EC 1152 - Using Big Data to Solve Economic and Social Problems

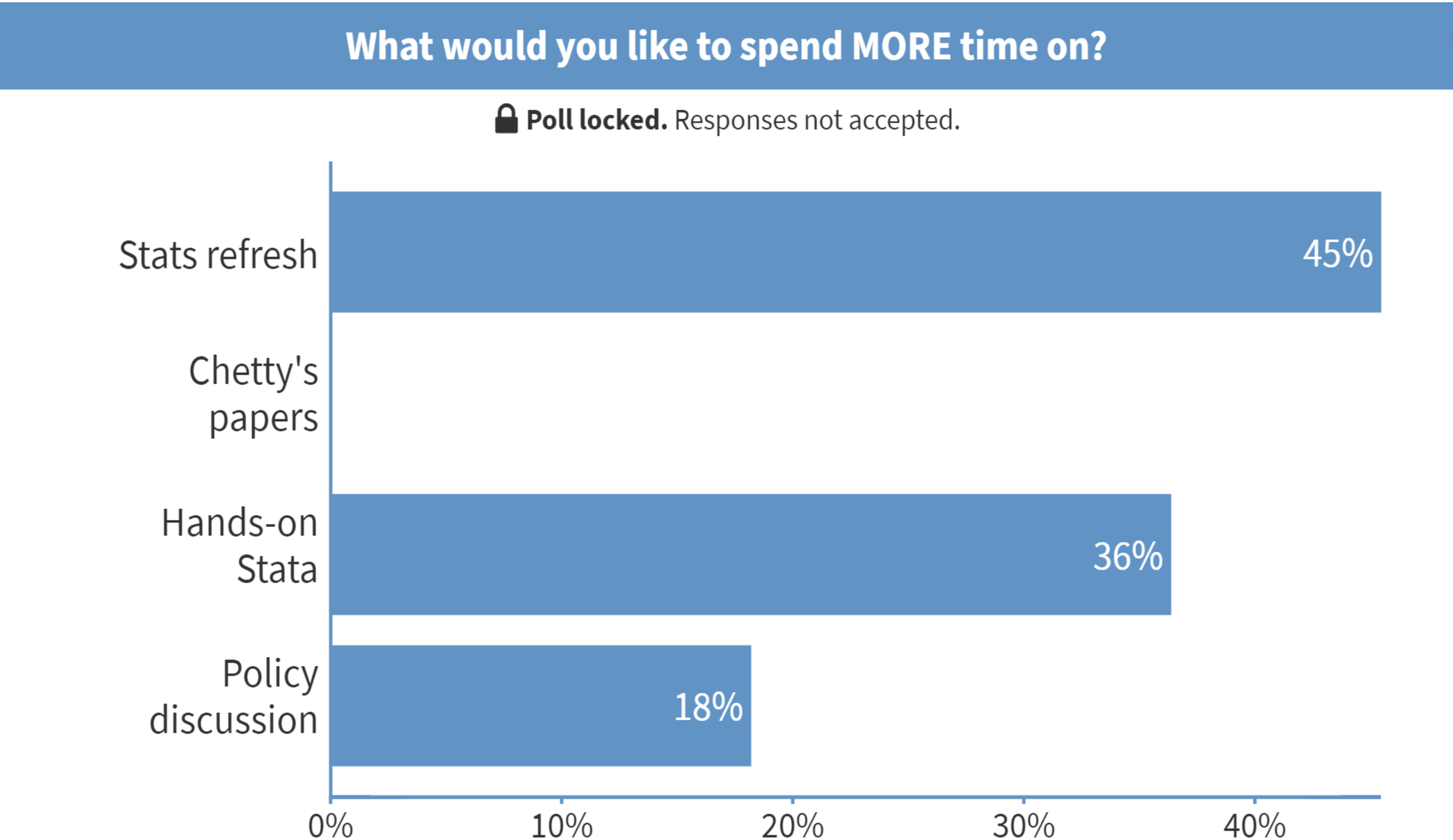
Review Session #1
TF: Diana Goldemberg

Prof: Raj Chetty
Harvard University
Spring 2019

Disclaimer: multiple TFs contributed to those slides. All mistakes are mine.

Logistics

- I’m Diana.
- Take 2 min to fill out this survey please [bit.ly/ec1152d998]
  Find this prez at: https://github.com/dianagold/Ec1152_diana
- We’ll meet every Friday @ 10.30-11.30am (Sever 201)
  - Advanced section, primarily for grad students but undergrads welcome!
- Office Hours:
  - Wednesdays @ 4.30-6.30pm (Barker 103)
  - I’m also available by appointment and after sections.
- Expectations:
  - Email (**diana_goldemberg@g.harvard.edu**) response times: within 24 hours M-F; 48 hours on the weekend
  - Google form to submit questions before section


What would you like to spend MORE time on?
Poll locked. Responses not accepted.
Stats refresh
45%
Chetty's papers
Hands-on Stata
36%
Policy discussion
18%
0%
10%
20%
30%
40%

Stats Refresh

Statistical concepts to refresh

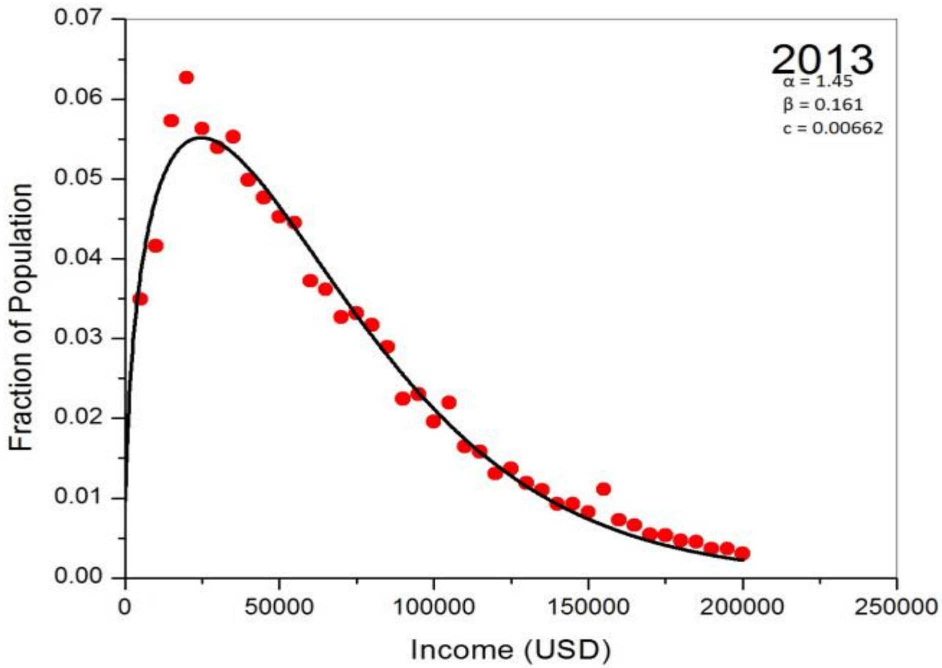
- Summary Stats: mean, median, mode, percentile, st deviation, variance, probability distribution function (pdf), cumulative distribution function (cdf)

- Statistical Inference: sample and population, estimate and st error, confidence intervals, hypothesis testing, p-values

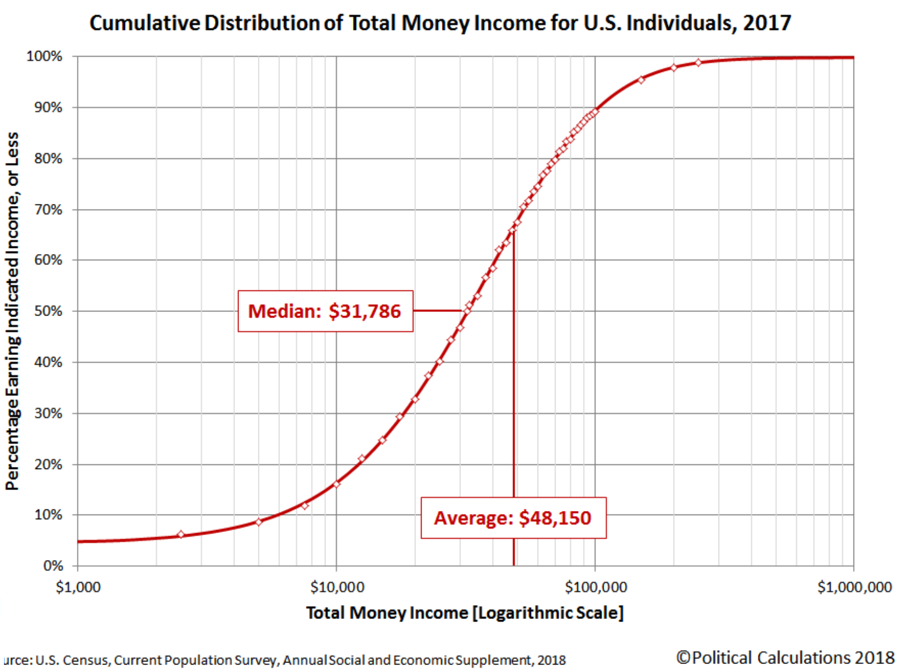
- Regression Analysis: motivation, interpreting coefficients (with/without standardizing variables), correlation is not causation

Note: look through the slides from the introductory section if you want a more basic (and spelled out) version

Summary Statistics



Probability Distribution Function



Cumulative Distribution Function

Anscombe's Data

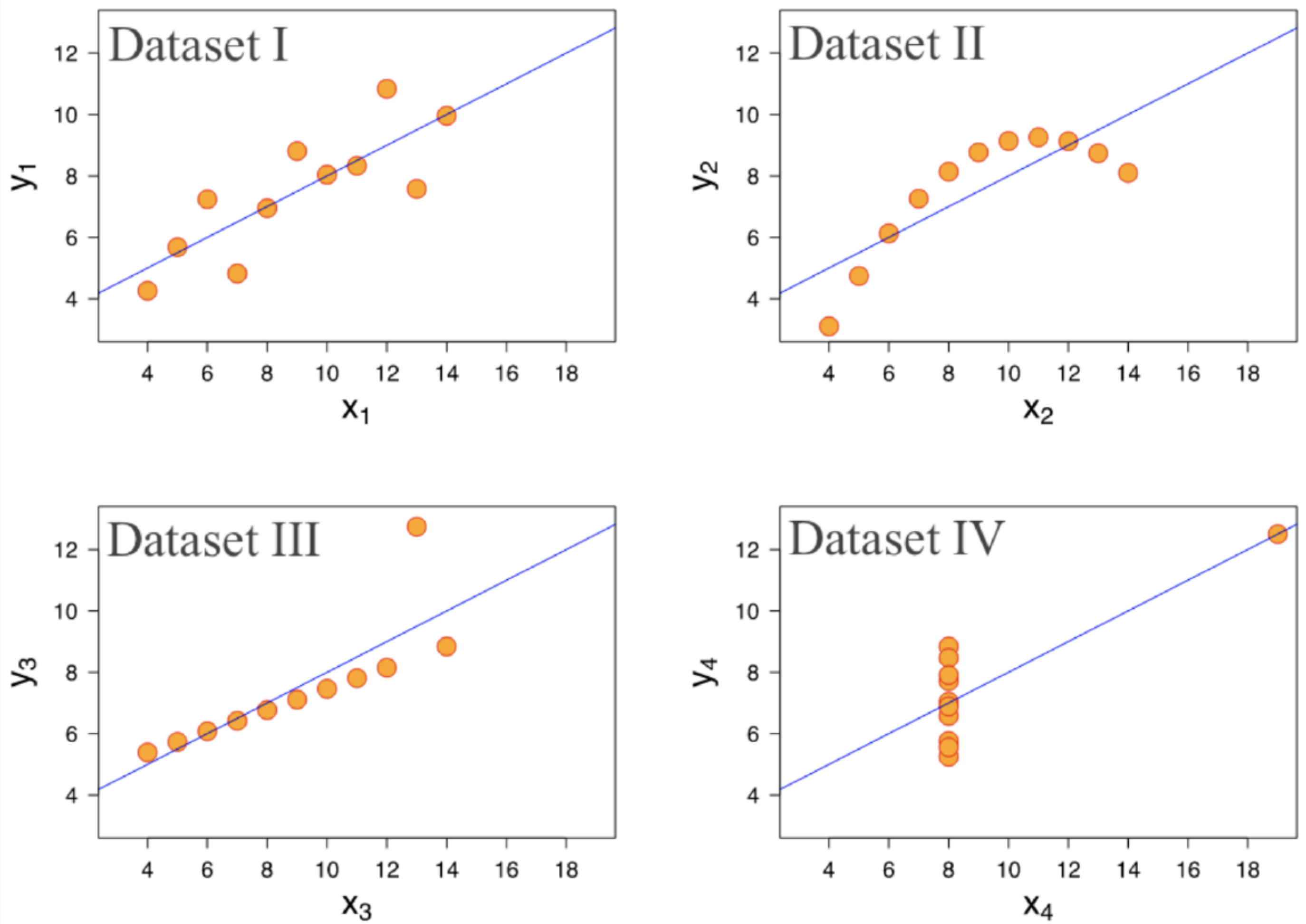
The following four datasets comprise the Anscombes Quartet (1973); all four sets of data have identical simple summary statistics

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Note: slide shamelessly copied from CS109A

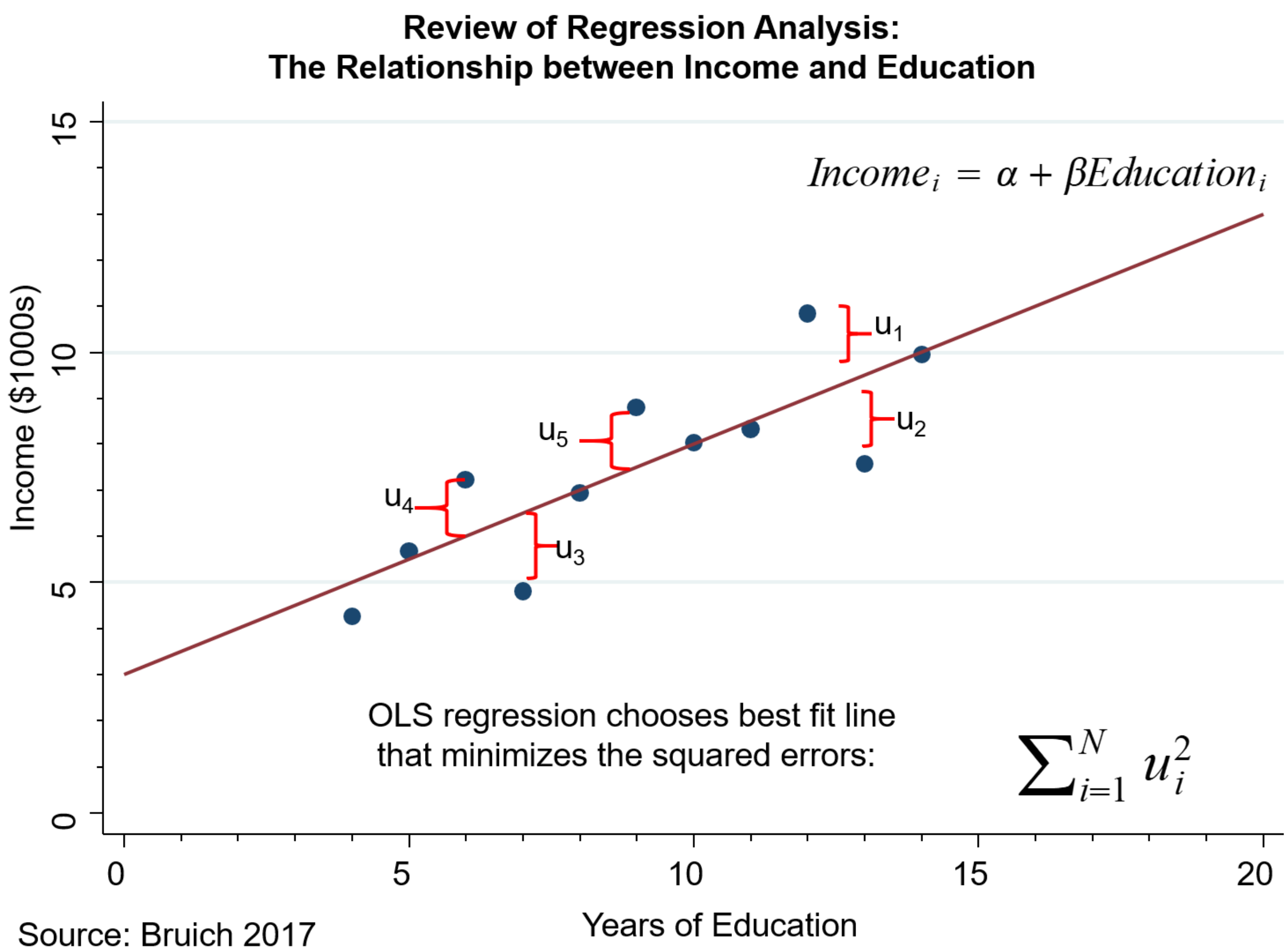
Anscombe's Data

Same summary statistics also mean you fit the same regression line.
But a picture can be worth a thousand words:



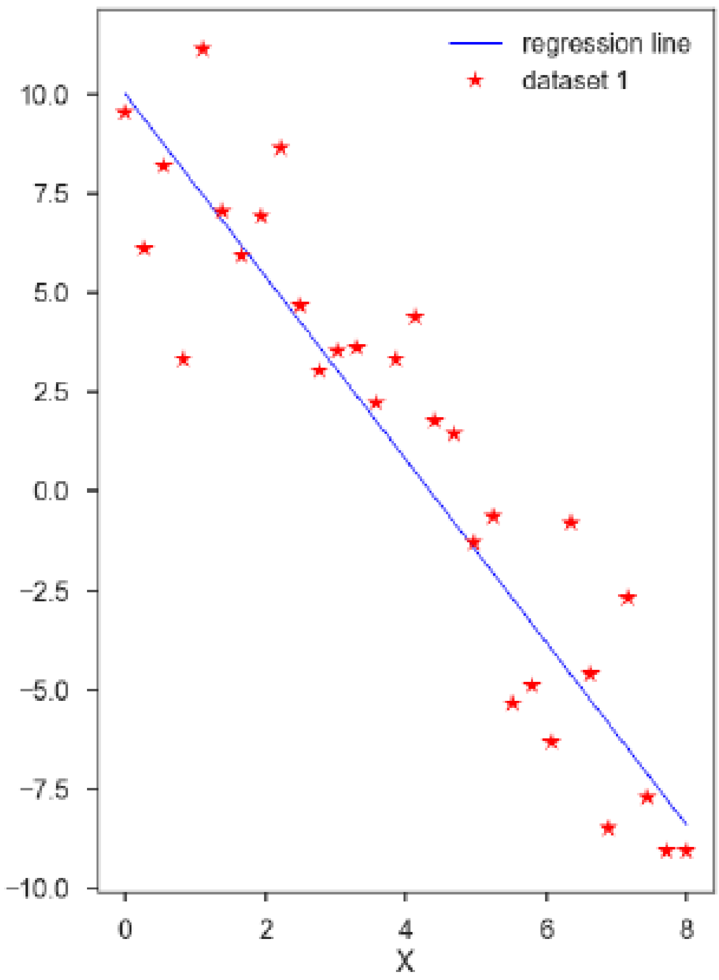
Note: slide shamelessly copied from CS109A

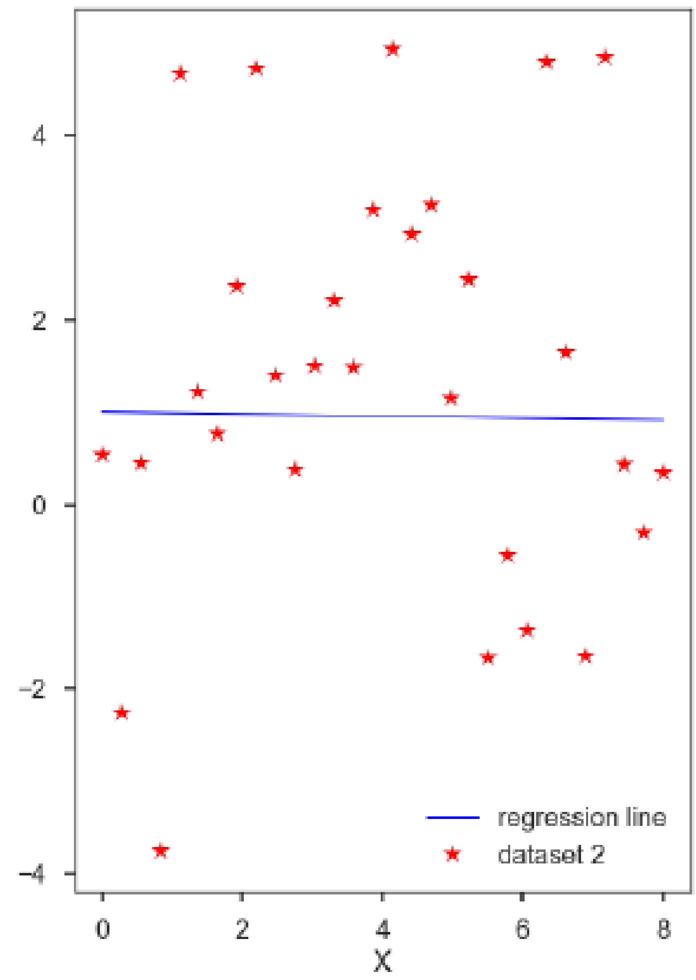
Regression Analysis



- What's an interpretation of α and of β?
- Inference for linear regression? (how well do we know α and β?)

Evaluating Significance of Predictors

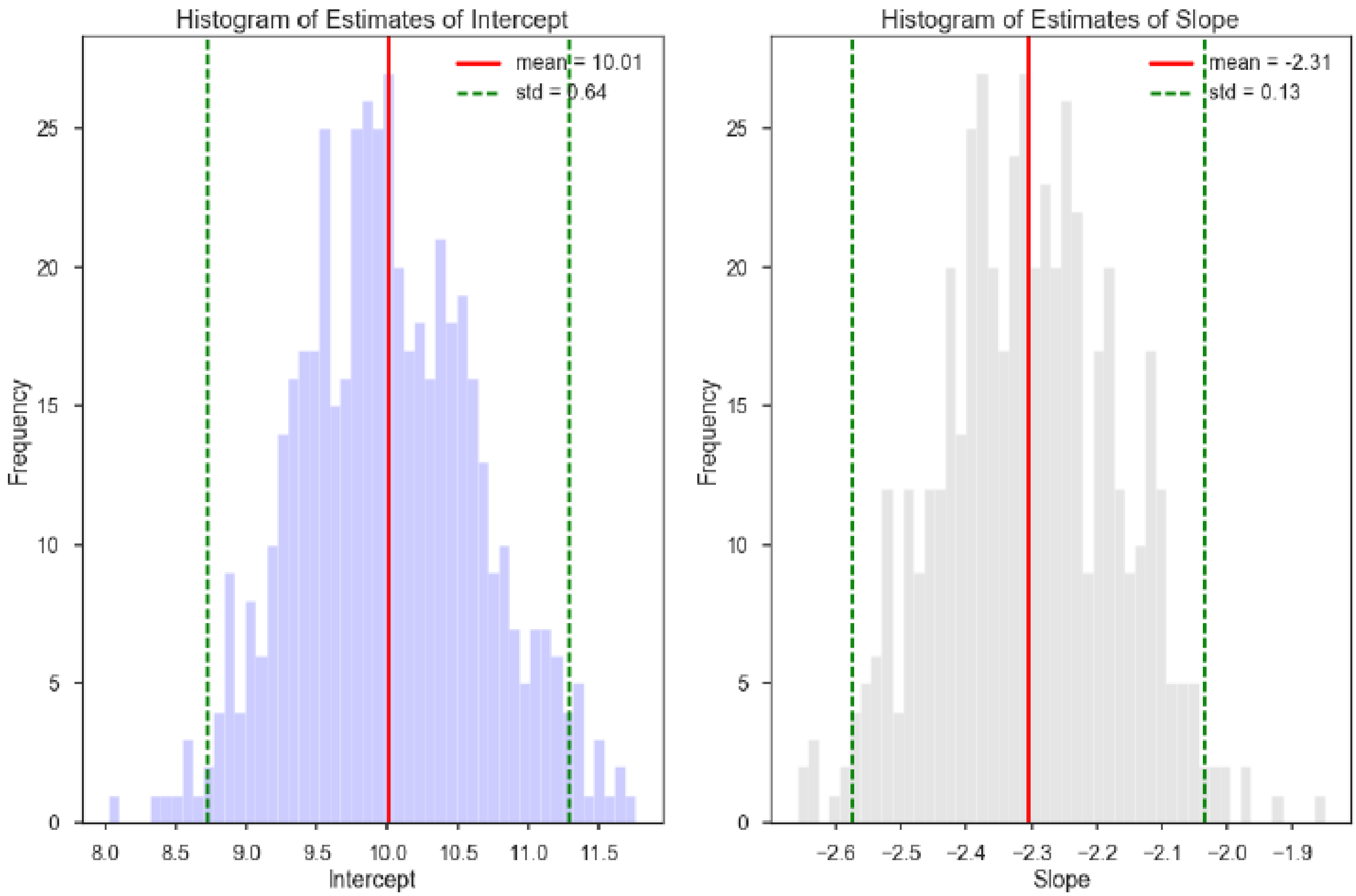




Note: slide shamelessly copied from CS109A

Evaluating Significance of Predictors

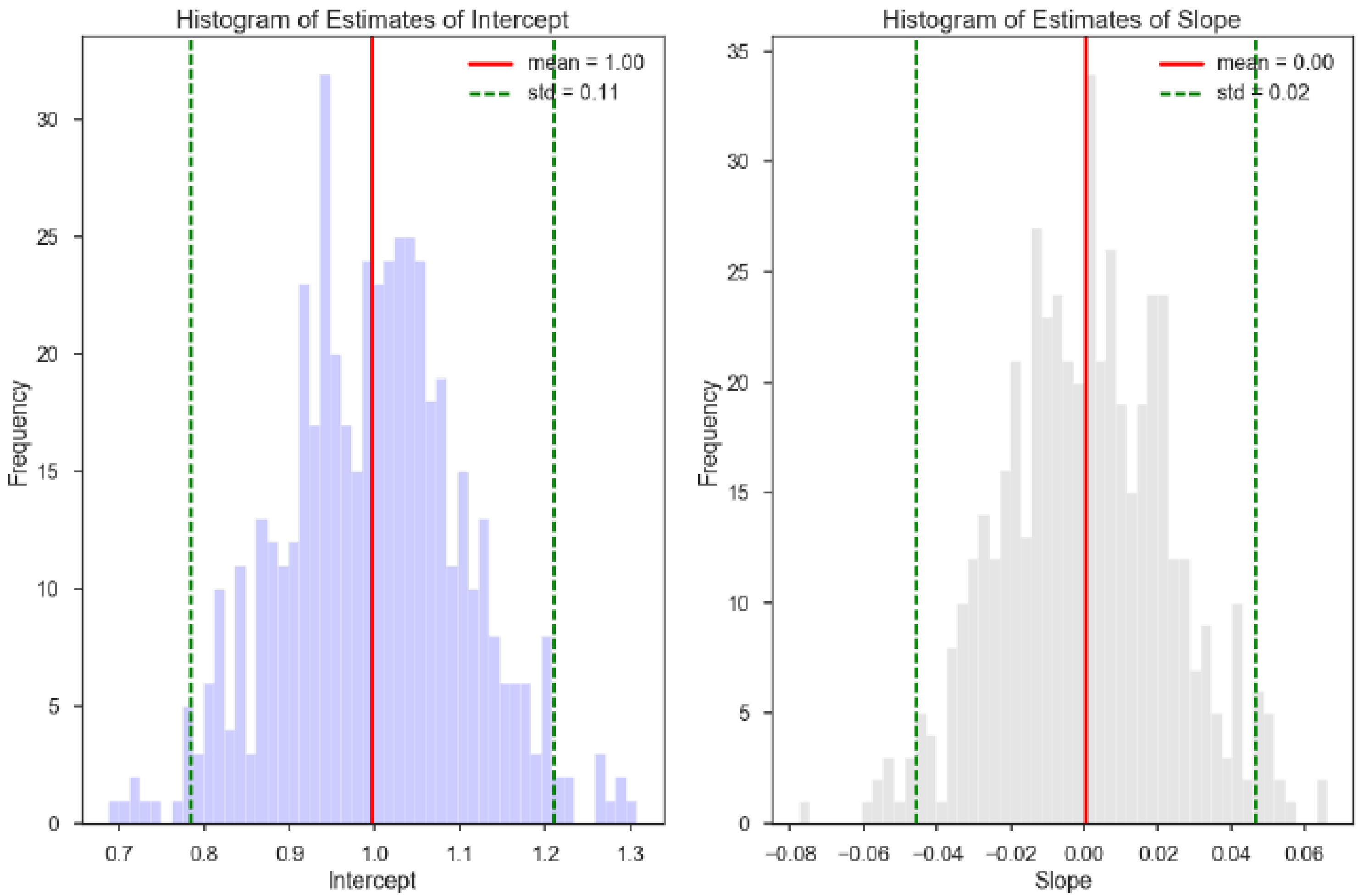
α-hat and β-hat for dataset 1



Note: slide shamelessly copied from CS109A

Evaluating Significance of Predictors

α-hat and β-hat for dataset 2



Note: slide shamelessly copied from CS109A

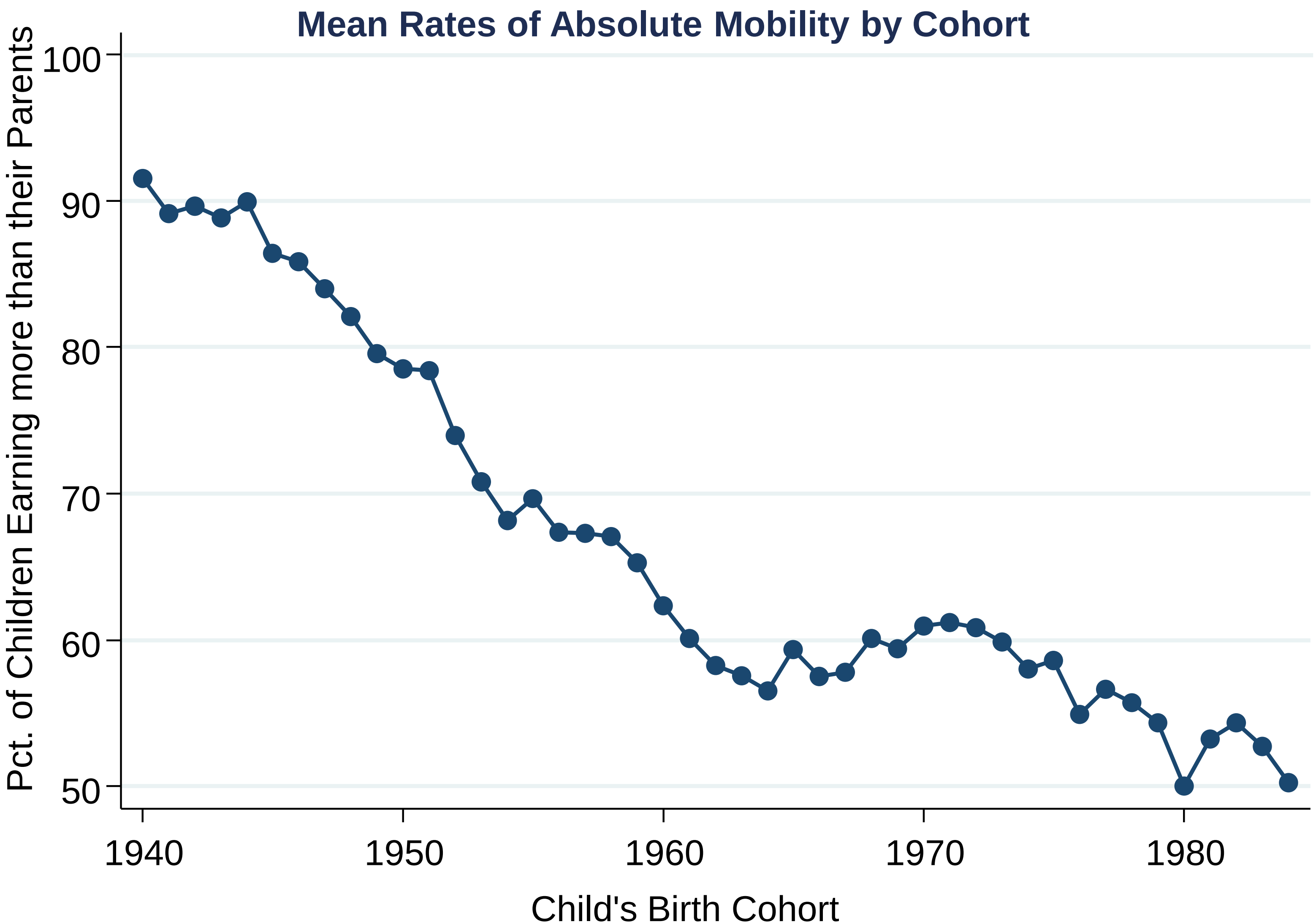
Closer look at Chetty's papers discussed in lecture

- *The Fading American Dream: Trends in Absolute Income Mobility Since 1940*. Raj Chetty, David Grusky, Maximilian Hell, Nathaniel Hendren, Jimmy Narang. Science 356(6336): 398-406, 2017
- *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.* Raj Chetty, John Friedman, Nathaniel Hendren, Maggie R. Jones, Sonya R. Porter. NBER Working Paper, 2018
- *The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment*. Raj Chetty, Nathaniel Hendren, Lawrence Katz. American Economic Review 106(4): 855-902, 2016

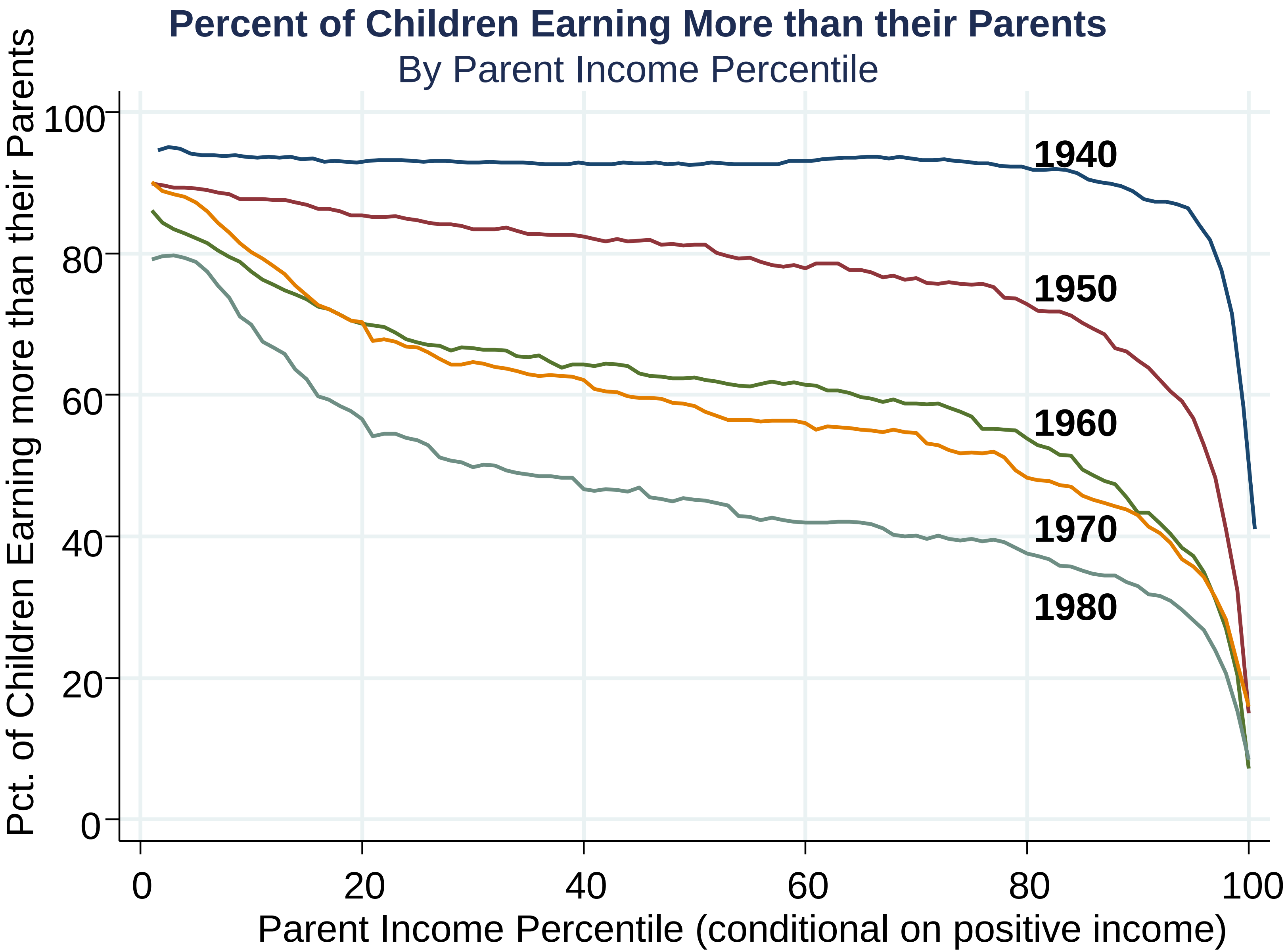
Disclaimer: slides in this section were copied from the corresponding ppts/papers in the Opportunity Insights website

Absolute Mobility

- What data is behind each of those points?

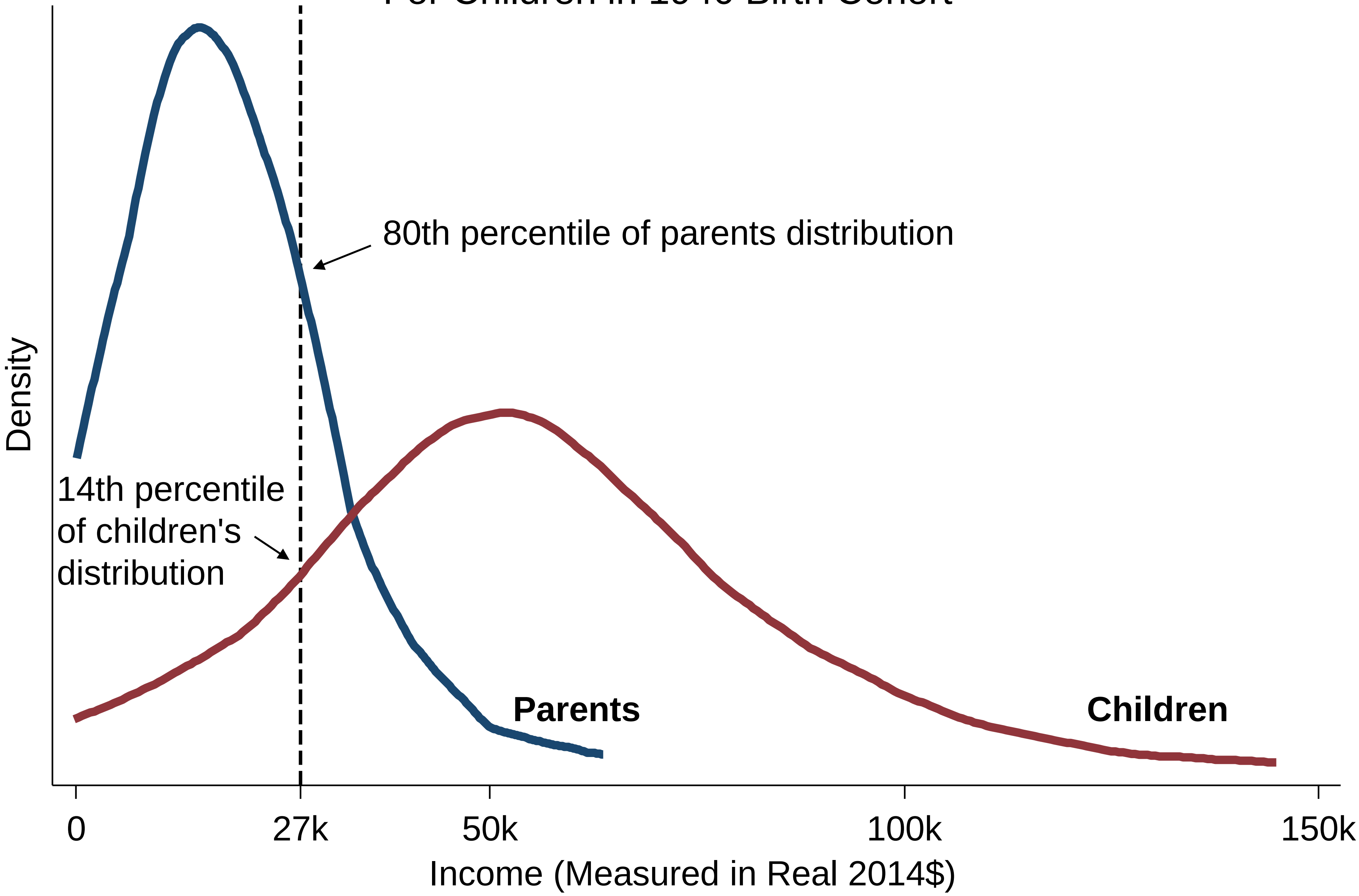


Absolute Mobility



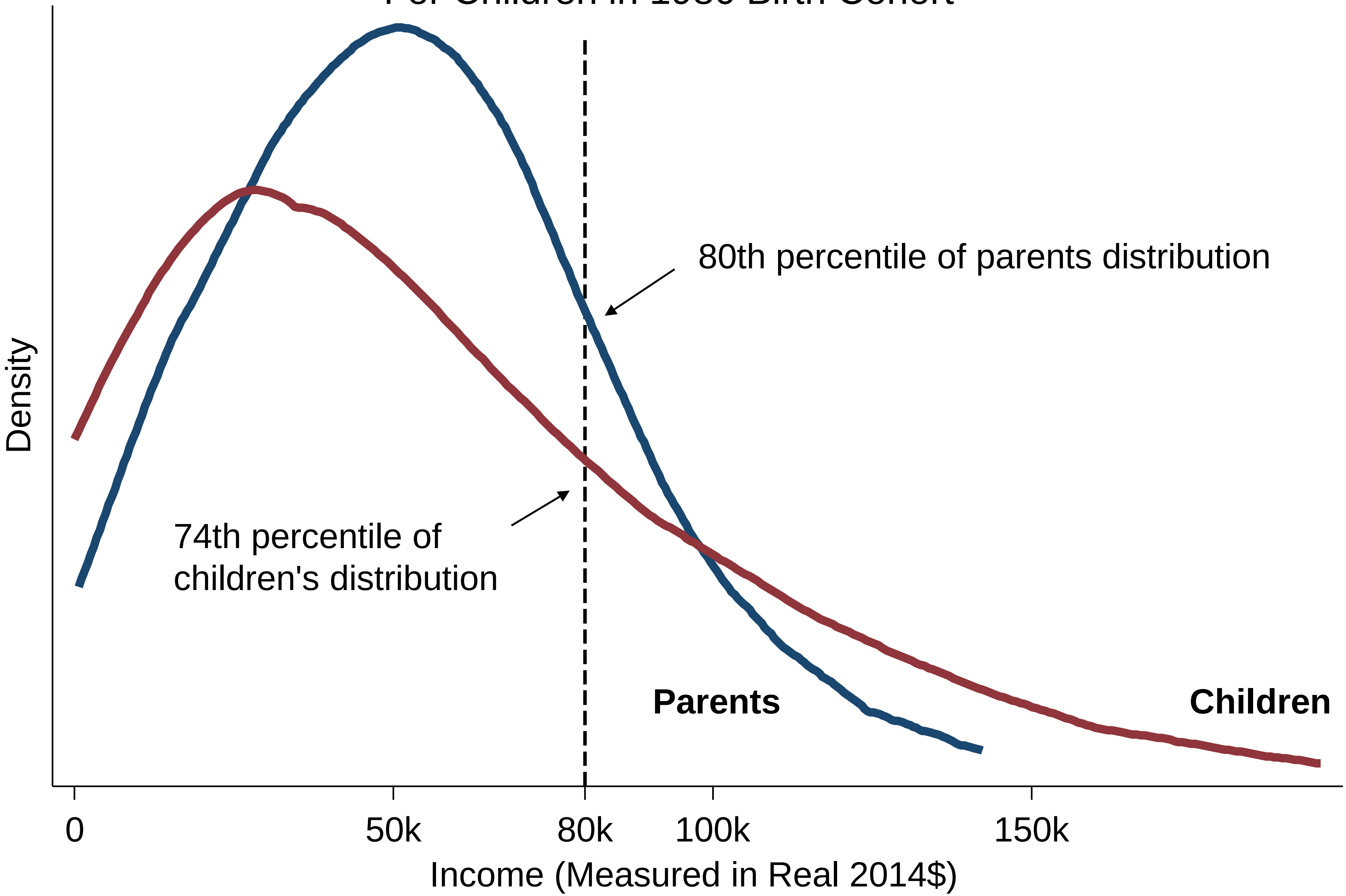
Household Income Distributions of Parents and Children at Age 30

For Children in 1940 Birth Cohort

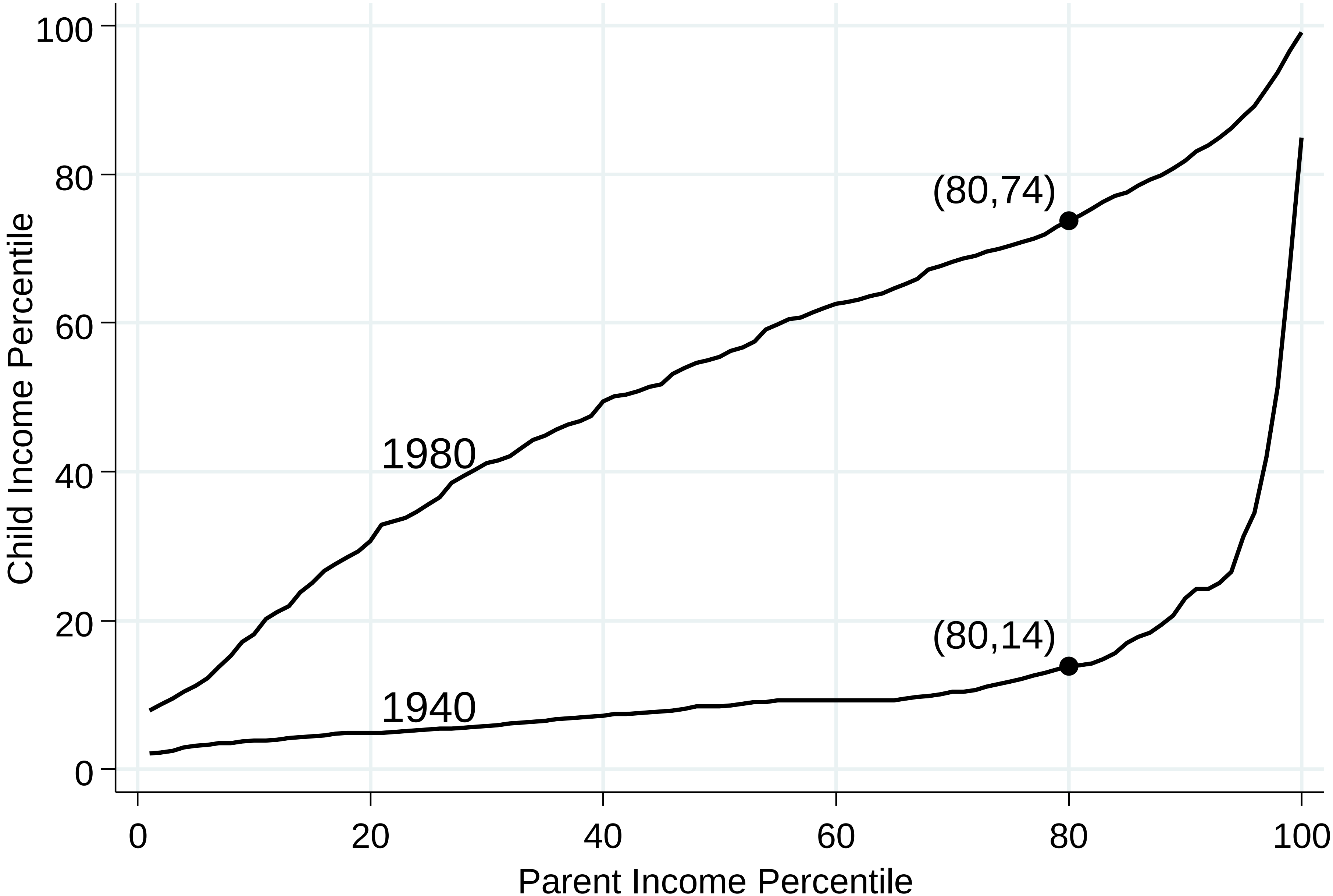


Household Income Distributions of Parents and Children at Age 30

For Children in 1980 Birth Cohort

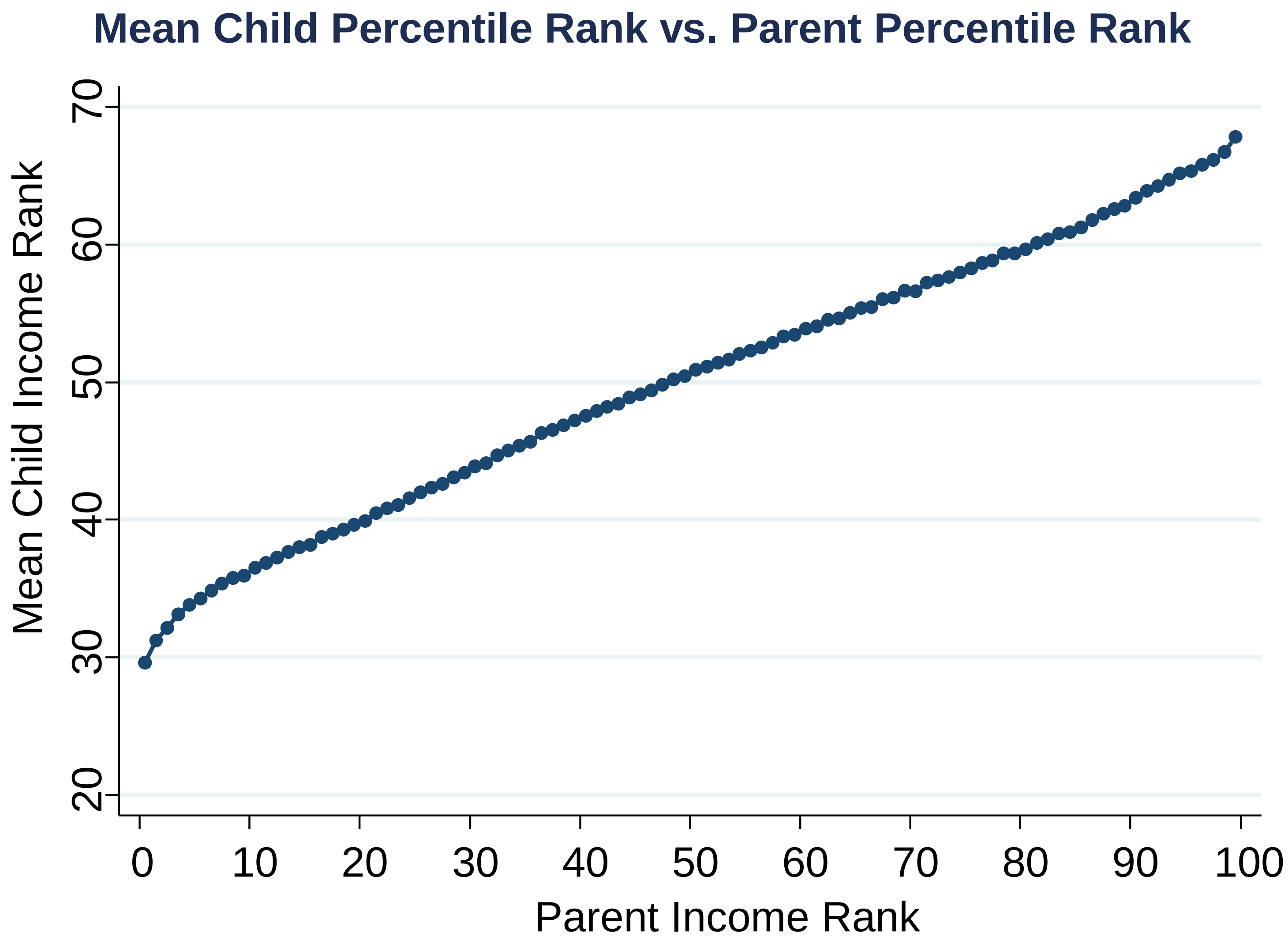


Child Rank Required to Earn More Than Parents



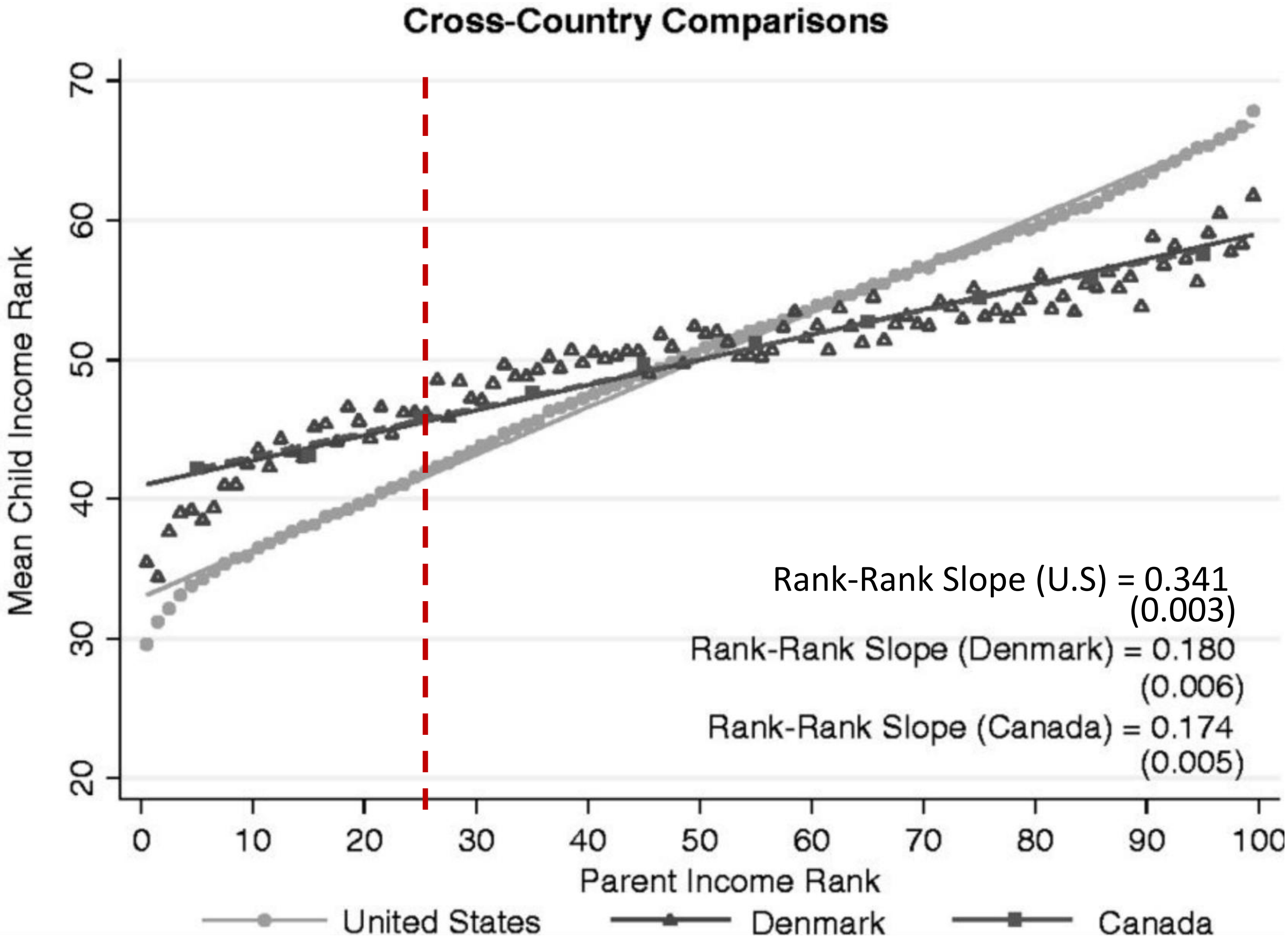
Intergenerational Mobility

- Think of measures that translate mobility



Intergenerational Mobility

- In simple terms: how well do kids from poor parents do?



MTO: Estimating Treatment Effects

- Regression specifications:

$$y_i = \alpha + \beta_E^{ITT} Exp_i + \beta_S^{ITT} S8_i + s_i\delta_s + \epsilon_i$$

Treatment Indicators

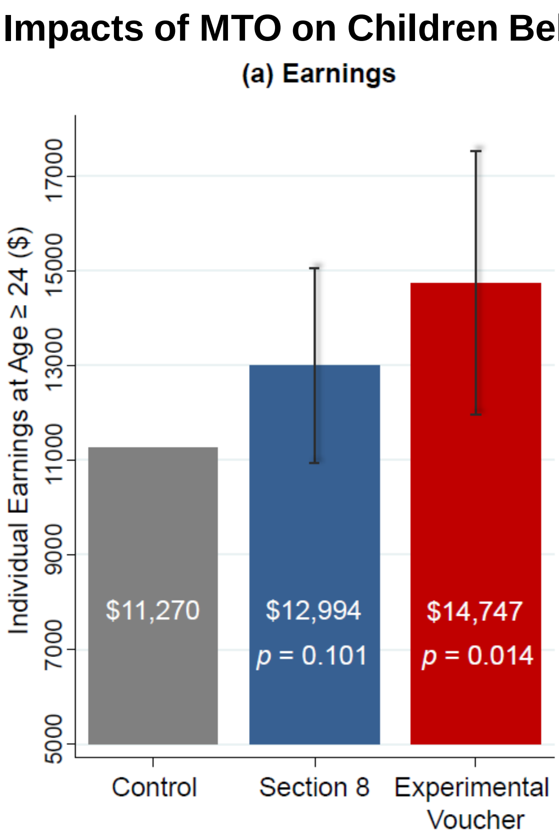
Site Indicators

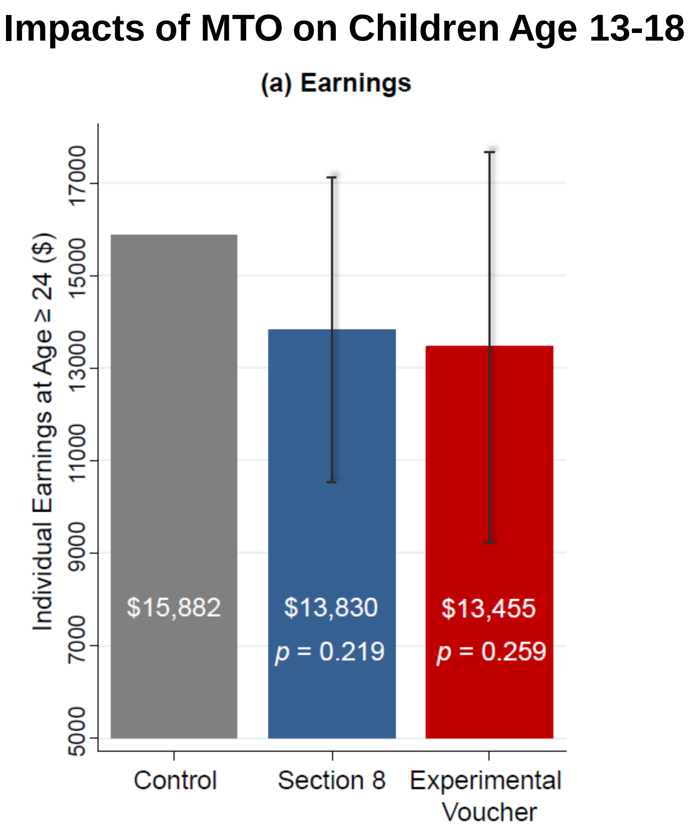
- These intent-to-treat (ITT) estimates identify effect of being *offered* a voucher to move through MTO

- From ITT to treatment-on-treated (TOT) estimates, needs to take into account voucher take-up (for young children: 48% for Exp and 66% for S8)

MTO: p-value is your friend!

- Explain as simply as possible what is the p-value translating





MTO: (

TABLE 1
Summary Statistics and Balance Tests for Children in MTO-Tax Data Linked Sample

	< Age 13 at Random Assignment			*Age 13-18 at Random Assignment*		
	Control Grp. Mean	Exp. vs. Control	Sec 8. vs. Control	Control Grp. Mean	Exp. vs. Control	Sec 8. vs. Control
	(1)	(2)	(3)	(4)	(5)	(6)
Linked to tax data (%)	86.4	-0.8	-0.4	83.8	1.5	-0.1
		(1.4)	(1.5)		(2.0)	(2.2)
Child's age at random assignment	8.2	-0.1	-0.0	15.1	0.1	-0.1
		(0.1)	(0.1)		(0.1)	(0.1)
Household Head Completed High School (%)	34.3	4.2+	0.4	29.5	5.0	0.7
		(2.4)	(2.6)		(3.1)	(3.3)
Household Head Employed (%)	23.8	1.0	-2.2	25.3	3.0	-0.4
		(2.1)	(2.2)		(2.9)	(3.0)
Household Head gets AFDC/TANF (%)	79.5	0.6	1.8	75.0	-0.8	-1.0
		(1.9)	(2.0)		(2.9)	(3.0)
Household Head never married (%)	65.1	-4.3+	-3.1	53.0	-3.1	-6.3+
		(2.3)	(2.6)		(3.2)	(3.4)
Household Head had teenage birth (%)	28.6	-0.9	-0.3	29.1	-3.6	-2.5
		(2.2)	(2.5)		(2.9)	(3.2)
N. of Children in Linked MTO-Tax Data	1613	1969	1427	686	959	686

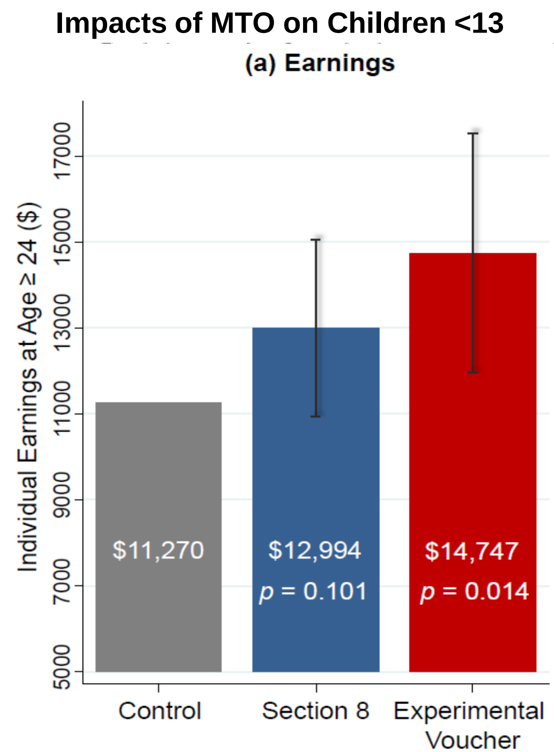
[there were more lines in here]

Notes: This table presents summary statistics and balance tests for match rates and a subset of variables collected prior to randomization; Appendix Table 1a replicates this table for all 52 control variables we use in our analysis. The estimates in the first row (fraction linked to tax data) are based on all children in the MTO data who were born in or before 1991. The estimates in the remaining rows use the subset of these observations successfully linked to the tax data. Columns 1-3 include children below age 13 at random assignment; Columns 4-6 include those above age 13 at random assignment. Columns 1 and 4 show the control group mean for each variable. Columns 2 and 5 report the difference between the experimental voucher and control group, which we estimate using an OLS regression (weighted to adjust for differences in sampling probabilities across sites and over time) of each variable on indicators for being assigned to the experimental voucher group, the section 8 voucher group, as well as indicators for randomization site. Columns 3 and 6 report the coefficient for being assigned to the section 8 group from the same regression. The estimates in Columns 2-3 and 5-6 are obtained from separate regressions. Standard errors, reported in parentheses, are clustered by family ($^{+}$ = p<0.10, * = p<0.05, ** = p<0.01). The final row lists the number of individuals in the control, experimental, and section 8 groups in the linked MTO-Tax data sample.

TABLE 3

Impacts of MTO on Children's Inc

Dep. Var.:	W-2 Earnings ($)	Indiv. Earnings 2008-12 ($)			Indiv	Inc. Growth ($)
	2008-12 ITT	ITT	ITT w/Cntrls.	TOT	Age 2	2008-12 ITT
	(1)	(2)	(3)	(4)	(5	(9)
Panel A: Children < Age 13 at Random Assignment						
Exp. vs. Control	1339.8*	1624.0*	1298.9*	3476.8*	175	1309.4*
	(671.3)	(662.4)	(636.9)	(1418.2)	(917	(518.5)
Sec. 8 vs. Control	687.4	1109.3	908.6	1723.2	551	800.2
	(698.7)	(676.1)	(655.8)	(1051.5)	(888	(517.0)
Num of Obs.	8420	8420	8420	8420	162	8420
Control Group Mean	9548.6	11270.3	11270.3	11270.3	113	4002.2
Panel B: Children Age 13-18 at Random Assignment						
Exp. vs. Control	-761.2	-966.9	-879.5	-2426.7	-53	-693.6
	(870.6)	(854.3)	(817.3)	(2154.4)	(79	(571.6)
Sec. 8 vs. Control	-1048.9	-1132.8	-1136.9	-2051.1	-15	-885.3
	(932.5)	(922.3)	(866.6)	(1673.7)	(84	(625.2)
Num of Obs.	11623	11623	11623	11623	23	11623
Control Group Mean	13897.1	15881.5	15881.5	15881.5	139	4128.1



Notes: Columns 1-3 and 5-9 report intent-to-treat (ITT) estimates from OLS regressions (weighted to adjust for differences in sampling probabilities across sites and over time) of an outcome on indicators for being assigned to the experimental voucher group and the section 8 voucher group as well as randomization site indicators. Column 4 reports treatment-on-the-treated (TOT) estimates using a 2SLS specification, instrumenting for voucher takeup with the experimental and section 8 assignment indicators. Standard errors, reported in parentheses, are clustered by family (+ = p<0.10, * = p<0.05, ** = p<0.01). Panel A restricts the sample to children below age 13 at random assignment; Panel B includes children between age 13 and 18 at random assignment. The estimates in Panels A and B are obtained from separate regressions. The number of individuals is 2,922 in Panel A (except in column 5, where it is 1,625) and 2,331 in Panel B. The dependent variable in Column 1 is individual W-2 wage earnings, summing over all available W-2 forms. Column 1 includes one observation per individual per year from 2008-12 in which the individual is 24 or older. Column 2 replicates Column 1 using individual earnings as the dependent variable. Individual earnings is defined as the sum of individual W-2 and non-W-2 earnings. Non-W-2 earnings is adjusted gross income minus own and spouse's W-2 earnings, social security and disability benefits, and UI payments, divided by the number of filers on the tax return. Non-W-2 earnings is recoded to 0 if negative and is defined as 0 for non-filers. Column 3 replicates Column 2, controlling for the characteristics listed in Online Appendix Table 1a. Column 4 reports TOT estimates corresponding to the ITT estimates in Column 2. In Column 5, we measure earnings in the year when the individual is 26 years old. In Column 6, we measure earnings in 2012, limiting the sample to those 24 or older in 2012. Columns 7-9 replicate Column 1 with the following dependent variables: employment (an indicator for having positive W-2 earnings), household income (adjusted gross income plus tax-exempt social security benefits and interest income for those who file tax returns, the sum of W-2 wage earnings, SSDI benefits, and UI benefits for non-filers, and 0 for non-filers with no W-2 earnings, SSDI, or UI benefits), and individual earnings growth (the change in individual earnings between year *t-5* and the current year *t*).

Stata hands-on demo

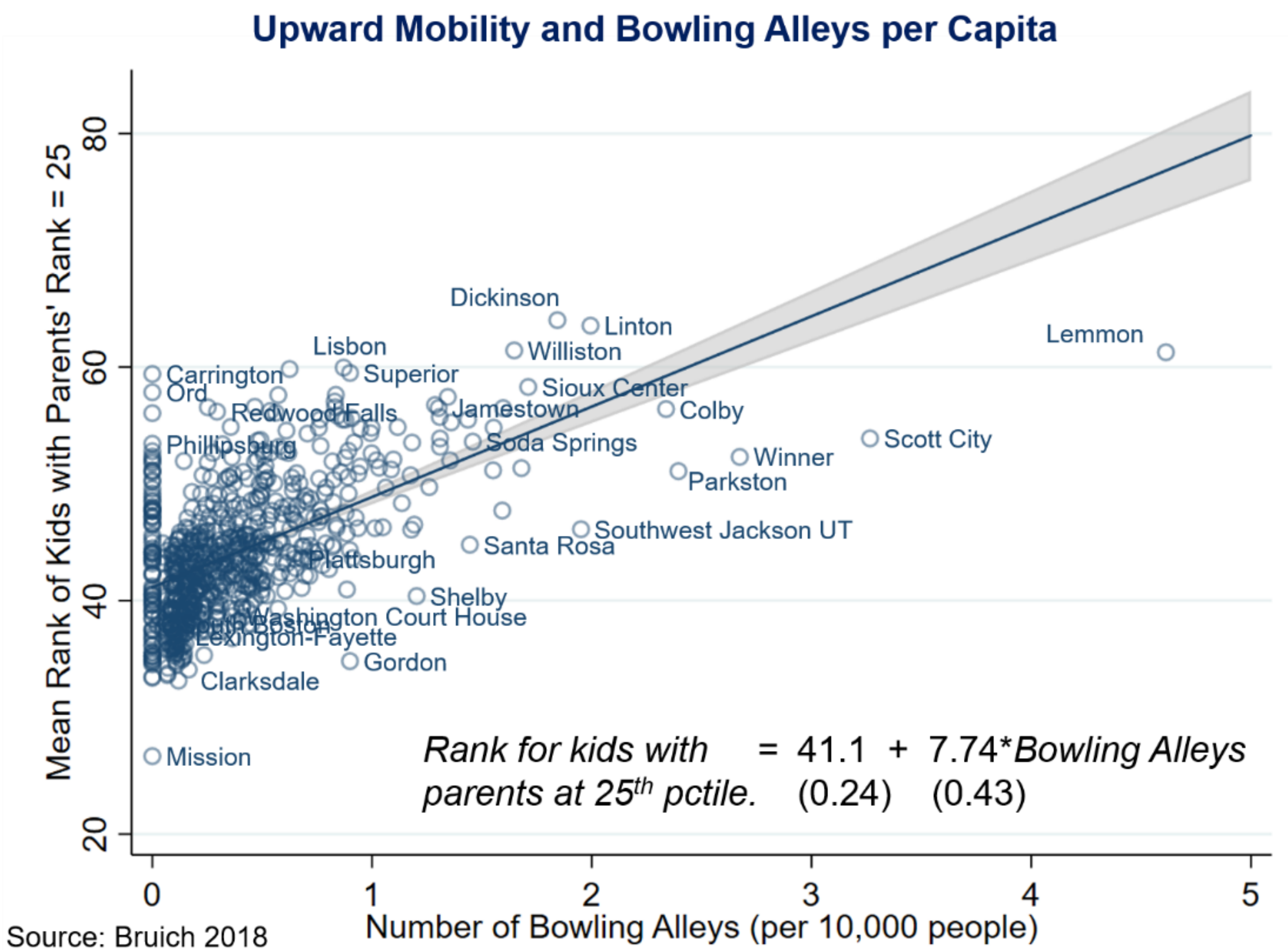
- *Stata will be used in section*
- *But you're very welcomed to follow the Jupyter notebooks for:*
  - *R*
  - *Python*
- *All files at:* https://github.com/dianagold/Ec1152_diana

Stata demo

Required files at: https://github.com/dianagold/Ec1152_diana

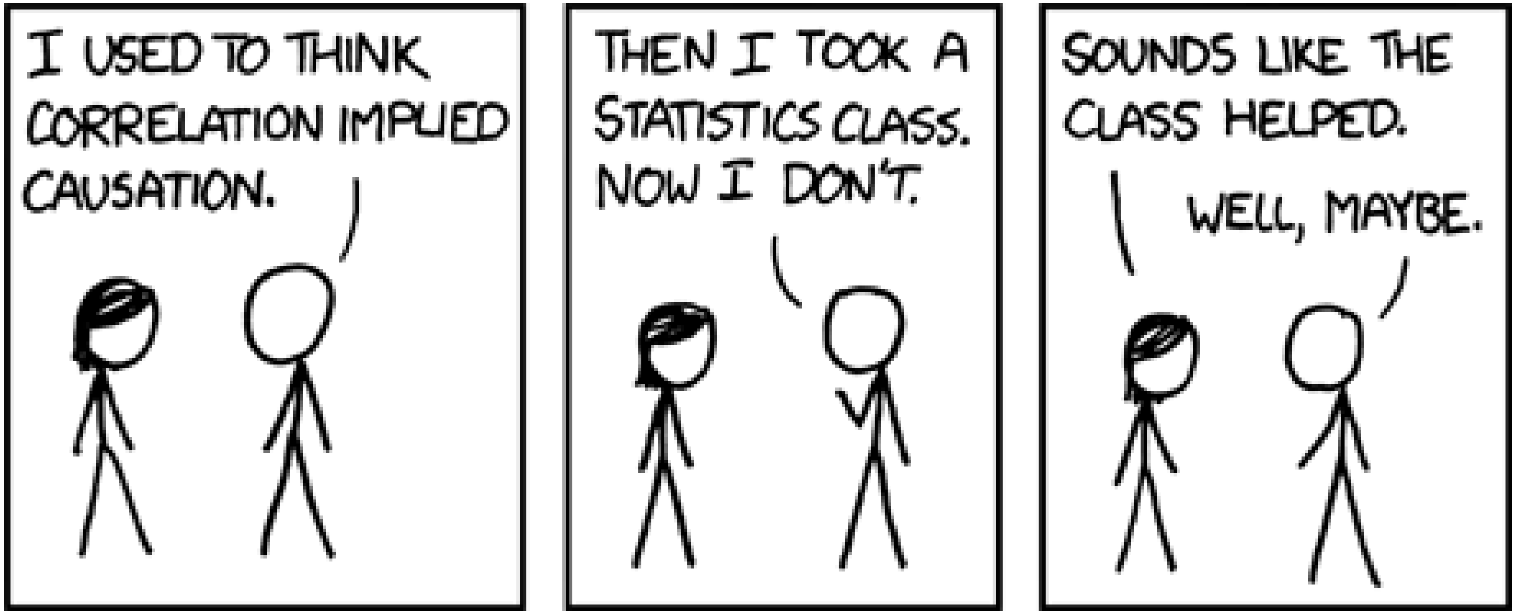
- If you have Stata in your computer, you may want to do it along
  - How to install & hints: https://canvas.harvard.edu/courses/19323
  - Optional workshop: Monday at 5:30 pm in Emerson Hall 105
- Why are we using Stata?
  - The most popular software used by economists for applied econometrics
  - Works for "big data": up to 20 billion observations and 32 thousand variables (contingent on RAM)
- Upward mobility (Y) as a linear regression of Bowling Alleys per capita (X)
- Tasks:
  - Get means and stdevs
  - Standardize Y and X
  - Use OLS to estimate correlation coefficients

Stata demo [output today]

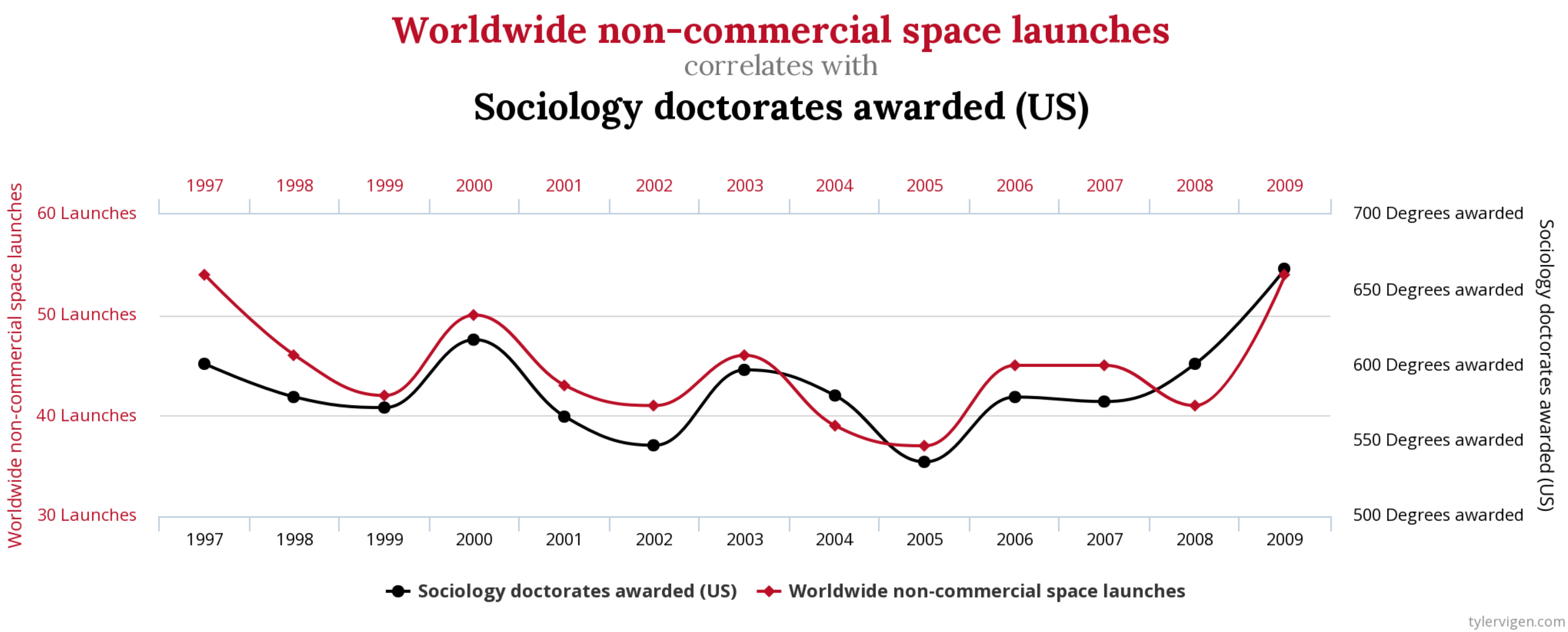


Correlation is not causation!

- Have you seen this meme before?
- Have you take a Stats class before?
- Correlation or causation?



Correlation is not causation!



Policy Discussion

Policy discussion (starters)

- Shaun Donovan on “the politics of getting it done”:
  - Can we spend money while actually saving money?
    The selling point for Culhane’s research
  - Mayors competing (what role for media & public attention?)
- Chetty’s lecture on MTO:
  - Experimental vouchers as “investment in the future generation”
    Intertemporal compromises
  - “Opportunity bargains”: information and transaction costs
    (making the process seamless)

MTO Conclusion slide: Policy Lessons

- How can we improve neighborhood environments for disadvantaged youth?

1. Short-term solution: Provide targeted housing vouchers at birth conditional on moving to better (e.g. mixed-income) areas

    - Taxpayers may ultimately gain from this investment
      [MTO experimental vouchers increased PDV of earnings by $100K for children who moved at young ages]

2. Long-term solution: improve neighborhoods with poor outcomes, concentrating on factors that affect children

    - Estimates here tell us which areas need improvement, but further work needed to determine which policies can make a difference