Lecture 23: Difference-in-Differences

Pierre Biscaye

Fall 2022

Agenda

1. Policy analysis with panel data
2. Difference-in-Differences estimation
3. Identification assumptions
4. Indonesia school construction example (Duflo 2001)

Policy analysis using panel data

- With panel data, we've replaced one MLR 4 with another.
- Now, instead of needing *levels* of x variables to be uncorrelated with u we need *changes* in x variables to be uncorrelated with u.
  - With long panels, we need strict exogeneity: $cov(x_{jit}, u_{is}) = 0 \text{ for all } t, s, \text{and} j.$
- This will be less attractive if we don't know *why* x and y are changing.
  - Harder to argue in this case nothing else is changing simultaneously.
  - Could be something else that is leading both x and y to change.
- One compelling case where we do know why x is changing: policy analysis.
  - When x is a policy that takes effect (or direct result of that policy), we know why there was a (big) change in x.

Policy analysis using panel data

- Suppose we have
  1. data *before* and *after* a new policy takes effect;
  2. and *not everyone* will benefit from the policy;
  3. *and* we can identify who will benefit from the policy (the "treatment" group) at baseline, i.e., before it takes effect.
- Then we can estimate

$$y_{it} = \beta_0 + \delta post_t + \alpha Treatment_i + \beta_1 (Treatment_i) * (post_t) + u_{it} \quad (1)$$

- Note that if we do a first differenced or fixed effects estimate, $Treatment_i$ will disappear.
- But, we can estimate this with a repeated cross-section as well.
- What do we call this?

Difference-in-Differences estimation

$$y_{it} = \beta_0 + \delta post_t + \alpha Treatment_i + \beta_1 (Treatment_i) * (post_t) + u_{it} \quad (2)$$

- This specification is called a "Difference-in-Differences" specification.
- Why? Because the treatment effect (β_1) is the difference between the treatment and control groups in the difference over time (between pre- and post-policy) in the outcome variable.
- Interpretations:
  1. β_0 is the baseline level of y for the control group.
  2. α is the mean difference in y between treatment and control people at baseline.
  3. δ is the change in y over time for the control group.
  4. β_1 is the difference in the change in y over time for the treatment group relative to the control group.

Difference-in-Differences

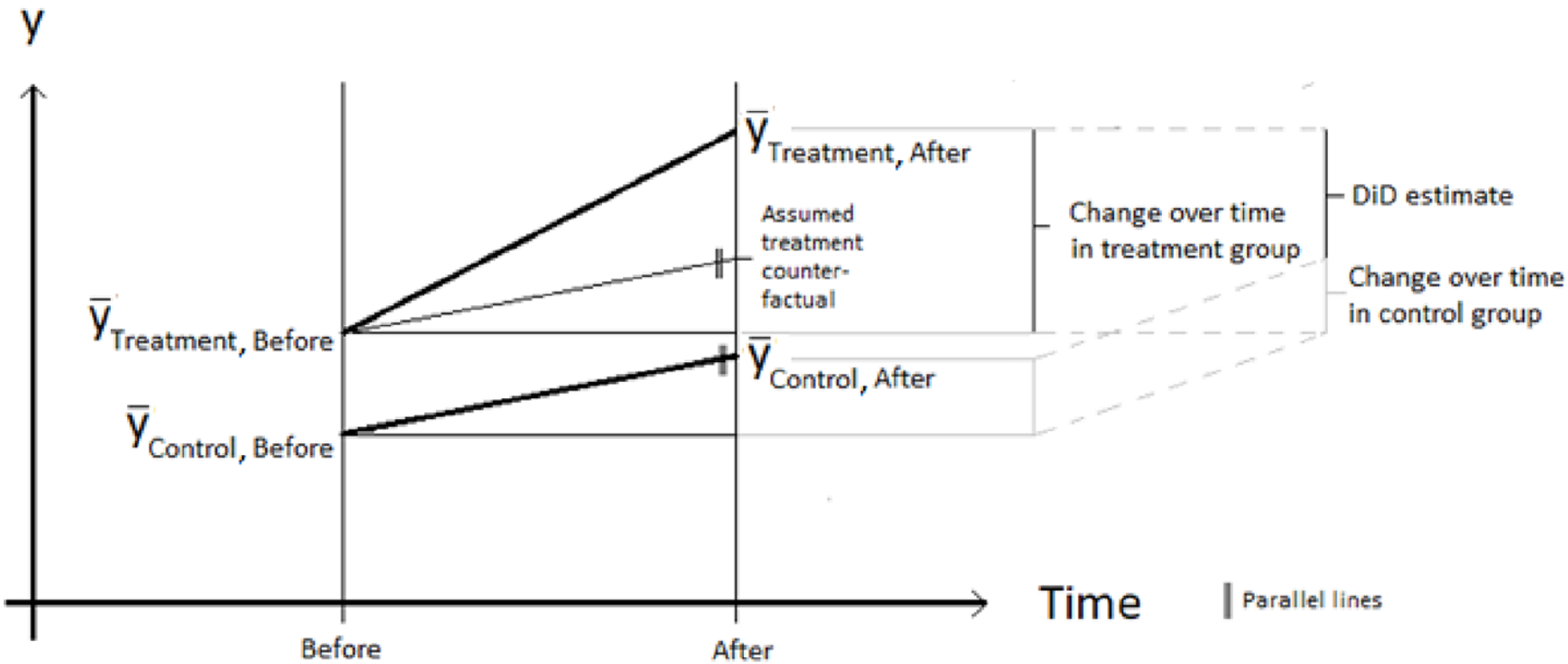
$$y_{it} = \beta_0 + \delta post_t + \alpha Treatment_i + \beta_1 (Treatment_i) * (post_t) + u_{it} \quad (3)$$

- Breaking down how β_1 represents the difference in the mean changes over time for the two groups:

$$\bar{y}_{Treatment,Post} = \beta_0 + \delta + \alpha + \beta_1 \quad (4)$$
$$\bar{y}_{Treatment,Pre} = \beta_0 + \alpha \quad (5)$$
$$\bar{y}_{Control,Post} = \beta_0 + \delta \quad (6)$$
$$\bar{y}_{Control,Pre} = \beta_0 \quad (7)$$
$$\beta_1 = \Delta\bar{y}_{Treatment} - \Delta\bar{y}_{Control} \quad (8)$$

DD, graphically

- $y_{it} = \beta_0 + \delta post_t + \alpha Treatment_i + \beta_1 (Treatment_i) * (post_t) + u_{it}$
- Where do we identify these parameters on the graph?



Source: Fredriksson, A. and Oliveira, G.M.d. (2019), "Impact evaluation using Difference-in-Differences"

DD in Words

- In other words, a Difference-in-Differences estimator asks how outcomes changed for exposed (treatment) people on average in a way that they did not change for unexposed (control).
- To do this, we will need some exposed, and some not exposed people.
  - We can't use a difference-in-differences approach to evaluate policies that impact *everyone*.
- We also need data before and after.
  - Otherwise we cannot separate baseline differences between the groups from the effect of the treatment.
- Identification assumption: The *only* thing that influenced y that changed differentially for the treatment group was the policy.
- In MLR4 terms: $E[\Delta u | Treatment, Post, Treatment * Post] = 0$
  - Refer to this as the *parallel trends* assumption.
  - If not, β_1 captures everything that changed differentially between the two groups over time, not just the treatment effect.

Example: traffic safety laws

- Two (now widespread) laws to encourage traffic safety:
  1. Open container laws: no open alcoholic containers in car cabins.
  2. Administrative *per se* laws: courts can suspend driver's licences at arrest.
- These were adopted by states in the 1980s.
- Did the adoption of these policies reduce traffic fatalities?

DD analysis of traffic safety laws

- We have data on states over time (1985 and 1990) on what traffic safety laws exist and the traffic deaths per 100 million miles driven.

$$dthrte_{it} = \beta_0 + \delta post_t + \alpha LawChange_i \quad (9)$$
$$+\beta_1 LawChange_i * post_t + u_{it} \quad (10)$$

- *LawChange* is a dummy taking a value of 1 for states that changed a traffic safety law over this time period. This is the "treatment".
- We know that $\beta_1 = \Delta dthrte_{LawChange_i=1} - \Delta dthrte_{LawChange_i=0}$
- To Jupyter!

What identification assumptions do we need?

- Recall MLR4 in a Difference-in-Differences analysis:
  - The *only* thing that changes death rates in states adopting administrative per se laws that doesn't change death rates in non-adopting states is the per se law.
- Is this likely to be true in this setting?

What identification assumptions do we need?

- First, what if pe se-adopting states adopted other laws too?
  - E.g., open container laws, speed limits.
- As before, we can always control for things we know about:

$$\begin{aligned} dthrte_{it} = \beta_0 + \delta post_t + \alpha_1 PerSeChange_i \\ + \beta_1 PerSeChange_i * post_t + \alpha_2 OpenContChange_i \\ + \beta_2 OpenContChange_i * post_t + u_{it} \end{aligned}$$

- Another approach, if you know there are some other simultaneous changes that could be correlated with treatment, would be to limit your sample to states that did not have other legal changes.
- To Jupyter!

What about other different trends?

- We still probably can't guess everything that might be different over time between states that adopt administrative per se laws and those that do not.
- A first test: see if there is anything obvious that is different about them at baseline.
  - Same Region?
  - Obviously different economically?
  - High Fatalities in 1985? Should this mean-revert if so?
- Can approach this the same way as we would testing for baseline balance with an RCT.
- Note that imbalances are not necessarily a problem, if they reflect time-invariant differences between the treatment and control groups.
- But large baseline differences across characteristics could also suggest that these or other factors may be likely to change differently over time across the groups.

MLR4 is still around

- Our Difference-in-Differences framework allows us to control for anything that does not change over time that is different between adopters and non-adopters (captured by the α parameter).
- But we do not control for things that change over time.
  - Similar to other panel data methods.
- Just as before with MLR4: we *cannot* test whether the assumption holds.
- Best case: provide supporting evidence.
  - One potentially particularly useful thing to provide: were the trends different *before* 1985?
  - If have data on time periods before the treatment and can show *parallel pre-trends* between treatment and control over time, this suggests that the MLR4 assumption of parallel trends over the treatment period could be plausible.

Another concern: treatment timing

- In the traffic laws dataset, some states already had open container or administrative per se laws in 1985.
- These are considered "control" states in the analysis, since they are not experiencing the treatment of changing laws between 1985-1990.
- But these may not be a good counterfactual for states that change their laws.
  - The true counterfactual for states that change their law is what would have happened in those states if they had not changed their law: best approximated by states that also did not have the law at baseline and did not enact it by the endline.
  - The trend over time for states that already had this law is very unlikely to be the same.
- Methods for DD analysis with different treatment timing is the subject of a growing econometric literature.
  - Beyond the scope of this class
  - Just be aware that any cases involving comparing not yet treated groups to already treated groups may lead to biased estimates.

Example 2: School construction in Indonesia

- This discussion based on (Duflo 2001)
- Between 1973-74 and 78-79, Indonesia constructed 61,807 primary schools.
  - Approx. 2 for every 1000 kids aged 5-14.
- Primary school enrollment grew from 69% in 1971 to 83% in 1978.
- Clearly an enormous change: did the school contruction *cause* the increase in enrollment?

Evaluating the school construction program

1. Did building schools bring in more kids?
2. Did those kids make more money as adults?

- Problem: schools were not built in random places.
- According to Policy Documents, Gov't built schools in places with lots of kids, and low enrollment rates.
- Duflo (2001) employs a Difference-in-Differences approach to analyze the program.

Difference-in-Differences in school construction

- Consider two groups:
  1. Districts where many new schools were constructed
  2. Districts where few new schools were constructed
- This gives one difference. We still need differences over time.
- One advantage of schooling: can use kids of different ages.
  - In 1974, some kids were already too old to go to primary school.
  - Others were young enough to be fully exposed to new schools.
  - Suggests using birth years to define pre- and post- periods.

2 differences

- We worry that districts that receive a lot of schools are different from districts who receive few.
- We also worry that kids born in later cohorts might be different from kids born in earlier cohorts.
- But maybe the trends in education among younger and older cohorts are similar in districts which receive many very few new schools.
- Let S be years of schooling, H represent high-construction districts, and L represent low construction districts

$$S_{id} = \beta_0^s + \delta^s [0-6]_i + \alpha^s High_d + \beta_1^s [0-6]_i * High_d + u_i \quad (11)$$

$$\beta_1^s = (\bar{S_H}^{0-6} - \bar{S_H}^{12-17}) - (\bar{S_L}^{0-6} - \bar{S_L}^{12-17}) \quad (12)$$

Mean differences

TABLE 3 -- MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in Region of birth			Level of program in Region of birth		
	High	Low	Difference	High	Low	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	-1.27	6.61	6.73	-0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	-1.39	6.87	7.02	-0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	-0.26	-0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
	8.00	9.41	-1.41	6.87	7.02	-0.15
Aged 12 to 17 in 1974	(0.054)	(0.042)	(0.078)	(0.0085)	(0.0069)	(0.011)
	7.70	9.12	-1.42	6.92	7.08	-0.16
Aged 18 to 24 in 1974	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
	0.30	0.29	0.013	0.056	0.063	0.0070
Difference	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Note: The sample is made of the individuals who earn a wage. Standard errors are in parentheses

Regression Coefficients

		Dependent variable					
		Years of education			Log(hourly wage)		
	Observations	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: Experiment of Interest: Individuals Aged 2 to 6 or 12 to 17 in 1974 (Youngest Cohort: Individuals Ages 2 to 6 in 1974)							
Whole sample	78,470	0.124	0.15	0.188			
		(0.0250)	(0.0260)	(0.0289)			
Sample of wage earners	31,061	0.196	0.199	0.259	0.0147	0.0172	0.0270
		(0.0424)	(0.0429)	(0.0499)	(0.00729)	(0.00737)	(0.00850)
PANEL B: Control Experiment : Individuals Aged 12 to 24 in 1974 (Youngest Cohort: Individuals Ages 12 to 17 in 1974)							
Whole sample	78,488	0.0093	0.0176	0.0075			
		(0.0260)	(0.0271)	(0.0297)			
Sample of wage earners	30,225	0.012	0.024	0.079	0.0031	0.00399	0.0144
		(0.0474)	(0.0481)	(0.0555)	(0.00798)	(0.00809)	(0.00915)
Control variables:							
Year of birth*enrollment rate in 1971		No	Yes	Yes	No	Yes	Yes
Year of birth* water and sanitation program		No	No	Yes	No	No	Yes

Notes: All specifications include region of birth, year of birth dummies and interactions between the year of birth dummies and the number of children in the region of birth (in 1971). The numbers of observations refer to the specification in columns 1 and 4.

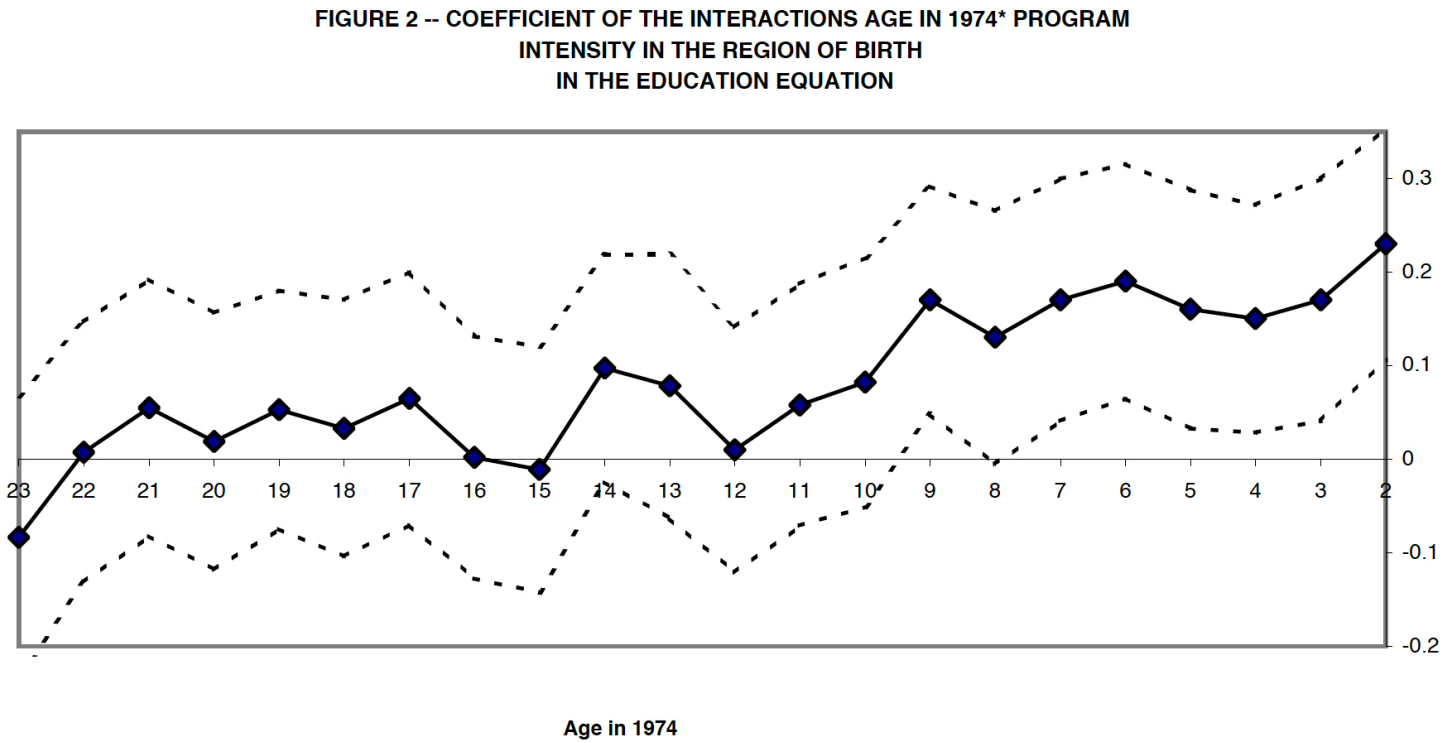
Standard errors are in parentheses.

Flexible age effects

- Of course, in a MLR we know we can have as many possible age effects as we want (as long as we have a large enough sample).
- No reason that we have to have just $[0-6]$ and $[12-17]$: we can measure the partially exposed as well.
  - And check whether the "fully exposed" 0 year olds look similar to the "fully exposed" 4 year olds.

$$S_{id} = \beta_0^s + \sum_{k=2}^{23} (\delta_k^s age = k) + \alpha^s * High_d + \sum_{k=2}^{23} (\beta_k^s (age = k) * (High_d)) + u_{id} \quad (13)$$

Flexible age effects



Wage Effects

- Once we have a strategy for MLR4, we can look at as many outcomes as we like.
- How about wages in adulthood?
- Can estimate ITT or ToT (not every child in high-construction regions was induced to increase education).

$$ITT : w_{id} = \beta_0 + \delta[0-6]_i + \alpha High_d + \beta_1^w [0-6]_i (High_d) + u_i \quad (14)$$

$$ToT : \frac{\beta_1^w}{\beta_1^s} \quad (15)$$

- Who are the compliers?

Mean differences

TABLE 3 -- MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in Region of birth			Level of program in Region of birth		
	High	Low	Difference	High	Low	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	-1.27	6.61	6.73	-0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	-1.39	6.87	7.02	-0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	-0.26	-0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
	8.00	9.41	-1.41	6.87	7.02	-0.15
Aged 12 to 17 in 1974	(0.054)	(0.042)	(0.078)	(0.0085)	(0.0069)	(0.011)
	7.70	9.12	-1.42	6.92	7.08	-0.16
Aged 18 to 24 in 1974	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
	0.30	0.29	0.013	0.056	0.063	0.0070
Difference	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Note: The sample is made of the individuals who earn a wage. Standard errors are in parentheses

Regression Coefficients

		Dependent variable					
		Years of education			Log(hourly wage)		
	Observations	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: Experiment of Interest: Individuals Aged 2 to 6 or 12 to 17 in 1974 (Youngest Cohort: Individuals Ages 2 to 6 in 1974)							
Whole sample	78,470	0.124	0.15	0.188			
		(0.0250)	(0.0260)	(0.0289)			
Sample of wage earners	31,061	0.196	0.199	0.259	0.0147	0.0172	0.0270
		(0.0424)	(0.0429)	(0.0499)	(0.00729)	(0.00737)	(0.00850)
PANEL B: Control Experiment : Individuals Aged 12 to 24 in 1974 (Youngest Cohort: Individuals Ages 12 to 17 in 1974)							
Whole sample	78,488	0.0093	0.0176	0.0075			
		(0.0260)	(0.0271)	(0.0297)			
Sample of wage earners	30,225	0.012	0.024	0.079	0.0031	0.00399	0.0144
		(0.0474)	(0.0481)	(0.0555)	(0.00798)	(0.00809)	(0.00915)
Control variables:							
Year of birth*enrollment rate in 1971		No	Yes	Yes	No	Yes	Yes
Year of birth* water and sanitation program		No	No	Yes	No	No	Yes

Notes: All specifications include region of birth, year of birth dummies and interactions between the year of birth dummies and the number of children in the region of birth (in 1971). The numbers of observations refer to the specification in columns 1 and 4.

Standard errors are in parentheses.

What do we think about MLR 4 here?

- Why are some places high regions and not others?
- If we don't know, we might be worried about many things.
- Possibilities?

Official explanation

- According to policy documents:

$$New\ Schools_d = \lambda Num\ Kids_d * disenrollment\ rate_d \tag{16}$$

$$log(New\ Schools_d) = log(\lambda) + log(Num\ Kids_d) + \log(disenrollment\ rate_d) \tag{17}$$

- How to test whether this rule was followed?

OLS

$$log(New\ Schools_d) = log(\lambda) + log(Num\ Kids_d) + \log(disenrollment\ rate_d) \quad (18)$$

$$log(New\ Schools_d) = \beta_0 + \beta_1 log(Num\ Kids_d) + \beta_2 log(disenrollment\ rate_d) + u_d \quad (19)$$

- What hypotheses can we test to explore whether the gov't followed its rule?

Testing school allocation

TABLE 2 -- THE ALLOCATION OF SCHOOLS

	log(INPRES schools)
Log of number of children aged 5-14 in the region	0.78
	(0.027)
Log(1-enrollment rate in primary school in 1973)	0.12
	(0.038)
Number of observations	255
R squared	0.78

Notes: Standard errors are in parentheses.
The dependent variable in the log of the number of INPRES schools built between 1973 and 1978.
The enrollment rate in primary school is the number of children enrolled in primary school in 1973 (obtained from the Ministry of education and Culture) divided by the number of children aged 5-14 in the region in 1973

Wrapping up Indonesia Schools

- It looks like the government *partially* followed the listed rule.
- We may worry about deviations from that rule.
  - What if they built schools in places where the returns to education were higher?
  - What if they built schools in places where the returns to education were growing faster?
  - What if they build schools in the home districts of influential politicians?
- Diff-in-Diff will rule out some omitted variables.
  - But will still need to explore other potential trends to be fully persuaded.