Paper -- see guidelines

- attempt at a correct regression analysis
- start with a question or hypothesis
  sharper the better
  doesn't have to be supported, just well tested

- any data -- in class data we find
  I will work with you on this

- structure -- walk through concept
7th class Part I

Diff-in-diff

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

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

Recall the basic idea of natural experiments

NOT: Dataset \( \rightarrow \) what to can I do with this dataset?

BUT: Event that creates exogenous variation \( \rightarrow \) Assemble data

measure/policy of interest to do the test

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

NOT: Assemble data on, say, PM5 and LBW and controls

LBW = a + b, PM5 + controls

became: why do poor live in clipt areas with different PM5

PM5 \( \rightarrow \) LBW

omitted variables bias
She uses the natural experiment of the roll-out of EZPass on the NJ and PA turnpikes.

lots of toll booths. Originally everyone stopped, then EZ Pass -- no slowing down. \( \Rightarrow \) huge reduction in emissions.

If emissions matter, then you should observe reduction in birth defects for people living closest to the toll booths.

So you want to examine birth outcomes:

1. Before and after the change to EZ Pass
2. For women living very close vs. a bit farther away from the tollbooth

\(<2 \text{ km vs } 2-5 \text{ km}\)
take all births < 5 km of a full birth
and run $BW = \alpha + b_1 E_2 \text{pass} + b_2 < 2 \text{km} + b_3 E_2 \text{pass} + b_4 \text{km}$

so: $b_1$, $b_2$ the main effect of $E_2$ pass
$b_2$ is the main effect if $< 2 \text{km}$ relative to 2-5
$b_3$ is the extra effect of $E_2$ pass if $< 2 \text{km}$ or
(we refer to the $E_2$ 2-5 eqs)

Go over table 3 $\Delta BW \text{ corr } = -0.0093$
base rate (see Table 1) is about 10% (.10)
so reduce is about $8\%$ $1 \text{ ppt } = -10\%$

Actual model is more sophisticated because it
includes multiple births to the same mother,
so add another different

3. For many births to the same women.

Fixed effects
Our other example: Katrina -- 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 where?

- All other kids in US? in LA?

No -- then kids in nearby school district with similar pre-Katrina test scores.

Assess data on

match test score how non-No

pre 2007 post

pre 2007 post
Test score = \( a + b_1 2004 + b_2 2005 + \cdots + b_n 2014 \)

+ \( c_1 \) while NO + \( d_1 \) NO x 2004 + \( d_2 \) NO x 2005

\( \cdots + d_n 2014 \)

Expect that \( d_1, d_2 \) will be close to zero

but \( d_1 \)'s for 2009-2014 used would be

certain 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.
Contrary to our look at using regression to estimate caused relationships.

**Fixed effects** --- adding dummy variables for all "units" in the sample.

- e.g. all schools in the sample
- "family" in sample
- "site" in multi-site experiment.

"Ecometric" fixed effects as opposed to HLM "fixed effects," which are fixed (vs. random coefficients).

**Two broad uses**

1. Show where the units are likely variance decomposit.

2. Powerful control for omitted variable bias.

   (Powerful because it controls for both measurable and unmeasurable sources of bias)

**Hence 101:** (Ecometric) fixed effects adjust for a powerful and underappreciated technique for reducing bias.

   (recommended in developmental studies)
Show more in text is Fryer at least Table 2 Fall K Math
Black - .663 (.025)
Hispanic - .738 (.024)

Constant .307 (.013) 

R² .11 ≤ 11% of variance in test score is accounted for with 4 dummy variables

Highly significant in a statistical sense but too small 90% of the variance is within races across race/ethnic groups.

But now look at Table 7: Does school quality explain Black Student’s harbor in grade? 

Math

Reading Math - .243 -.180 
              (.052)   (.061)

Reading  - .343 - .214
              (.057)   (.065)

include school fixed effect? No Yes
-0.243 vs. -0.180 => 3/4 of the growth in the gap occurs within schools rather than between schools.

Most of the action is not caused by the fact that black and white students attend different schools.

Not always in case
Michelle had lead in water.
Across part had substantially higher lead in school water for black students.
Within or across the schools.
With or across class decades within schools.
Fixed effect as a way of reducing uncontrolled variable bias.

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

NCATSB: prenatal testing scale at 9 months
extensor Baby Motor Scale at 2 years
Bellevue Baby Scale at 2 years

Suppose points on extend at 9 months with all of Group 1.

Parent teaching High

Brain development not with.
Bucky = a + b, NCATS + with Feb 1

For each unit allows no setup, intercept with total

Note: no charts & NCATS continue Feb 1

Stokes are intended to be the same with all filters.

If you have 1,000 tiles, you don't want 1,000 interlace terms.

Cassie's results for NCATS

US results by Birthright
7th Class Part II

Fixed effect models

1. Add dummy variable for every unit
2. Use `xtreg` and specify unit
3. Transform all variable so they lie on deviations from the unit mean.

Why does this work.

\[ \text{PLAT_{math}} = a + b_1 \text{Breast Fed} + b_2 \frac{\text{Fixed}}{\text{chance}} + b_3 \text{other stuff} \]

Suppose sibling data, 2 sibs A: B

\[ \text{PLAT}_A = a + (b_1) \text{BF}_A + b_2 \text{Fedy}_A + b_3 \text{OS}_A \]
\[ \text{PLAT}_B = a + (b_1) \text{BF}_B + b_2 \text{Fedy}_B + b_3 \text{OS}_B \]

But `Fedy_A` is the same as `Fedy_B`.

\[ \text{PLAT}_B - \text{PLAT}_A = a - (b_1) (\text{BF}_B - \text{BF}_A) + 0 \]
\[ + b_3 (\text{OS}_B - \text{OS}_A) \]
b's have the same intensity.

Go to Table 4

Big reduction. Should run an longer
a cost $4

Go back to Camp at Walker.