Econometrics

TA Session 9

Lucia Sauer

2025-11-27

Overview

- Difference-in-Differences (DiD) recap
- Assumptions for DiD
- Application DiD

Difference-in-Differences (DiD) recap

- DiD is a quasi-experimental design that exploit **variation in time** (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.
- The key idea is to control for **unobserved confounders** that are constant over time and common trends affecting both groups.
- Data Requirements: We need data from periods before and after treatment to use DiD (and some periods where no unit is treated).

The DiD Estimator

- The DiD estimator is calculated as:

$$\text{DiD} = (\bar{Y}_{T,post} - \bar{Y}_{T,pre}) - (\bar{Y}_{C,post} - \bar{Y}_{C,pre})$$

  where:

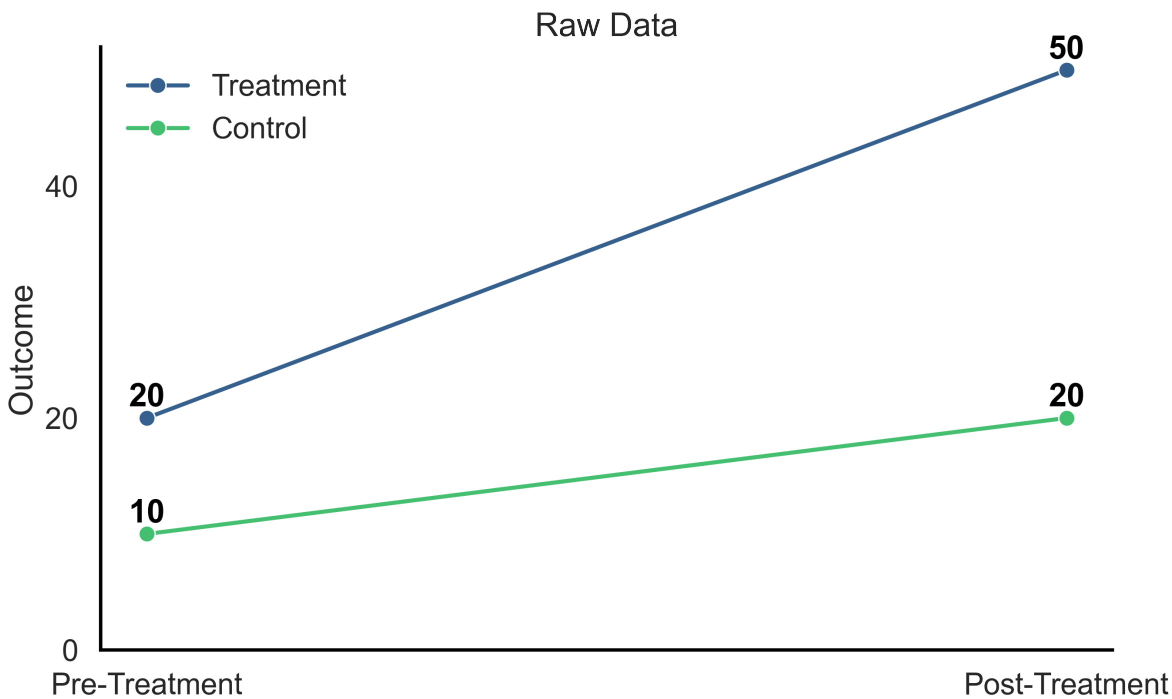
  - $\bar{Y}_{T,post}$: sample mean outcome of the treatment group after treatment

- $\bar{Y}_{T,pre}$: sample mean outcome of the treatment group before treatment
- $\bar{Y}_{C,post}$: sample mean outcome of the control group after treatment
- $\bar{Y}_{C,pre}$: sample mean outcome of the control group before treatment

---

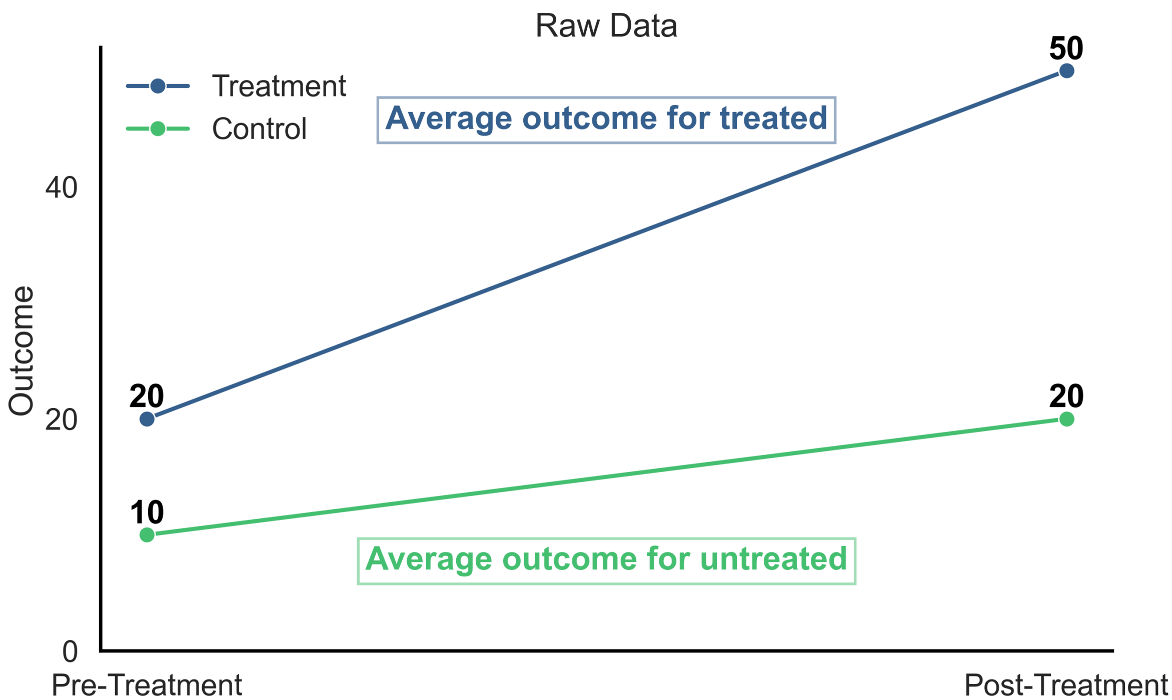
Assumptions for DiD

Parallel Trends Assumption: In the absence of treatment, the average change in outcomes for the treatment group would have been the same as that for the control group.



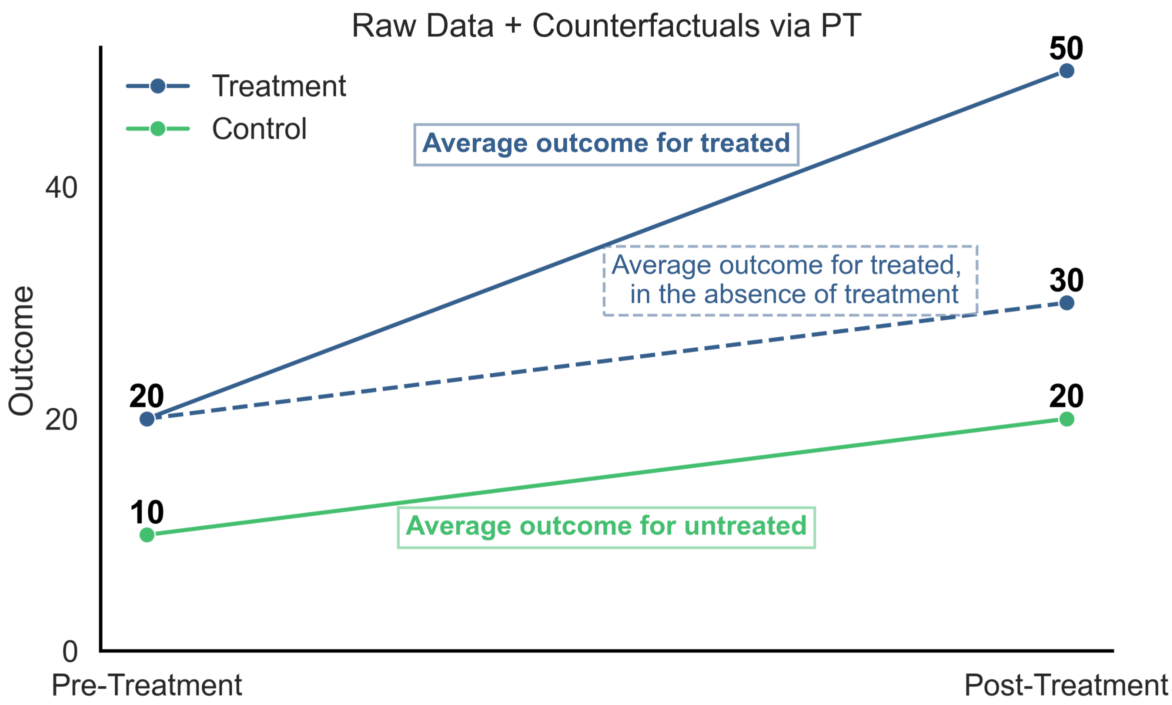
Assumptions for DiD

Parallel Trends Assumption: In the absence of treatment, the average change in outcomes for the treatment group would have been the same as that for the control group.



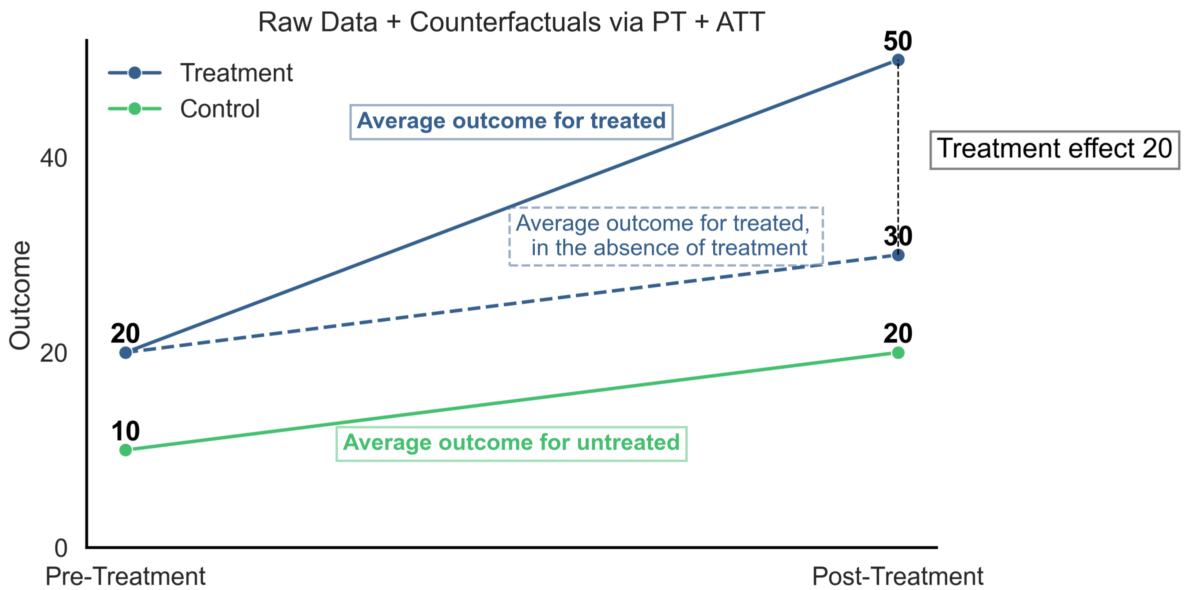
Assumptions for DiD

Parallel Trends Assumption: In the absence of treatment, the average change in outcomes for the treatment group would have been the same as that for the control group.



Assumptions for DiD

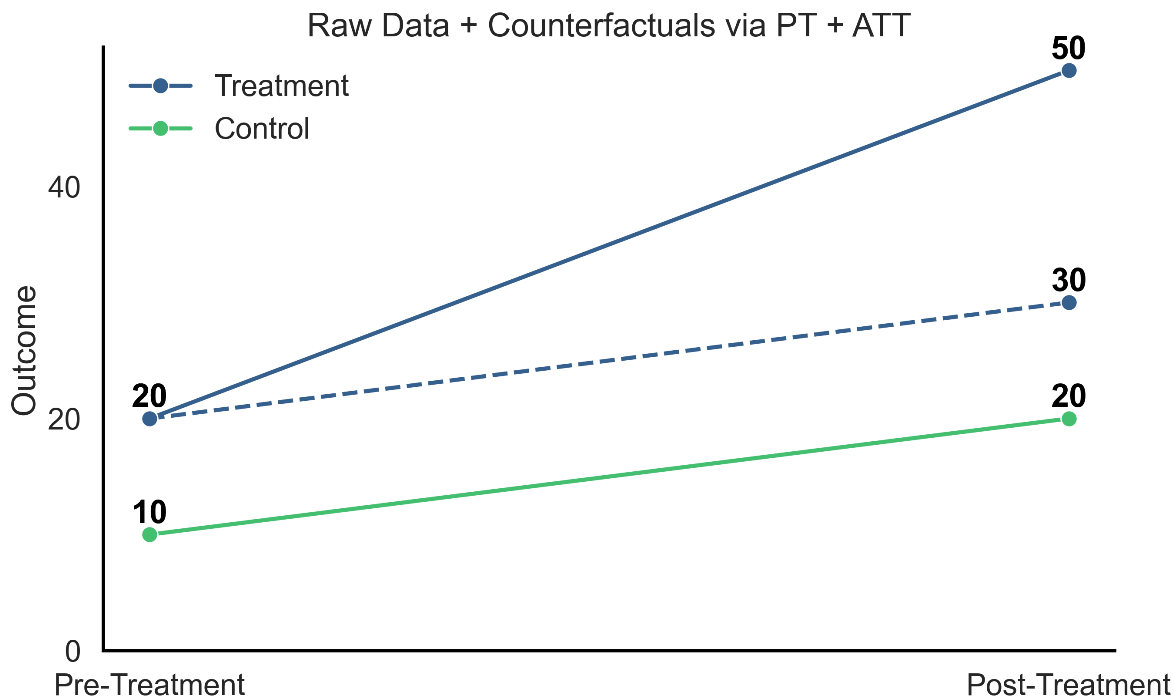
Parallel Trends Assumption: In the absence of treatment, the average change in outcomes for the treatment group would have been the same as that for the control group.



DiD and Regression

The DiD model can be specified as:

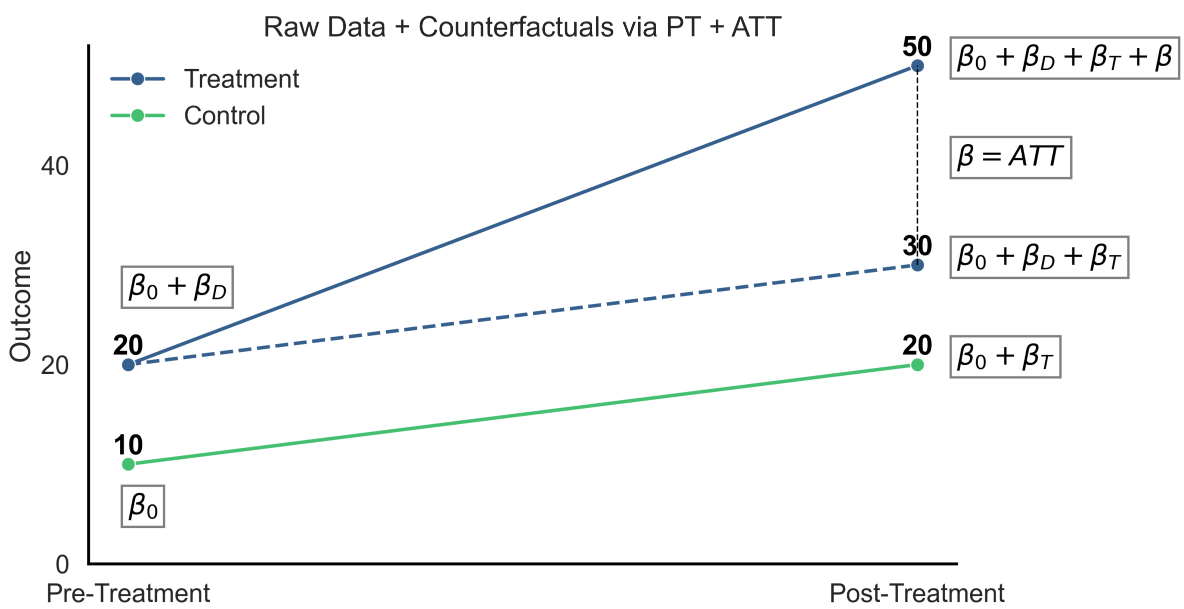
$$Y_{it} = \beta_0 + \beta_D \mathrm{D}_i + \beta_T \mathrm{Post}_t + \beta(\mathrm{D}_i \times \mathrm{Post}_t) + \epsilon_{it}$$



DiD and Regression

The DiD model can be specified as:

$$Y_{it} = \beta_0 + \beta_D \mathrm{D}_i + \beta_T \mathrm{Post}_t + \beta(\mathrm{D}_i \times \mathrm{Post}_t) + \epsilon_{it}$$



DiD and Regression

$$\beta = (\bar{Y}_{T,post} - \bar{Y}_{T,pre}) - (\bar{Y}_{C,post} - \bar{Y}_{C,pre})$$

Rearranging the terms, we get:

$$\beta = (\bar{Y}_{T,post} - \bar{Y}_{C,post}) - (\bar{Y}_{T,pre} - \bar{Y}_{C,pre})$$

Which can be equivalently put as:

- $(\bar{Y}_{T,post} - \bar{Y}_{C,post})$ = treatment effect + selection bias
- $(\bar{Y}_{T,pre} - \bar{Y}_{C,pre})$ = selection bias
- β = treatment effect

Application DiD

What is the effect of tutoring on students' GPA?

```python
data = {
    "Student": [1,2,3,4,5,6,7,8,9,10,
                1,2,3,4,5,6,7,8,9,10],
    "Time": [0,0,0,0,0,0,0,0,0,0,
             1,1,1,1,1,1,1,1,1,1],
    "GPA": [2.7,2.6,2.9,3.0,2.8,2.8,3.0,3.2,3.1,3.5,
            2.75,2.6,3.0,2.9,3.1,2.9,3.3,3.3,3.2,3.8],
    "Tutoring": [0,0,0,0,0,1,1,1,1,1,
                 0,0,0,0,0,1,1,1,1,1]
}

df = pd.DataFrame(data)
df.sample(10)
```

	Student	Time	GPA	Tutoring
10	1	1	2.75	0
6	7	0	3.00	1
16	7	1	3.30	1
7	8	0	3.20	1
5	6	0	2.80	1
0	1	0	2.70	0
9	10	0	3.50	1
8	9	0	3.10	1
4	5	0	2.80	0
15	6	1	2.90	1

- We have data on two groups of students: those who received tutoring (treatment group) and those who did not (control group).
- We observe their GPA before and after the tutoring program was implemented.

$$\beta = (G\bar{P}A_{tut,post} - G\bar{P}A_{tut,pre}) - (G\bar{P}A_{control,post} - G\bar{P}A_{control,pre})$$

```python
gpa_pre_tutoring = df[(df['Time'] == 0) & (df['Tutoring'] == 1)]['GPA'].mean()
gpa_post_tutoring = df[(df['Time'] == 1) & (df['Tutoring'] == 1)]['GPA'].mean()
gpa_pre_no_tutoring = df[(df['Time'] == 0) & (df['Tutoring'] == 0)]['GPA'].mean()
gpa_post_no_tutoring = df[(df['Time'] == 1) & (df['Tutoring'] == 0)]['GPA'].mean()

diff_pre_treatment = gpa_pre_tutoring - gpa_pre_no_tutoring
```

```
diff_post_treatment = gpa_post_tutoring - gpa_post_no_tutoring

did_estimator = diff_post_treatment - diff_pre_treatment
round(did_estimator, 3)
```

```
np.float64(0.11)
```

The same DiD estimator can be obtained by estimating the following regression model:

$$GPA_{it} = \beta_0 + \beta_{\text{tutoring}}\text{tutoring}_i + \beta_T\text{Post}_t + \beta(\text{tutoring}_i \times \text{Post}_t) + \epsilon_{it}$$

where:

- GPA_{it} is the GPA of student i at time t.
- tutoring_i is a binary variable indicating whether student i received tutoring (1 if yes, 0 if no).
- Post_t is a binary variable indicating the time period (1 for post-treatment, 0 for pre-treatment).
- β is the DiD estimator, capturing the effect of tutoring on GPA.

We can also transform the outcome variable to represent the change in GPA for each student over time, and then regress this change on the treatment indicator.

$$\Delta GPA_i = GPA_{i,post} - GPA_{i,pre}$$

The model then becomes:

$$\Delta GPA_i = \alpha + \beta\text{tutoring}_i + \epsilon_i$$

Organ Donor Registration and Policy Interventions

In the US there are different politics to enourage organ donations. When people sign up for a driver's license, they can choose between:

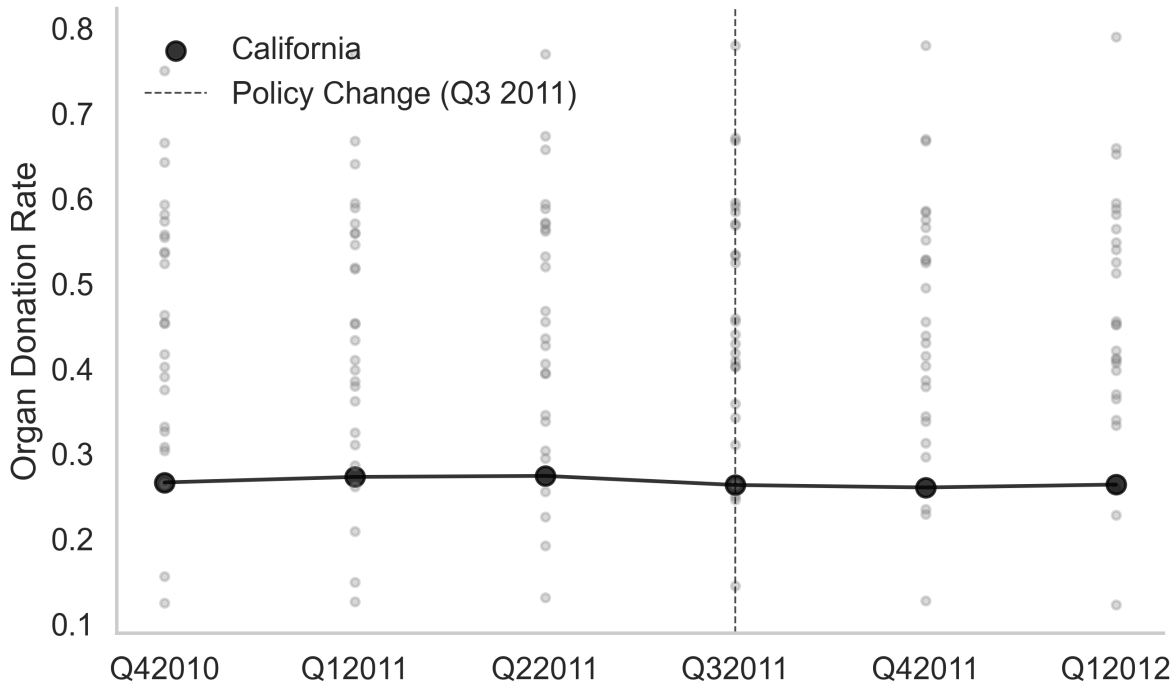
- *Opt-in*: the default is not to donate; individuals must actively agree.
- *Active choice*: individuals are required to make an explicit yes/no decision
- *Opt-out*: individuals are automatically registered unless they decline.

In **July 2011**, the state of California shifted from a traditional **opt-in** system to an **active-choice** approach.

> Kessler, J. B., & Roth, A. E. (2014). *Don't Take "No" for an Answer: An Experiment with Actual Organ Donor Registrations.* National Bureau of Economic Research.

Organ Donation Rates Over Time

- California already doesn't have a great organ donation rate, sitting near the bottom of the pack.
- California's rate didn't rise much after the policy went into effect - in fact, it seems to have dropped slightly.



Two-Way Fixed Effects Model

- We have more than two groups and two time periods.
- The goal here is to control for **group** differences, and also control for **time** differences.

$$rate_{it} = \alpha_i + \delta_t + \beta(\text{California}_i \times \text{Post}_t) + \epsilon_{it}$$

where:

- $rate_{it}$ is the organ donation rate in state i at time t.
- α_i captures state-specific fixed effects, controlling for time-invariant differences between states.
- δ_t captures time-specific fixed effects, controlling for common shocks affecting all states at time t.
- California_i is a binary variable indicating whether state i is California.
- Post_t is a binary variable indicating whether time t is after the policy change.
- β is the DiD estimator, capturing the effect of the active-choice policy on organ donation rates.

Estimation Results

```python
from linearmodels import PanelOLS
#| echo: true
od["Post"] = od["Quarter_Num"] > 3
od["California"] = od["State"] == "California"
od["Treated"] = (od["Post"] & od["California"]).astype(int)

# Set panel index
od = od.set_index(["State", "Quarter_Num"])

# DID model with entity and time fixed effects
model = PanelOLS.from_formula(
    "Rate ~ Treated + EntityEffects + TimeEffects",
    data=od
)

results = model.fit(cov_type="clustered", cluster_entity=True)
print(results)
```

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                   Rate   R-squared:                        0.0092
Estimator:                   PanelOLS   R-squared (Between):             -0.0010
No. Observations:                 162   R-squared (Within):              -0.0021
Date:                Wed, Nov 26 2025   R-squared (Overall):             -0.0010
Time:                        22:15:02   Log-likelihood                    388.57
Cov. Estimator:             Clustered
                                        F-statistic:                      1.2006
Entities:                          27   P-value                           0.2752
Avg Obs:                       6.0000   Distribution:                   F(1,129)
Min Obs:                       6.0000
Max Obs:                       6.0000   F-statistic (robust):             11.525
                                        P-value                           0.0009
Time periods:                       6   Distribution:                   F(1,129)
Avg Obs:                       27.000
Min Obs:                       27.000
Max Obs:                       27.000
```

```
                              Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
Treated       -0.0225     0.0066    -3.3949     0.0009     -0.0355     -0.0094
==============================================================================

F-test for Poolability: 191.71
P-value: 0.0000
Distribution: F(31,129)

Included effects: Entity, Time
```

Event Study Analysis

- An event study allows us to visualize the dynamics of the treatment effect over time.
- Each estimated coefficient represents the effect of the treatment at different time points relative to the policy change.

$$rate_{it} = \alpha_i + \delta_t + \sum_{k \neq -1} \beta_k D_{i,t+k} + \epsilon_{it}$$

where:

- $D_{i,t+k}$ is a binary variable that equals 1 if state i is in period k relative to the treatment time (with $k = -1$ as the omitted category).
- β_k captures the effect of the treatment at time k relative to the treatment time.

Eventy Study Results