Lecture 7 – Natural Experiments

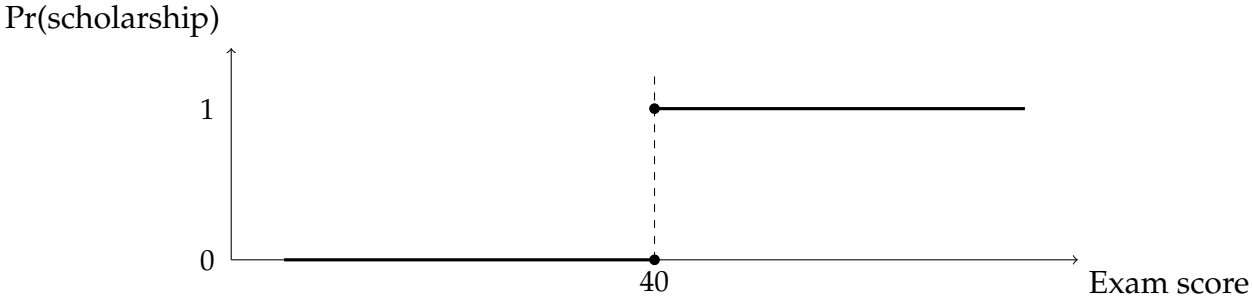
Sometimes it is impossible to randomize something (e.g. school funding). But we have alternatives that get us most the way there.

Regression Discontinuity

Often programs are assigned on the basis of a score on a scale. The Pell Grant depends on Expected Family Contribution, state merit scholarships depend on a test score, class size caps depend on enrollment, and so on.

It may be surprising that these rules can generate "as if random" variation. The basic intuition is that if people can't pick their score on the scale, which side of the cutoff they fall on is close to random.

Say a state offers a merit scholarship to any student scoring 40 or higher on a statewide exam, and nothing below 40.



Scholarship eligibility jumps from 0 to 1 at the cutoff.

Will a student who scored 39 be different from a student who scored 40? Can they precisely manipulate their score? In expectation the only difference between the two is that one got the scholarship and one did not. For students near the line, a cutoff rule generates something close to random assignment.

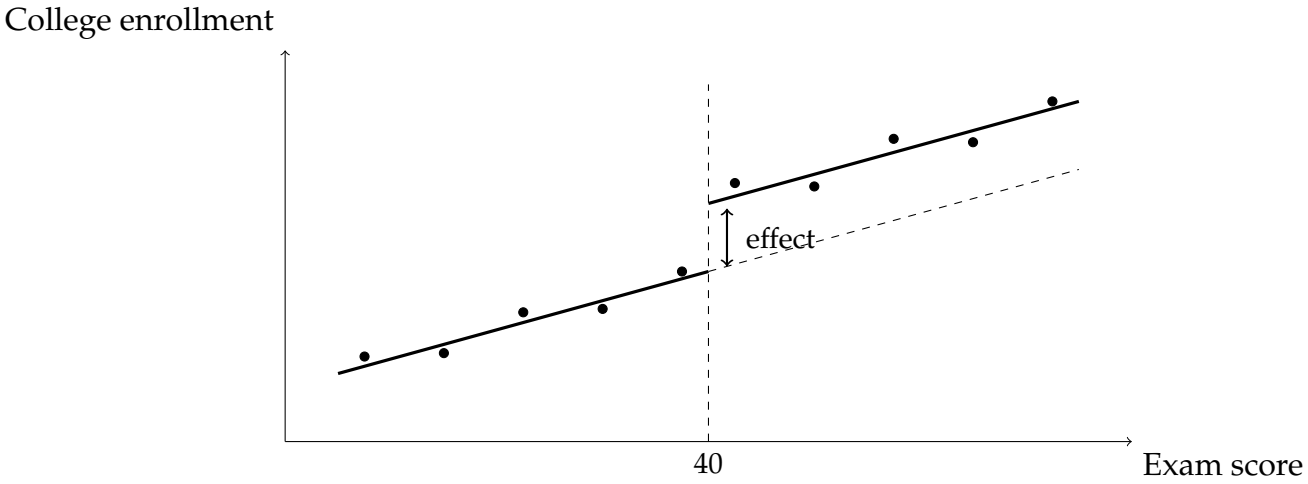
What we look for

Assumptions

You can think of the assumption as it being as good as random which side of the cutoff you're on.

When would that be false? It would be false if something else also switches at 40. Suppose the state uses the same score to flag students for a different intervention. Then you are measuring a bundle and cannot separate the pieces, which is the same problem as a bundled RCT.

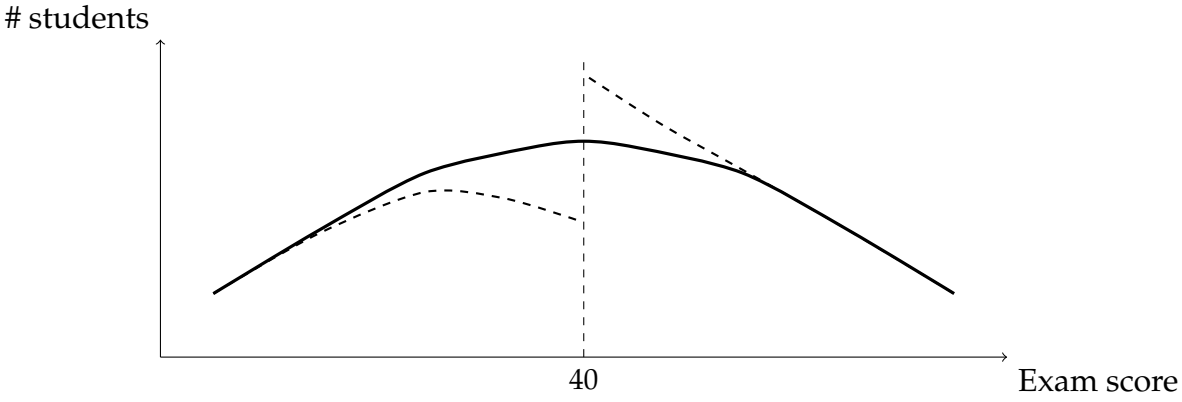
You can think of two testable implications of this assumption:



Each dot is the average outcome for students at that score. The dashed line is what we think would have happened without the jump.

1. People cannot precisely manipulate their score. If they can, we are back to selection bias because people can decide to be treated.
2. Students just below and just above the cutoff look similar on predetermined covariates.

To test the first, you check whether there is bunching in the density of the score. Could a teacher grading a borderline exam nudge a student from 39 to 40 so they get the scholarship? If that is happening, you will see students missing from just below the cutoff and piled up just above.



Solid: no manipulation. Dashed: a hole just below the cutoff and a pile-up just above.

To test the second, you compare students on either side of the cutoff using characteristics measured beforehand, such as prior scores, demographics, and attendance.

The intuition is the same as in an RCT, where we check for balance on observables and assume balance on unobservables.

Suppose students just above 40 turn out to have higher 4th grade scores than students just below. What does that tell you? Is it a problem you can fix by controlling for 4th grade scores?

If you run these tests and they fail, you cannot interpret the RD as estimating a causal effect.

Sharp versus fuzzy

A sharp RD is when the score is the only thing that determines treatment, so everyone at 40 or higher gets the scholarship. A fuzzy RD is when the rule only shifts the odds, so students above 40 are eligible but only 60% ever claim the money.

Fuzzy RD is the compliance problem from RCTs. You use clearing the cutoff as an instrument for actually receiving the scholarship and divide by the first stage. Many of the terms are the same as they were in IV because IV sits on top of RD.

Differences in Differences

The basic intuition for differences in differences is that there are two groups that would follow the same trend absent treatment. Then treatment occurs and one of the groups moves relative to that trend.

Suppose a state raises teacher pay in 2024 and we want to know what it did to teacher turnover. The table below shows the turnover rate in that state and in a neighboring state that did not act.

	2023	2025
Pay-raise state	11	13
Comparison state	14	21

One estimate would be to just compare turnover before and after in the pay-raise state, or $13 - 11 = 2$. Turnover went up. What would be the problem with this comparison?

Other things changed between 2023 and 2025, namely inflation that eroded teacher salaries in both states, so the change in turnover reflects the pay raise and everything else that happened in those two years.

The approach in differences in differences is to find a comparison group that was not treated. If we assume the control group is a good counterfactual, we can get an estimate. Here the comparison state went from 14 to 21, or $21 - 14 = 7$.

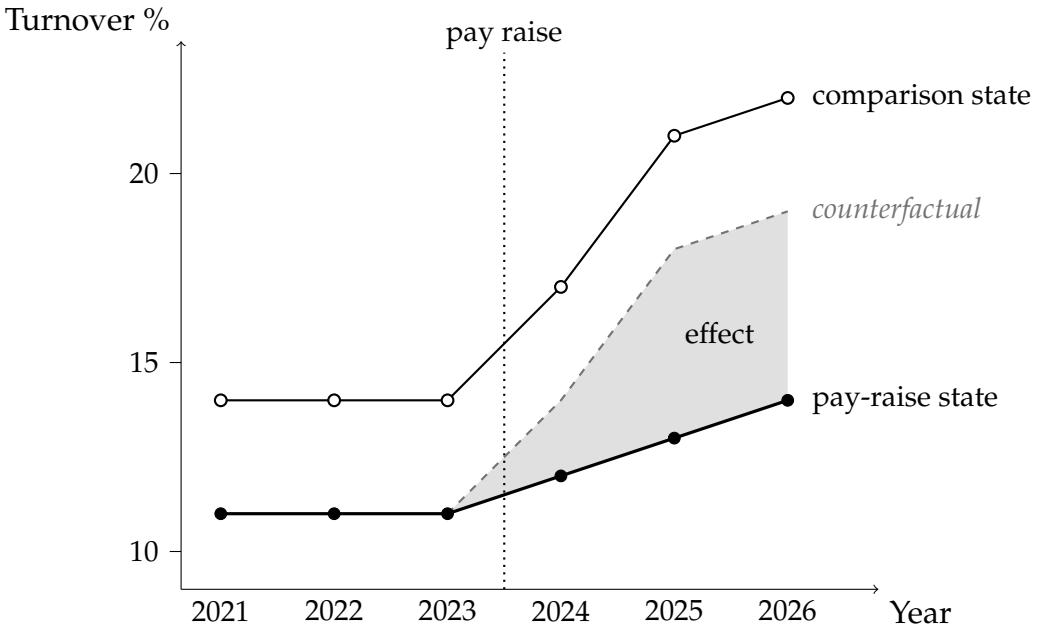
We get the effect of the raise by taking the difference of these two differences, $2 - 7 = -5$. Turnover in the pay-raise state rose 5 points less than it would have otherwise.

Notice that the simple before and after comparison did not just get the magnitude wrong, it got the sign wrong.

Notice also that the two states do not have the same level of turnover, and that is fine. Differences in differences requires the same trend, not the same level.

Parallel trends

We call the assumption here parallel trends. The treatment group would have followed the same trend as the control group if there had not been treatment.



Before the raise, turnover is flat in both states and sits 3 points apart. The dashed line applies the comparison state's change to the pay-raise state, so it stays 3 points below the comparison state throughout. The shaded wedge between it and what actually happened is the estimate, 5 points.

How strong is parallel trends? Pretty strong, because you never observe what the trend would have been without treatment.

What supporting evidence can you provide? You can show that the two states actually did follow the same path before 2024, which is why every DD paper shows a figure like the one above. Plotting the gap year by year is called an event study, and in the pre-period you want those differences to be flat.

Another thing to worry about is other things changing at the same time. If the state also cut class sizes in 2024, could you interpret the estimate as the effect of the pay raise? (No) This is often described as requiring no other contemporaneous shocks, and the instinct is the same as an exclusion restriction.

Another thing to worry about is whether treatment is endogenous, meaning it happens in places that are already different in some unobserved way. If the raise was a response to a turnover crisis that was already building, parallel trends was in trouble before you started.

Regression version

We can implement this via regressoin

Wrapping Up

Five strategies, and each time the question is the same. Where is the variation in treatment coming from, and is it contaminated?

Method	**Variation used**	**What you have to believe**
RCT	A coin flip	The randomization was run properly
Controls	Whatever is left after holding measured things fixed	Nothing *unmeasured* drives both treatment and outcome
IV	A lever that shifts treatment	The lever touches the outcome no other way
RD	Landing just above vs. just below a cutoff	Nobody places themselves precisely, and nothing else changes at the cutoff
DD	A policy change in one group and not another	The groups would have trended together

RD uses cutoff rules. Check for manipulation, check balance at the cutoff, and zoom in close.

DD uses a policy change plus a comparison group. Check pre-trends, rule out other shocks, and ask why the policy passed where it did.

In both cases the estimate is a jump, either at a cutoff or at a date, and everything else in the picture is background.

When you read a study, you should be able to identify the assumption.