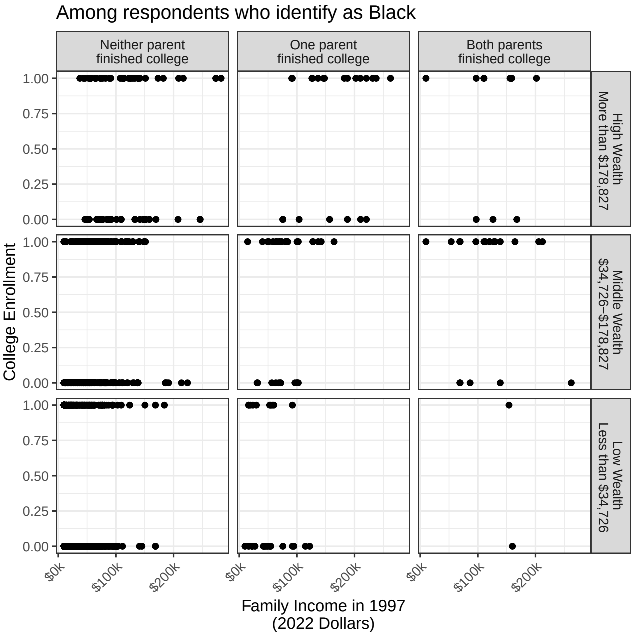
Lundberg & Brand. The Nonlinear and Heterogeneous Effects of Parental Income on Children's Educational Attainment

- Data
- Theory and hypotheses
- Causal inference
- Estimation + Results
- Discussion

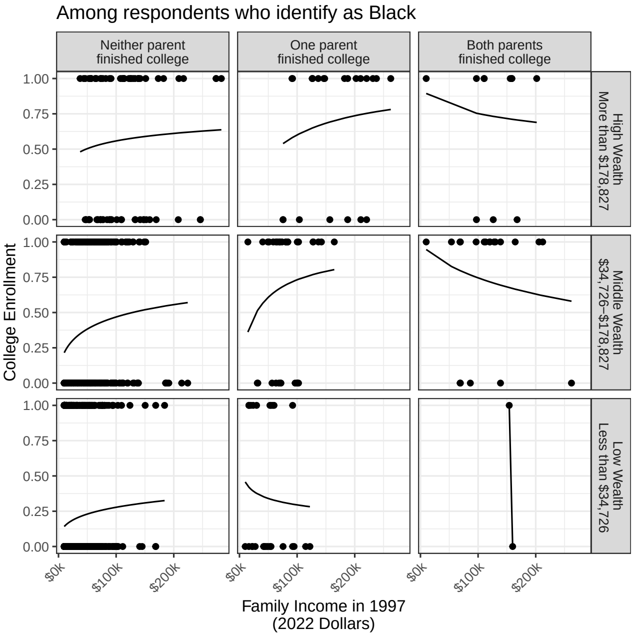
How might we move on from regression?



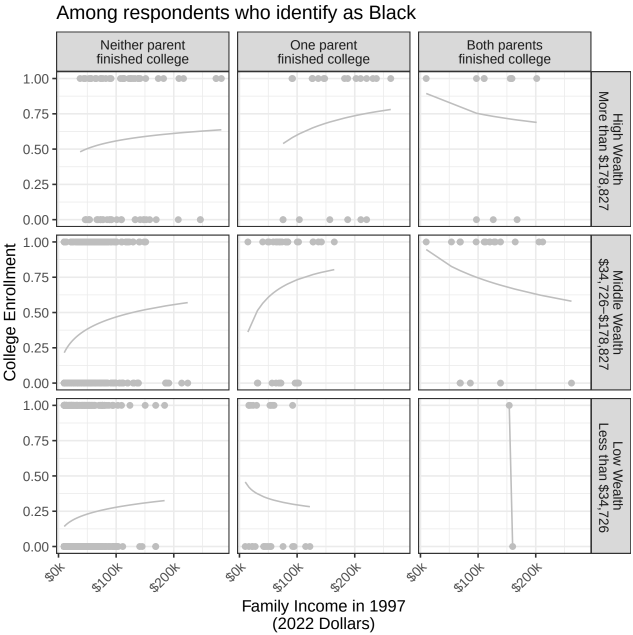
Among respondents who identify as Black
Neither parent finished college
One parent finished college
Both parents finished college
High Wealth More than $178,827
Middle Wealth $34,726–$178,827
Low Wealth Less than $34,726
College Enrollment
1.00
0.75
0.50
0.25
0.00
$0k
$100k
$200k
Family Income in 1997 (2022 Dollars)

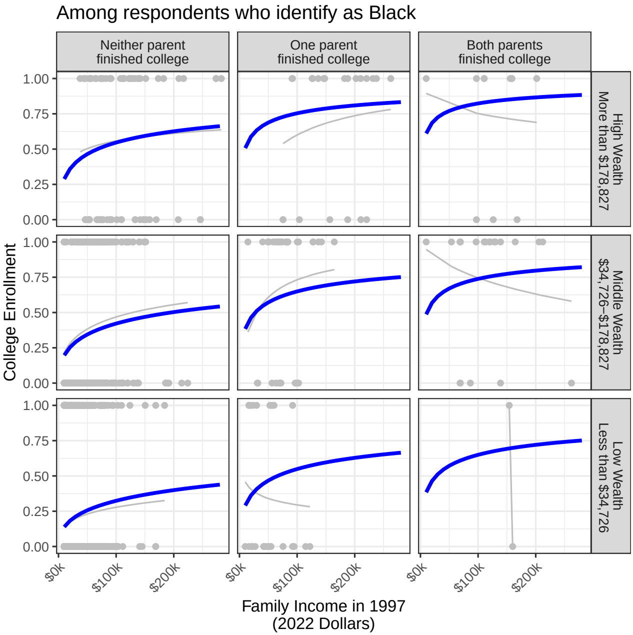


Among respondents who identify as Black
Neither parent finished college
One parent finished college
Both parents finished college
High Wealth More than $178,827
Middle Wealth $34,726–$178,827
Low Wealth Less than $34,726
College Enrollment
1.00
0.75
0.50
0.25
0.00
$0k
$100k
$200k
Family Income in 1997 (2022 Dollars)

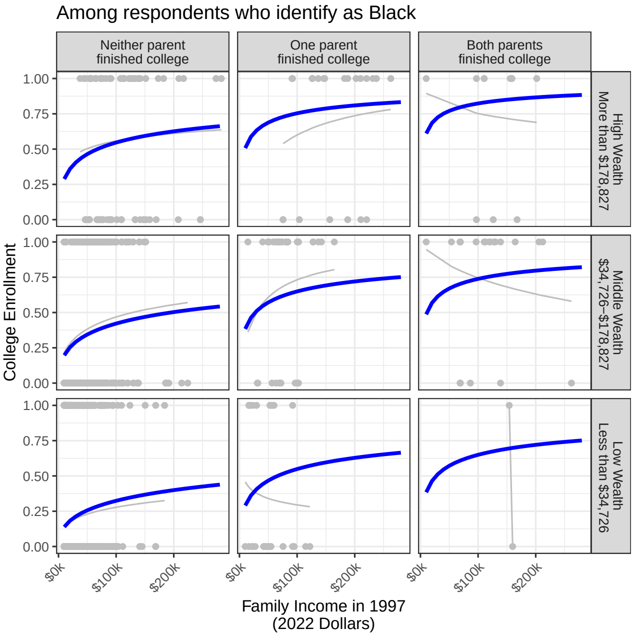


Among respondents who identify as Black
Neither parent finished college
One parent finished college
Both parents finished college
High Wealth More than $178,827
Middle Wealth $34,726–$178,827
Low Wealth Less than $34,726
College Enrollment
1.00
0.75
0.50
0.25
0.00
$0k
$100k
$200k
Family Income in 1997 (2022 Dollars)


Among respondents who identify as Black
Neither parent finished college
One parent finished college
Both parents finished college
High Wealth More than $178,827
Middle Wealth $34,726–$178,827
Low Wealth Less than $34,726
College Enrollment
1.00
0.75
0.50
0.25
0.00
$0k
$100k
$200k
Family Income in 1997 (2022 Dollars)



Among respondents who identify as Black
Neither parent finished college
One parent finished college
Both parents finished college
High Wealth More than $178,827
Middle Wealth $34,726–$178,827
Low Wealth Less than $34,726
College Enrollment
1.00
0.75
0.50
0.25
0.00
$0k
$100k
$200k
Family Income in 1997 (2022 Dollars)

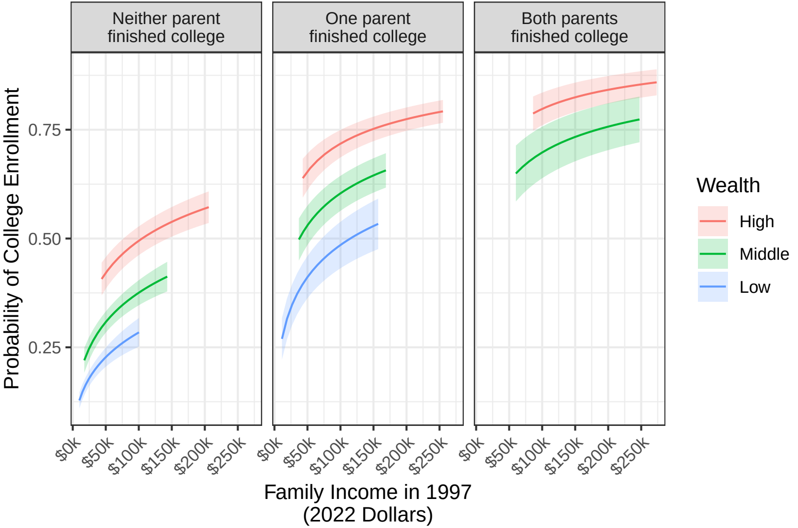


Want a balance between **structure** and **flexibility**

Structure: Additive logistic regression

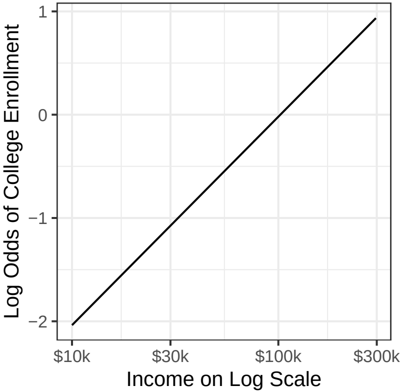
$$
\begin{aligned}
\mathrm{logit}\left(\frac{\mathsf{P}(Y=1 \mid \vec{X})}{1-\mathsf{P}(Y=1 \mid \vec{X})}\right) = \alpha &+ \beta \times \log(\text{Income}) \\
&+ \vec{\gamma} \times (\text{Parents' Education}) \\
&+ \vec{\eta} \times (\text{Race}) \\
&+ \vec{\lambda} \times \log(\text{Wealth}) \times \text{WealthTercile}
\end{aligned}
$$

Structure: Additive logistic regression



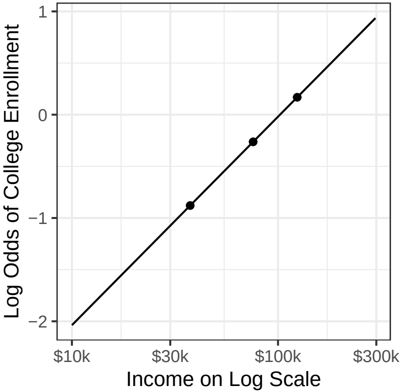
Flexibility: Generalized additive model + interactions

Wood 2017, R package `mgcv`



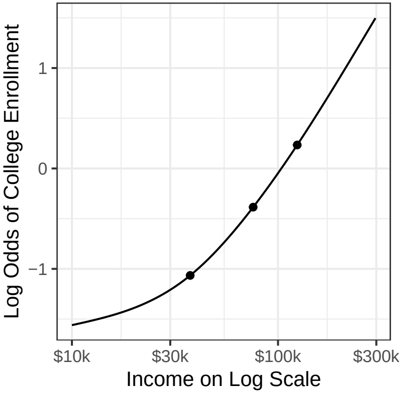
Flexibility: Generalized additive model + interactions

Wood 2017, R package `mgcv`



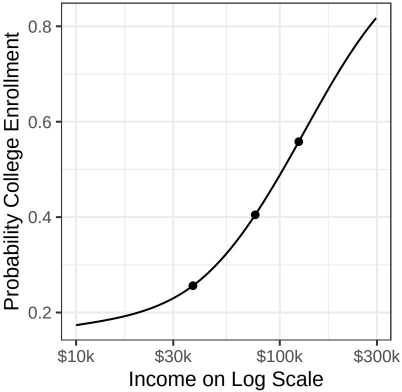
Flexibility: Generalized additive model + interactions

Wood 2017, R package `mgcv`



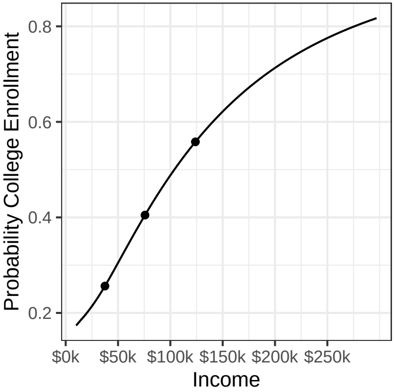
Flexibility: Generalized additive model + interactions

Wood 2017, R package `mgcv`

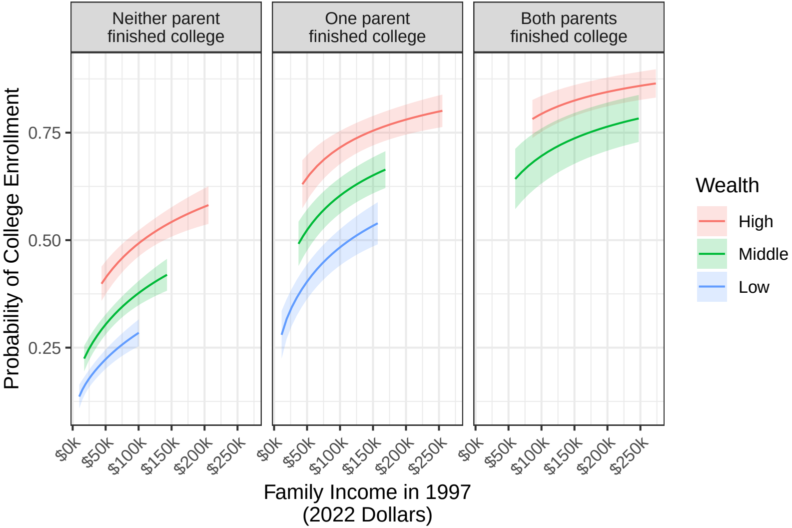


Flexibility: Generalized additive model + interactions

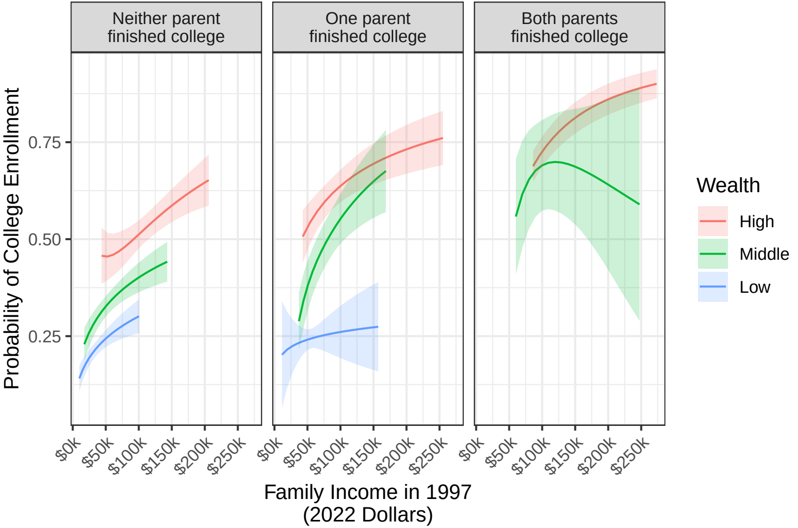
Wood 2017, R package `mgcv`



Linear model



Smooth model with interactions

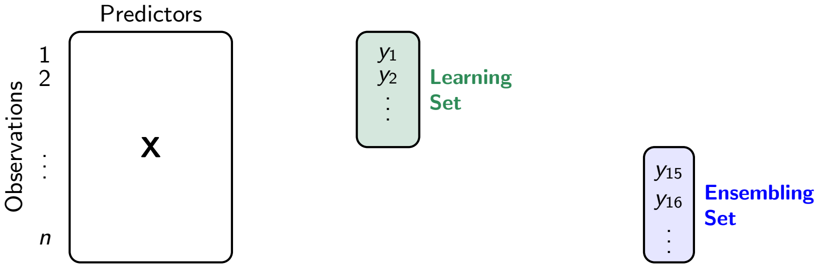


Many base learners

- Standard logistic regression
  - Additive
  - Income $\times$ Education
  - Income $\times$ Wealth
  - Income $\times$ Race
  - Income $\times$ Education $\times$ Wealth
  - Income $\times$ Education $\times$ Race
- Each of the above, but with smooths

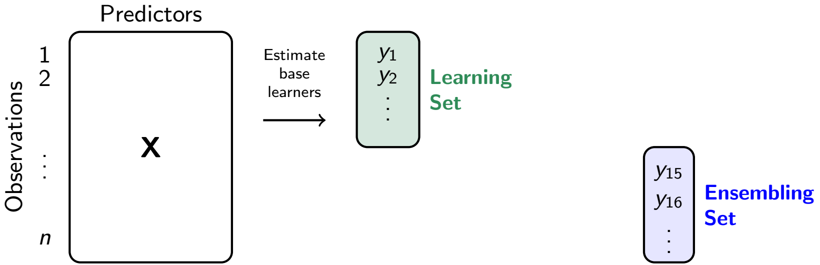
Ensemble

See **Super Learner**
Van der Laan et al. 2007



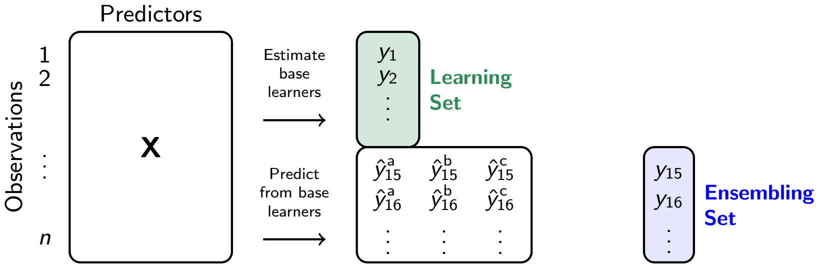
Ensemble

See **Super Learner**
Van der Laan et al. 2007



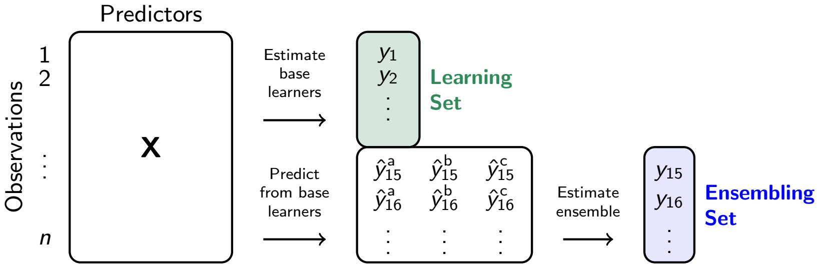
Ensemble

See **Super Learner**
Van der Laan et al. 2007



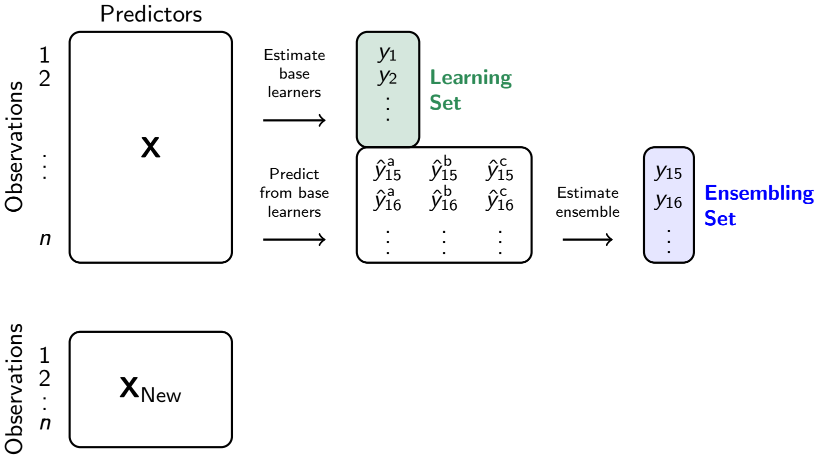
Ensemble

See **Super Learner**
Van der Laan et al. 2007



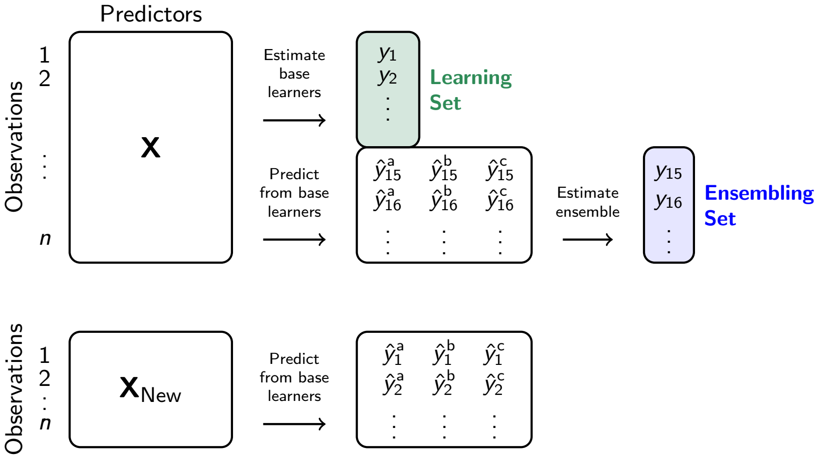
Ensemble

See **Super Learner**
Van der Laan et al. 2007



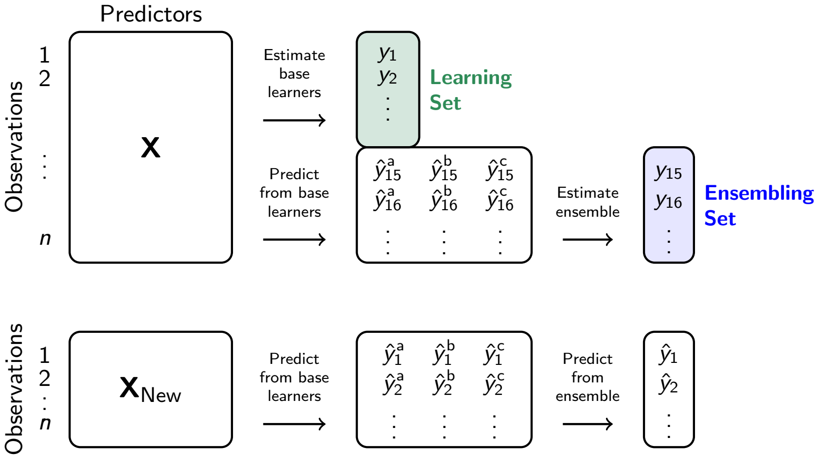
Ensemble

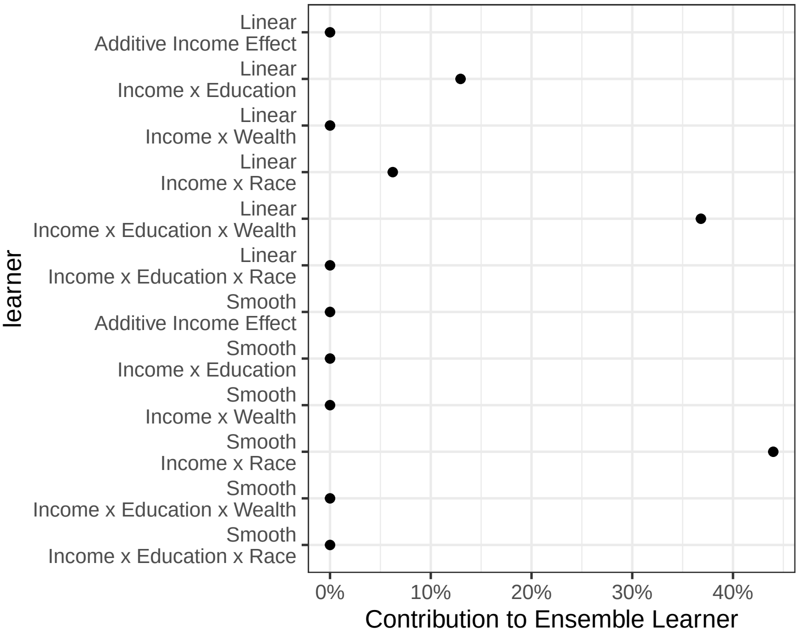
See **Super Learner**
Van der Laan et al. 2007

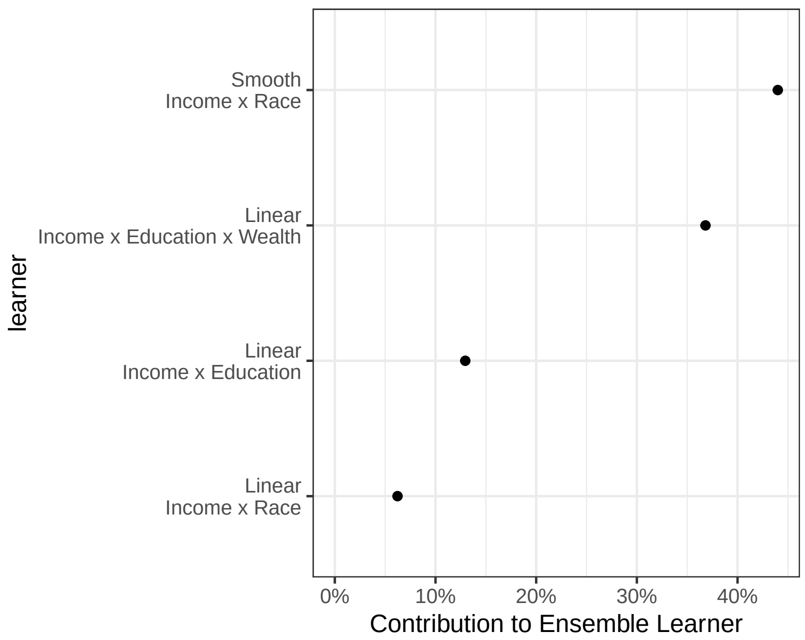


Ensemble

See **Super Learner**
Van der Laan et al. 2007

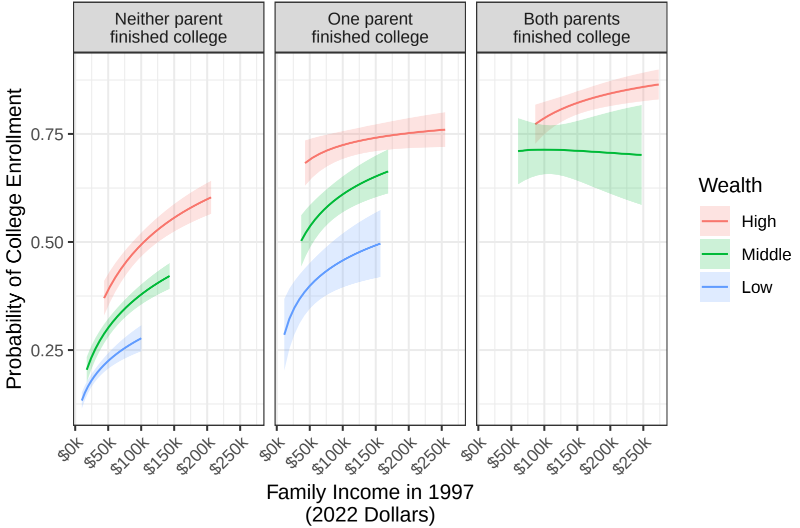



Linear
Additive Income Effect
Linear
Income x Education
Linear
Income x Wealth
Linear
Income x Race
Linear
Income x Education x Wealth
Linear
Income x Education x Race
Smooth
Additive Income Effect
Smooth
Income x Education
Smooth
Income x Wealth
Smooth
Income x Race
Smooth
Income x Education x Wealth
Smooth
Income x Education x Race
learner
0%
10%
20%
30%
40%
Contribution to Ensemble Learner


Smooth
Income x Race
Linear
Income x Education x Wealth
Linear
Income x Education
Linear
Income x Race
learner
0%
10%
20%
30%
40%
Contribution to Ensemble Learner

Ensemble results

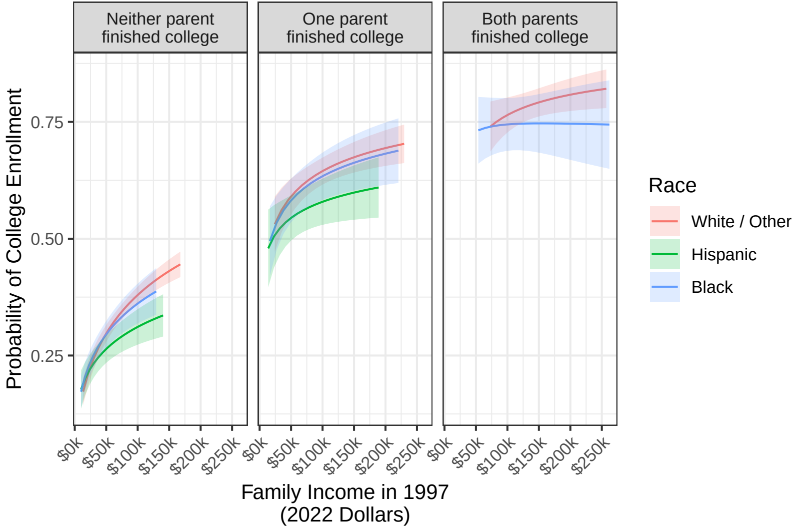
Estimate omitted if $n < 25$



;

Ensemble results

Estimate omitted if $n < 25$



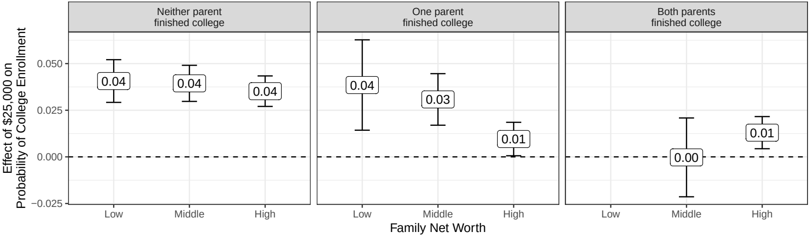
;

Aggregate point estimates

1. Predict at observed values
2. Predict with $25,000 extra
3. Difference and average

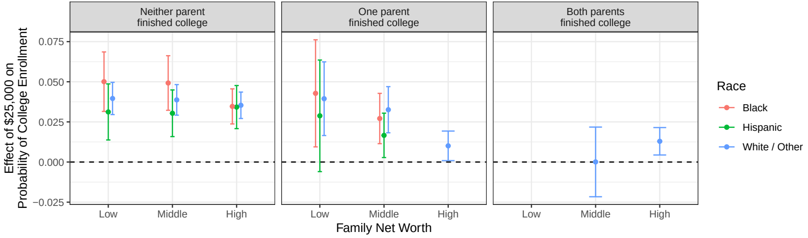
Ensemble results

Estimate omitted if $n < 25$



Ensemble results

Estimate omitted if $n < 25$



DISCUSSION

A story of cultural capital and financial capital

A story of cultural capital and financial capital

For those with vast cultural capital,	(parents completed college)
financial capital is inconsequential	(enroll in college regardless)

A story of cultural capital and financial capital

For those with vast cultural capital, financial capital is inconsequential	(parents completed college) (enroll in college regardless)
For those with less cultural capital, financial capital helps a bit	(parents did not complete college) \$25,000 $\rightarrow +4$ percentage points

Small effects: Policy versus scientific understanding

Small effects: Policy versus scientific understanding

Imagine a policymaker

Small effects: Policy versus scientific understanding

Imagine a policymaker

- Take \$625,000

Small effects: Policy versus scientific understanding

Imagine a policymaker

- Take \$625,000
- Distribute it across 25 families

Small effects: Policy versus scientific understanding

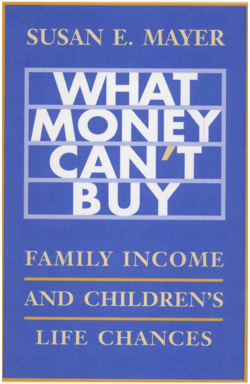
Imagine a policymaker

- Take $625,000
- Distribute it across 25 families
- Cause 1 to enroll in college

Small effects: Policy versus scientific understanding

Imagine a policymaker

- Take \$625,000
- Distribute it across 25 families
- Cause 1 to enroll in college

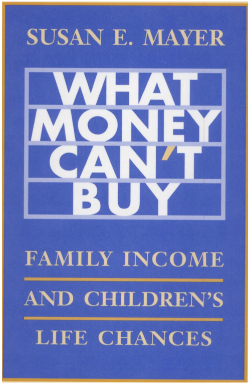


Small effects: Policy versus scientific understanding

Imagine a policymaker

- Take \$625,000
- Distribute it across 25 families
- Cause 1 to enroll in college

Small effects are still worthwhile



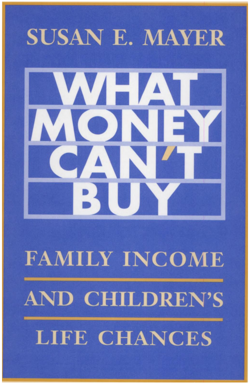
Small effects: Policy versus scientific understanding

Imagine a policymaker

- Take \$625,000
- Distribute it across 25 families
- Cause 1 to enroll in college

Small effects are still worthwhile

- We care for understanding



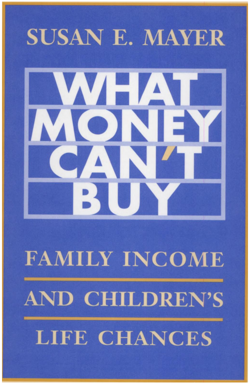
Small effects: Policy versus scientific understanding

Imagine a policymaker

- Take $625,000
- Distribute it across 25 families
- Cause 1 to enroll in college

Small effects are still worthwhile

- We care for understanding
- Many collateral benefits of income



In a world where effects are small,

In a world where effects are small,
it is all the more important to find
the subgroups for whom they are larger

In a world where effects are small,
it is all the more important to find
the subgroups for whom they are larger

Do this with trees

Athey & Imbens 2016
Brand et al. 2021

In a world where effects are small,
it is all the more important to find
the subgroups for whom they are larger

Do this with trees	Athey & Imbens 2016
	Brand et al. 2021
Do this with forests	Wager & Athey 2018

In a world where effects are small,
it is all the more important to find
the subgroups for whom they are larger

Do this with trees	Athey & Imbens 2016
	Brand et al. 2021
Do this with forests	Wager & Athey 2018
Do this with targeted learning	Van der Laan & Rose 2018

Scholars of social stratification face a tension

$$\begin{array}{rcl} \text{theory} & \rightsquigarrow & \text{belief in effect heterogeneity} \\ \text{small samples} & \rightsquigarrow & \text{estimation of additive models} \end{array}$$

Scholars of social stratification face a tension

$$\begin{array}{rcl} \text{theory} & \rightsquigarrow & \text{belief in effect heterogeneity} \\ \text{small samples} & \rightsquigarrow & \text{estimation of additive models} \end{array}$$

Middle Ground

Scholars of social stratification face a tension

theory	$\rightsquigarrow$	belief in effect heterogeneity
small samples	$\rightsquigarrow$	estimation of additive models

Middle Ground

search for heterogeneity
(flexible model)

Scholars of social stratification face a tension

$$\begin{array}{rcl} \text{theory} & \rightsquigarrow & \text{belief in effect heterogeneity} \\ \text{small samples} & \rightsquigarrow & \text{estimation of additive models} \end{array}$$

Middle Ground

search for heterogeneity (flexible model)	assume structure (additive model)

Scholars of social stratification face a tension

theory	$\rightsquigarrow$	belief in effect heterogeneity
small samples	$\rightsquigarrow$	estimation of additive models

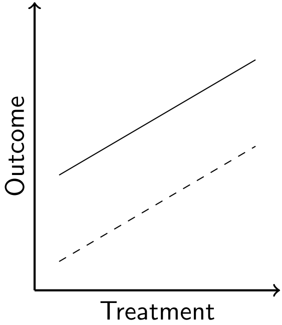
Middle Ground

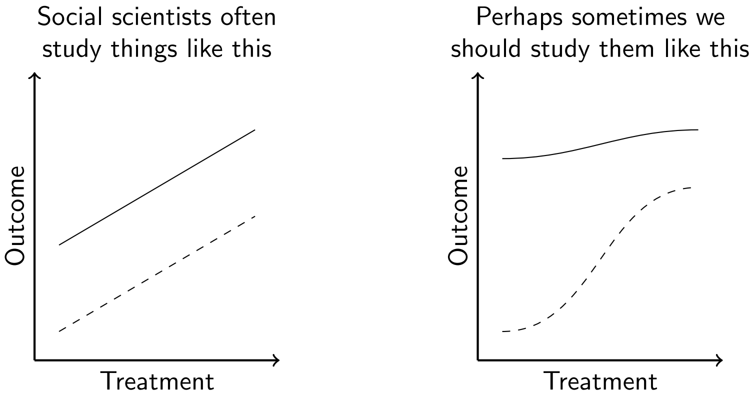
search for heterogeneity (flexible model) $\searrow$

$\swarrow$ assume structure (additive model)

data-driven weighted average

Social scientists often study things like this



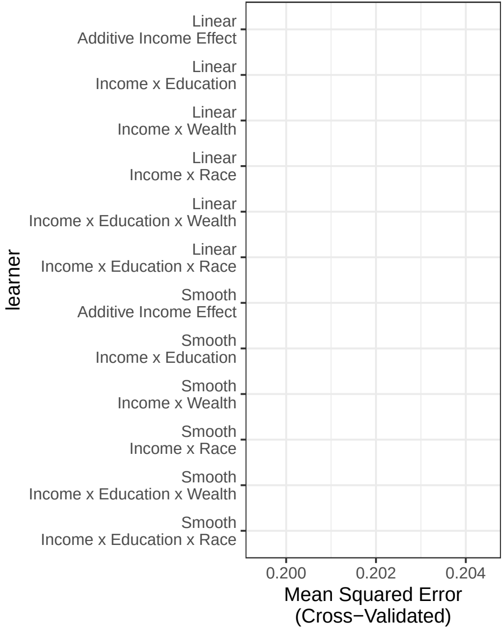

Social scientists often study things like this
Outcome
Treatment
Perhaps sometimes we should study them like this
Outcome
Treatment

Thanks!

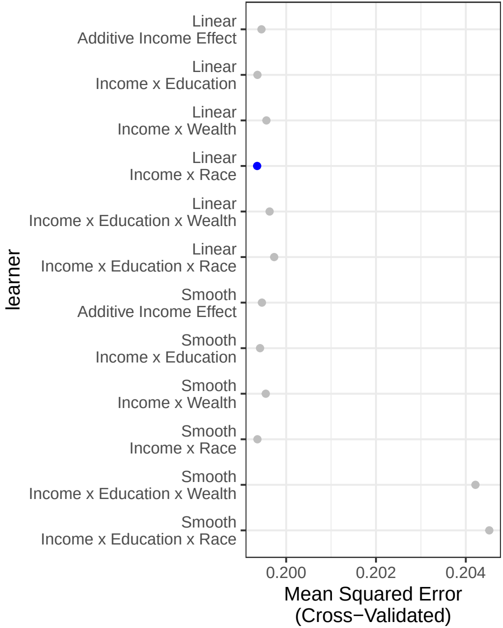
Ian Lundberg
ilundberg@cornell.edu

Jennie E. Brand
brand@soc.ucla.edu

How did the base learners perform?

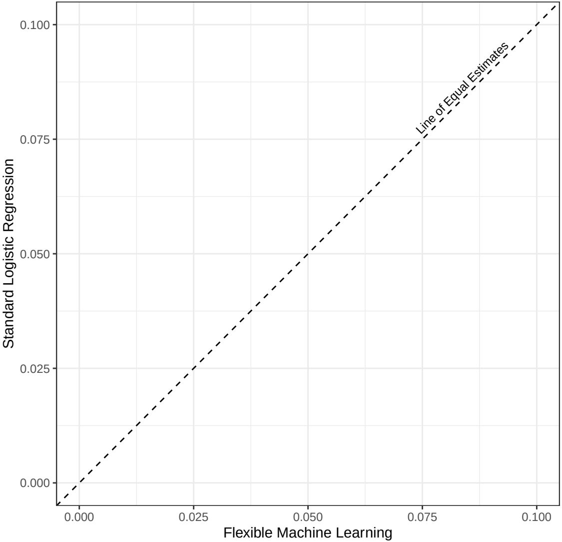


How did the base learners perform?



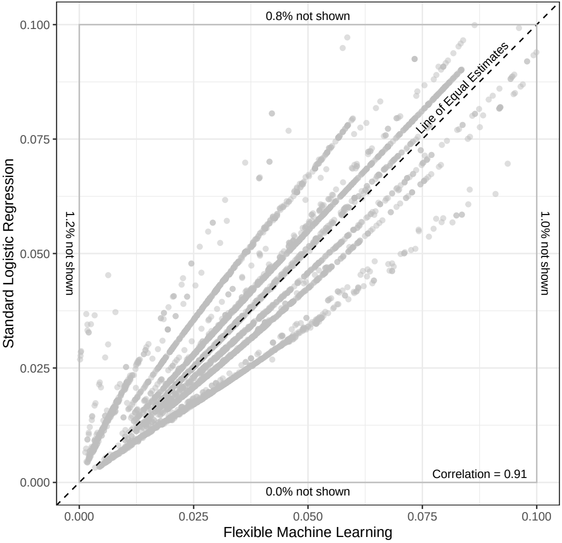
How different was machine learning?

Effect of extra $25k



How different was machine learning?

Effect of extra $25k



Sample restrictions

Raw sample	8,984
Observed at age 25+	8,408
With valid income	6,198
With income not top-coded	6,074
Non-missing wealth	5,418
Valid enrollment outcome	4,856
Valid completion outcome	4,777

Standard errors

Let $\hat{\theta}_1, \hat{\theta}_2$ be estimates that are random variables due to sampling variability. Let π_1, π_2 be weights on them. For simplicity, we will take the ensemble weights π_1, π_2 to be known, although this misses one source of uncertainty.

$$\begin{aligned}
\mathsf{V}(\pi_1\hat{\theta}_1 + \pi_2\hat{\theta}_2) &= \pi_1^2\mathsf{V}(\hat{\theta}_1) + \pi_2^2\mathsf{V}(\hat{\theta}_2) + 2\pi_1\pi_2\mathsf{Cov}(\hat{\theta}_1, \hat{\theta}_2) \\
&= \pi_1^2\mathsf{V}(\hat{\theta}_1) + \pi_2^2\mathsf{V}(\hat{\theta}_2) + 2\pi_1\pi_2\mathsf{Cor}(\hat{\theta}_1, \hat{\theta}_2)\mathsf{SD}(\hat{\theta}_1)\mathsf{SD}(\hat{\theta}_2) \\
&\leq \pi_1^2\mathsf{V}(\hat{\theta}_1) + \pi_2^2\mathsf{V}(\hat{\theta}_2) + 2\pi_1\pi_2\mathsf{SD}(\hat{\theta}_1)\mathsf{SD}(\hat{\theta}_2)
\end{aligned}$$

First difference, by education and race