2nd class Part I

Settle down:

1. We're in this together
2. We're using exactly the same syllabus ~~and~~ + problem sets as forever
3. Stata
4. Friday AM 11-12?

Heresy 101 review.

Causality

The causal effect of treatment T (e.g. attending Head start, \$5K increase in family income) for individual i on outcome Y:

is the difference between Y_i in the presence and absence of the treatment

George Bailey

Note causality is individual specific and never observed [pre-post doesn't work]

Nothing, not even random assignment, can prove causality for individual i

Huh? RA approximates causality for groups of individuals

which, for many purposes, is just fine

Why we focus on causal analysis

- policy recommendations

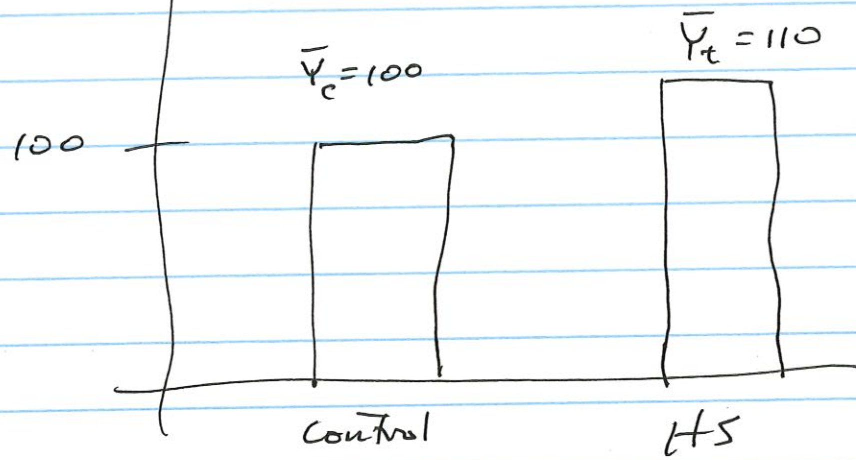
2nd class

$\boxed{\text{Every regression is an experiment}}$

Suppose group-based causation is a fine goal and we have an experiment in which

100 ~~200~~ kids assigned to HS from a waitlist
100 ~~200~~ kids not given slots in HS

Track kids for the year and administer a cognitive test with mean = 100 and sd = 15



"effect" is 10 units = 2/3 of a sd.
is this significant? Do t-test on difference in means.

$t = \cancel{4.54}$ ~~p<.0001~~

$t = -4.7 \quad p < .00001$

<u>OR</u> run a regression $Y = a + b_1 T$

where $T = 1$ if child is in treatment group
$T = 0$ if not

2nd class

what will be the estimates of $a = 100$
and $b_1 = 10$

$$Y = 100 + 10T$$

you will also get (2.1) standard error on the coefft.

coefft / s.e. gives you a t test which is exactly the same as the t test for difference in means.

$$\frac{10}{2.1} = 4.7$$

Breakout groups
Suppose two treatment groups + control:
what regression gives you T_1 vs C?
T_2 vs C?
T_1 vs. T_2?

WHY BOTHER?

Because a regression framework allows you to do a lot more

1. Adjust for imbalance. If, say, luck of the draw differences give you more low income kids in the HS group than controls,
run $Y = a + b_1 T + b_2$ Family Income

2. Reduce s.e.'s and get more statistical power

$$\text{s.e.} = \sqrt{\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}}$$

residual sum of squares (kind of $1 - R^2$) or unexplained variance in Y

week 2

If instead of

$$Y = 100 + 10T$$
$$(2.1)$$

You have $Y = 100 + 10T$
$$(1.0)$$

you have a much tighter confidence interval

not 2 to 18 or 6 − 14

but 6 − 14   9 − 11

so $Y = a + b_1 T + b_2 \text{Fam Inc} + \underline{b_3 Y_{\text{baseline}}}$

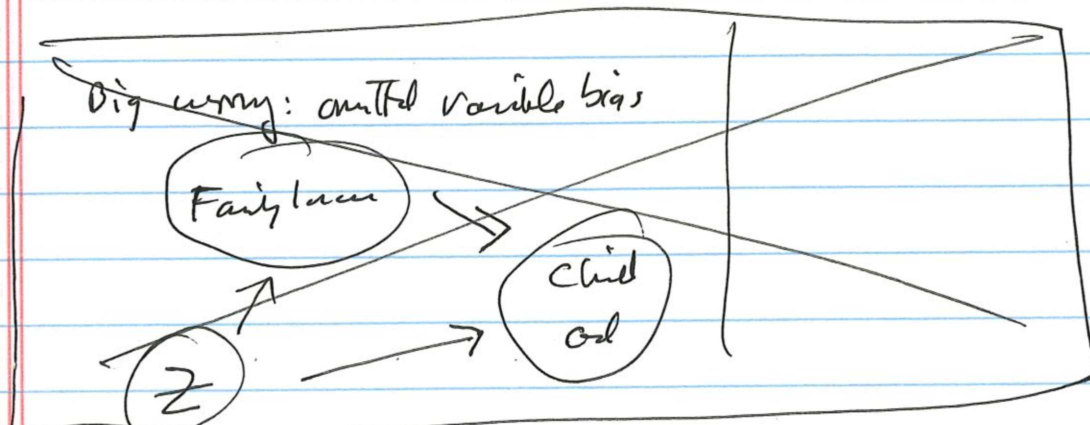
correlation between Y at 2 points is high, so R^2 is higher and σ^2 is smaller

and s.e.s are smaller

what if we don't have random assignment?

e.g. what if we have

Child Ed = a + b_1 Fam Income + controls?

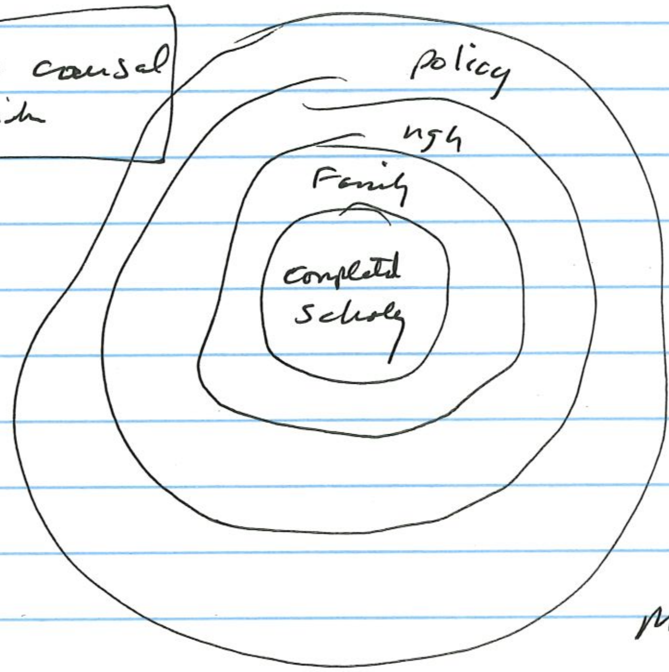


Week 2

How to model Child Ed?

Reverse causal explanation

~~Forward~~ Rev



Bronfenbrenner

Goal: Explain it as completely as possible

maximize R^2 (explained variance)

Worthy, but all but worthless for policy
Knowing that family "action" is most important doesn't tell you ~~that~~ whether family-related policies will help

Forward causal inference approach ← our focus

$$\text{Child Ed} = a + b_1 \text{Family Income} + b_2 \text{Controls}$$

~~Q~~ Can we make b_1 into an estimate of the causal effect of Family Income on Child Ed?

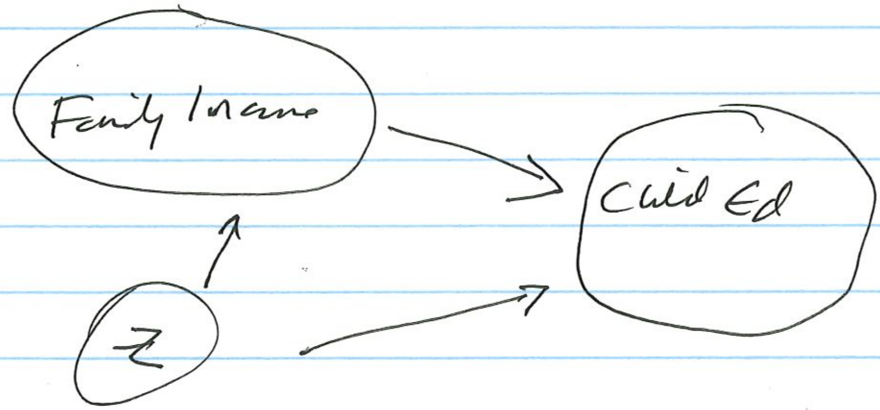
Week 2

Forward Causal Inference	Reverse Causal Explanation
Experimentalists, policy analysts, economists	Non-experimental psychologists, most ed researchers, most epidemiologists
Manipulate X to see its effect on Y - explicitly - quasi-experimentally	Explain Y as completely as possible with many Xs
R^2 doesn't matter Most coefficients are completely uninteresting	R^2 is important most coefficients are interesting
Testing hypotheses	Generating hypotheses $\boxed{\text{No policy conclusions allowed!}}$

Week 2

What is the problem?

In observational data



More formally

$$\text{Child Ed} = a + b_1 \text{Family Income} + e_i$$

b_1 is unbiased if Family Income is uncorrelated with the error term e_i

But if Z, then

$$\text{Child Ed} = a + b_1 \text{Fam Inc} + \overbrace{\left(b_2 Z + u_i\right)}^{e_i}$$

If you don't include Z explicitly and therefore control for it, then Fam Inc is correlated with e_i

-8-

Week 2

What to do?

1. Randomly assign Family Income

2. Load up the regression equation with lots of would be Z's

3. Adopt a "quasi-experimental" approach.

For examples

(1) Akee Casino revenue for Cherokee Indians + Great Smoky Mountain study.

$\boxed{\text{Age 19} = a + b_1\,\text{Age 9} + b_2\,\#\text{ of Parent}}$

Age 9 measurement → Casino comes in → Age 19/21 outcome

(2) Dahl + Lochner EITC expansion made more generous for 2^+ children families than 1 child families.

$$\underset{\text{pre-post}}{\Delta\,\text{Ach}} = a + b_1\, 2^+ \text{ children}^{(1,0)}$$

$\boxed{\text{Heritage Key B}}$