7th class Part I Paper and Diff-in diff

Paper -- see guidelines

- attempt at a coherent regression analysis.
- start with a question or hypothesis
  sharper the better
  doesnt have to be supported, just well tested.
- any data -- the class data are fine
  I will work with you on this.
- structure - walk through handout.

7^{th} class Part I

Diff-in-diff

Using regression for causal analysis RD, diff-in-diff, fixed effects

Today diff-in-diff to take advantage of natural experiments

Recall the basic idea of natural experiments

NOT: Data set → what can I do with this data set?

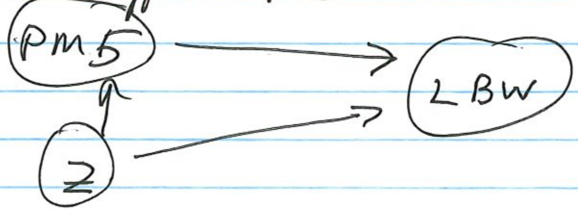
BUT: Event that creates exogenous variation in a measure/policy of interest → Assemble data and construct model to do the test

Example: Currie and Walker: what is the effect of ~~air~~ auto exhaust pollution on child health

NOT: Assemble data on, say, PM5 and ~~for~~ LBW and correlate

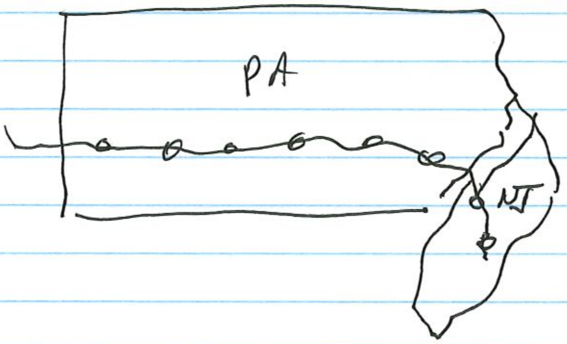
$LBW = a + b_1 PM5 + \text{controls}$

because: why do people live in different areas with different PM5



omitted variable bias

She uses the natural experimt of the roll-out of EZPass on the NJ and PA turnpikes



lots of TOLL boothes. originally everyone stopped, then EZPass -- no slowing down. ⟹ Huge reduction in emissions

If emissions matter, then you should observe reduction in ~~Chi~~ bad child outcomes for people living closest to the toll booths.

So you want to compare birth outcomes

~~(1) birth~~

1. Before and after the change to EZPass
2. For women living very close vs. a ~~bitt~~ bit further way from them

< 2 km vs 2-5 km

7th class Part I

take only births < 5 km of a toll booth

and run $BW = a + b_1 \text{ EZpass} + b_2 < 2\text{km}$

$+ b_3 \text{ EZpass} + b_2 \text{ km}$

so: b_1 is the main effect of EZ pass

b_2 is the main effect of < 2km relative to 2-5

b_3 is the extra effect of EZ pass for the < 2 km ers.
(relative to the ~~25~~ 2-5 ers)

Go over table 3 LBW coeff = -.0093

base rate (in Table 1) is abt 10% (.10)

so reduction is abt ~~10%~~ 1 ppt = ~ 10%

Actual model is more sophisticated because it includes multiple births to the same mother.

so add another difference

3. ~~For~~ Among births to the same women.

Fixed effects

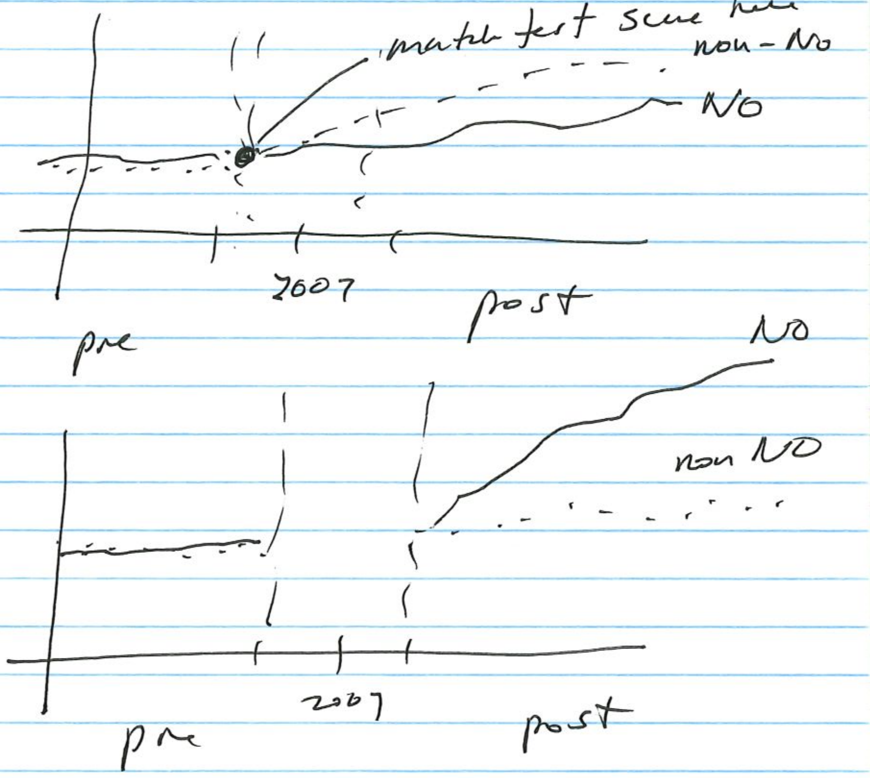
One other example. Optional -- impact of charter schools in New Orleans.

Hurricane Katrina hit in 2007 and NO school district became nearly 100% charter schools. Did kids do better?

Do better than whom?

- All other kids in US? in LA?
  No -- than kids in neighboring school districts with similar pre-Katrina test scores.

Assemble data on



Test score = $a + b_1\,2004 + b_2\,2005 + \cdots b_n$ ~~200~~ 2014

$+ c_1 *$ white NO $+ d_1$ NO $* 2004 + d_2$ NO $* 2005$

$\ldots + d_n\,2014$

Expect that d_1, d_2 would be close to zero

but d's for 2009-2014 ~~[illegible]~~ would be either positive or negative.

Applied Regression Paper Format Suggestions

In general, I am hoping that you will be spending almost as much time writing the paper as analyzing the data in it. Too often students spend almost all of their time futzing with their analyses and very little time writing it up. Our PhD program wants to develop both your analytic and writing skills.

1.) Abstract: When I sit down to write a paper, I force myself to write a ~150-word abstract first, because it you can't write the abstract, you don't yet have a story to tell. Every paper tells a story, with motivation for why what you are examining matter, your research questions, your methods and your results. You should be able to say that in 150 words.

2.) Introduction: Here you want to set up your hypothesis. In a few paragraphs, explain why your research is interesting and needed, your research question/hypothesis, and very briefly discuss any background research or theories that inform your research question (and I not expecting as much here as I would in a second-year paper).

3.) Describe your methods: sample, measures, and study procedures. Your explanation should be clear and complete enough so that I could replicate what you did if I had access to your data. Use tables of descriptive statistics (means, standard deviations, proportions, with your key measures at the top of the table) within this section. If you are looking at differences between groups, do descriptive statistics by those groups and provide p-levels from statistical tests of group differences. Describe all your independent variables in this way.
Codes to use:
tab, sum, estout
*you may need to create dummy variables for your groups. Use codebook in stata and/or the data documentation to find the appropriate identifiers for your dummy variables (Example in HW 2). Take a look at some of our course readings for ways of doing this concisely.

4.) Describe the kind of analyses you will use to test your hypotheses. Frame your description in light of the types of analyses we have discussed in class: OLS, logistic, spline, quadratic, fixed effects etc. (also residualized change and simple change, which we will cover in future classes). Including formulas may be useful.
How will you handle missing data? Are you limiting your sample in any way? How does the limited sample compare to the entire sample (use descriptives). Are your variables highly correlated? Will this correlation present a problem (multicollinearity).

5.) Present your key results in text and tables, identifying which of your coefficients tests your key hypotheses. Your tables should be publication-ready: in APA or another format with variable descriptors that make sense (not the actual names of your variables within stata). In text, report the associations between variables as we have in class and lab—what do the numbers mean?

6.) Briefly summarize your results in light of your question, hypothesis and the theories or background that inform your paper. What are the implications for your results? Are there any limitations to your study (there should be)? I am expecting this to be much shorter than would usually be the case in an article-type paper.

7th Class   Part II Fixed Effects

Continuing with our look at using regression to estimate causal relationships

Fixed effects -- adding dummy variables for all "units" in the sample
e.g. all schools in the sample
" families " " sample
" sites " a multi-site experiment.

"Econometric" fixed effects as opposed to HLM "fixed effects" which are fixed (vs. random coeffects"

Two broad uses

1. Show where the action is. eg Variance decomposition

2. Powerful control for omitted variable bias
(Powerful because it controls for both measurable AND unmeasurable sources of bias)

Heresy 101: (Econometric) fixed effects adjustments are a powerful and underappreciated technique for reducing bias

(rare in developmental studies)

7th Class Part II

Show where the action is

Fryer and Levitt Table 2

	Fall K Math
Black	−.663 (.025)
Hispanic	−.738 (.024)
⋮	
Constant	.307 (.013)
R^2	.11 ⟸ 11% of variation in test score is accounted for with 4 dummy variables

Highly significant in a statistical sense but ~~thus~~ ~ 90% of the variation is within rather than across race/ethnic groups.

But now look at Table 7: Does school quality Explain Black Students' Losing ground?

~~Math~~

change in Gap K to 3rd

~~Reading~~ Math	−.243 (.052)	−.180 (.061)
Reading	−.343 (.057)	−.214 (.065)
include school fixed effects?	No	Yes

-.243 vs. -.180 => 3/4 of the growth in the gap occurs <u>within</u> schools rather than ~~bett~~ between schools

most of the action is **not** caused by the fact that Black and White students attend different schools

Not always the case
Michelle and lead in water
Across Portland substantially higher lead in school water for Black students
within or across ~~cl~~ schools
within or across class rooms within schools

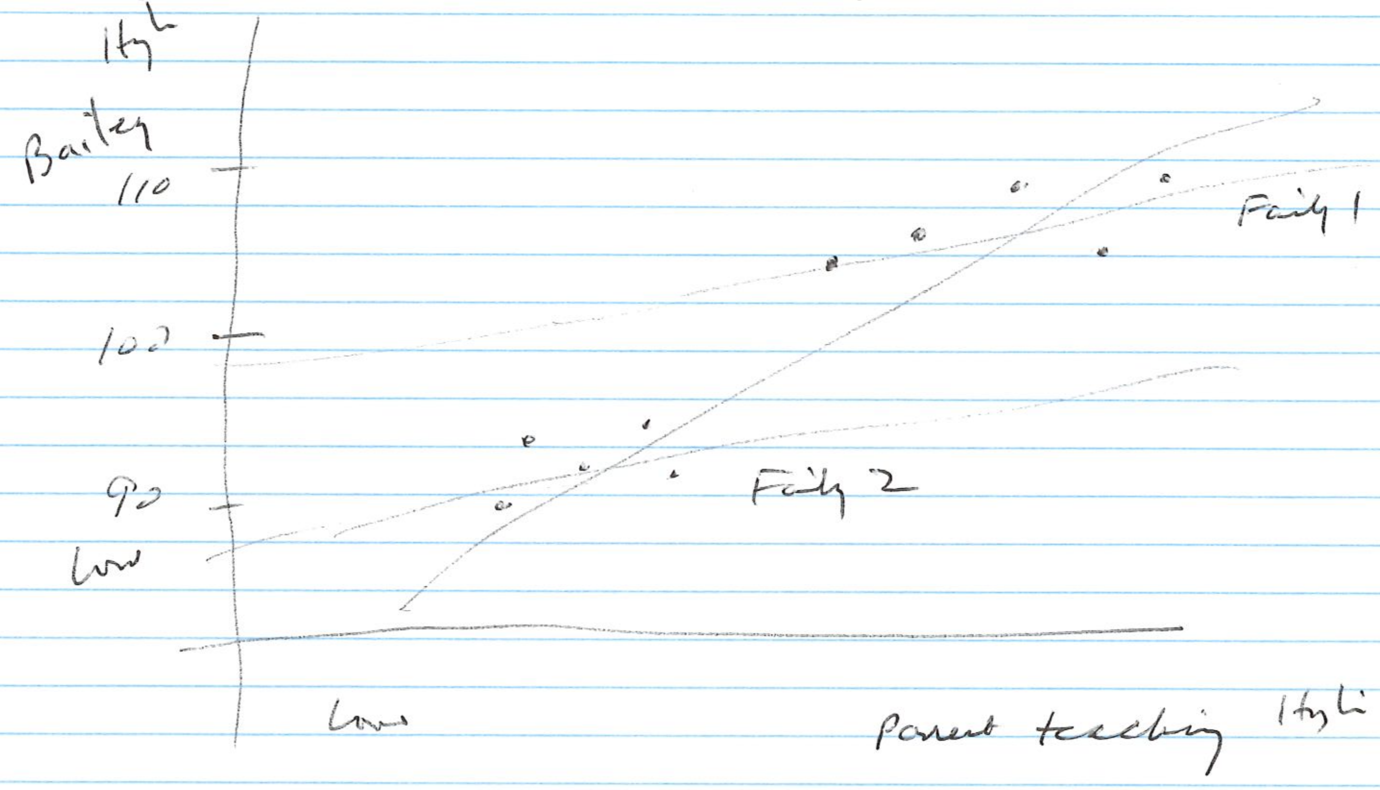
Fixed effects as a way of reducing omitted variable bias.

Suppose you are investigating the relationship between parent cognitive stimulation and child cognitive development

NCATS - parent teaching scale at 9 months

outcomes Bailey Mental Scale at 2 years

Behavior Rating Scale at 2 years.



Suppose parents are observed at 9 months with all of their children

Biased across but not within

7th Class Part II

Bailey = a + b, NCATS + whth Family 1 + . . .

allows for separate intercept ~~within time~~ for each unit

Note: no whether Family 1 × NCATS interaction

slopes are assumed to be the same within all families.

if you have 1000 families you don't want 1,000 interaction terms

Cassie's results for NCATS
vs results for Birthright

7th Class Part II

Fixed effects models

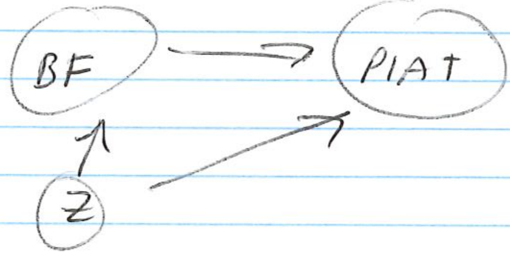
(1) Add dummy variable for every unit

(2) Use xtreg and specify unit

(3) Transform all variables so that they are deviations from their unit means.

Why does that work.

$$PIATmath = a + b_1\, Breast\ Fed + b_2\, Fixed\ Family\ charact. + b_3\, other\ stuff$$



Suppose sibling data, 2 sibs A & B

$$PIAT_A = a + \boxed{b_1} BF_A + b_2\, Family_A + b_3\, OS_A$$

$$PIAT_B = a + \boxed{b_1} BF_B + b_2\, Family_B + b_3\, OS_B$$

but Family A is the same as Family B

$$PIAT_B - PIAT_A = a + \boxed{b_1}(BF_B - BF_A) + 0 + b_3(OS_B - OS_A)$$

7th Class Part II

b_1's ~~ea the~~ have the same interpretation

Go to Table 4

Big reduction. Standard errors are larger

a cost of

Go back to Currie and Walker