

ECON 594: Applied Economics

Panel Data and Difference-in-Differences

Sam Norris

University of British Columbia

Where we are

The arc of the week

- Yesterday: how to use AI
- Today: panel data and difference-in-differences
- Next class: event studies (the dynamic version of DD)

These are review lectures

- I assume you've seen this before
- Focus is on the practical decisions you'll make in your thesis

Why panel data?

Running example: minimum legal drinking age (MLDA) and mortality

- Some US states had MLDA = 18 in the 1970s, others MLDA = 21
- Y_{st} : deaths per 100,000 among 18–20 year olds in state s , year t

Cross-sectional regression:

$$Y_{st} = \beta \cdot \mathbb{1}[\text{MLDA} < 21]_{st} + \varepsilon_{st}$$

OLS requires $E[\varepsilon_{st} \mid \text{MLDA}_{st}] = 0$

- Drinking culture, road quality, demographics all live in ε
- And they are almost certainly correlated with state policy

Panel data lets us split: $\varepsilon_{st} = \alpha_s + \nu_{st}$

- Just control for α_s !
- α_s absorbs anything fixed about state s
- Now only need ν_{st} uncorrelated with policy after netting out the state

Types of data

Cross-section: many units at one point in time

Repeated cross-section: new random sample each period

- E.g. American Community Survey, Canadian Census

Panel / longitudinal: same units across many periods

- E.g. PSID, NLSY, administrative data with stable identifiers

Quasi-panel: repeated cross-section collapsed to group means

- E.g. state-by-year averages from the Current Population Survey
- Treat the group \times year as the panel “unit”

Notation

Two subscripts: unit (i or s) and time (t)

- Y_{st} = outcome for state s in year t

Variables come in three flavours:

- Unit-only: X_s (e.g. state area in km^2)
- Time-only: X_t (e.g. federal funds rate)
- Both: X_{st} (e.g. state-level policy in year t)

Useful exercise: for every variable in your data, ask which subscripts it can have

- Determines what variation it carries
- Determines what fixed effects can absorb it

The fixed effects model

Two ways to write the same regression:

With a state-specific intercept α_s :

$$Y_{st} = \beta X_{st} + \alpha_s + v_{st}$$

With a dummy for every state but one:

$$Y_{st} = \alpha + \beta X_{st} + \sum_{k=2}^N \gamma_k \mathbb{1}[\mathbf{s} = \mathbf{k}] + v_{st}$$

These produce the same $\hat{\beta}$

- The α_s notation is just shorthand for the dummy regression
- “Fixed effect on state s ” = “dummy variable for state s ”

The within transformation

Take state-level means of the regression:

$$\bar{Y}_s = \beta \bar{X}_s + \alpha_s + \bar{v}_s$$

Subtract from the original:

$$Y_{st} - \bar{Y}_s = \beta(X_{st} - \bar{X}_s) + (v_{st} - \bar{v}_s)$$

α_s has dropped out. OLS on the demeaned data gives the same $\hat{\beta}$ again

- Nice way to think about what an FE regression actually does
- This is what `reghdfe` does internally with millions of units

What you cannot identify

Anything that doesn't vary within a state gets demeaned to zero:

$$Y_{st} - \bar{Y}_s = \beta_1(X_{1,st} - \bar{X}_{1,s}) + \beta_2 \underbrace{(X_{2,s} - X_{2,s})}_{=0} + \dots$$

β_2 on a time-invariant regressor is not identified once you include state FE

- Examples: state area, distance to coast, founding year

If you care about the effect of something fixed within state, FE can't help you

- Often a sign you need a different research design

Stata: `reghdfe`

`reghdfe` (high-dimensional FE) is the workhorse

- Absorbs FE without estimating them, scales to millions of units

```
ssc install reghdfe
```

```
* One-way FE
```

```
reghdfe y x, absorb(state) cluster(state)
```

```
* Two-way FE
```

```
reghdfe y x, absorb(state year) cluster(state)
```

```
* Three-way FE (firm, worker, year)
```

```
reghdfe y x, absorb(firm worker year) cluster(firm)
```

Alternatives

- `areg` works for one set of FE
- `xtreg, fe` works but is slower for large panels

The simplest case: 2×2 DD

Alabama lowered its drinking age from 21 to 19 in 1975

- Neighbouring Arkansas kept its drinking age at 21
- Two groups, two periods:

	Pre (1974)	Post (1976)
Alabama (treated)	$Y_{T,Pre}$	$Y_{T,Post}$
Arkansas (control)	$Y_{C,Pre}$	$Y_{C,Post}$

Goal: estimate the effect of lowering the drinking age

The 2×2 calculation

Two obvious estimators, neither of them good:

- Cross-section in Post: $Y_{T,Post} - Y_{C,Post}$
Confounded by fixed differences between Alabama and Arkansas
- Time-series in Alabama: $Y_{T,Post} - Y_{T,Pre}$
Confounded by other things changing over time

DD combines them:

$$\delta_{DD} = (Y_{T,Post} - Y_{T,Pre}) - (Y_{C,Post} - Y_{C,Pre})$$

- Differences out fixed unit differences, and differences out common time shocks

Potential outcomes notation

For each unit at each time, two potential outcomes:

- $Y_{it}(1)$ if the unit is treated
- $Y_{it}(0)$ if the unit is not treated

We observe one of them, not both:

$$Y_{it} = D_{it} \cdot Y_{it}(1) + (1 - D_{it}) \cdot Y_{it}(0)$$

Want to estimate the ATT (average treatment effect on the treated):

$$ATT = E[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1]$$

$Y_{it}(0)$ for treated units in Post is the missing counterfactual

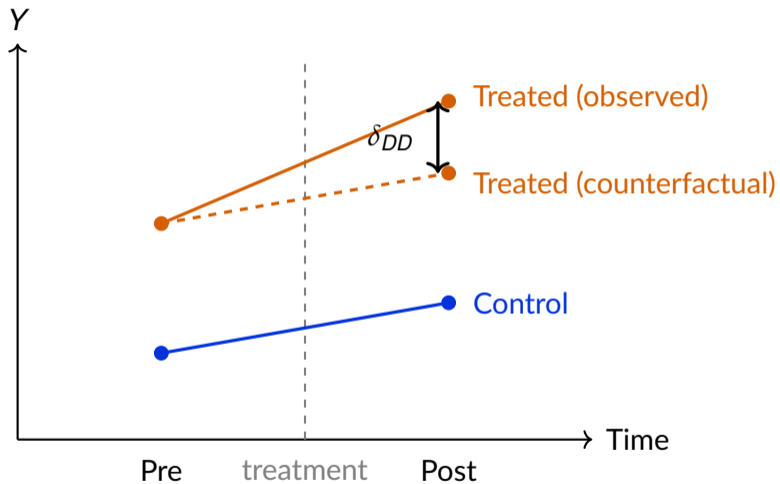
- Need an assumption about how it would have evolved

The diff-in-diff assumption

$$\begin{aligned}\delta_{DD} &= \underbrace{(Y_{T,Post} - Y_{T,Pre})}_{\text{Pre-to-post change in T}} - \underbrace{(Y_{C,Post} - Y_{C,Pre})}_{\text{Pre-to-post change in C}} \\ &= (Y_{T,Post}(1) - Y_{T,Pre}(0)) - (Y_{C,Post}(0) - Y_{C,Pre}(0)) \\ &= \underbrace{Y_{T,Post}(1) - Y_{T,Post}(0)}_{\text{ATT}} \\ &\quad + \underbrace{Y_{T,Post}(0) - Y_{T,Pre}(0) - (Y_{C,Post}(0) - Y_{C,Pre}(0))}_{= 0 \text{ under parallel trends}}\end{aligned}$$

Parallel trends: if treatment hadn't happened, outcomes in T and C would have moved by the same amount between Pre and Post.

Parallel trends, visualised



Parallel trends \Leftrightarrow the dashed line is the right counterfactual.

The 2×2 as a regression

With one treated and one control unit and two periods, run:

$$Y_{st} = \alpha + \beta \cdot \mathbb{1}[\text{Treat}_s] + \gamma \cdot \mathbb{1}[\text{Post}_t] + \delta \cdot \mathbb{1}[\text{Treat}_s] \times \mathbb{1}[\text{Post}_t] + e_{st}$$

Take expectations in each cell:

$$E[Y \mid C, Pre] = \alpha$$

$$E[Y \mid T, Pre] = \alpha + \beta$$

$$E[Y \mid C, Post] = \alpha + \gamma$$

$$E[Y \mid T, Post] = \alpha + \beta + \gamma + \delta$$

$\hat{\delta}$ on the interaction equals the DD calculation

- Bonus: regression gives standard errors for free (more on those later)

We need more states, but...

Let's add one more state: Tennessee

- Dropped its MLDA to 18 in 1971
- Raised it to 19 in 1979

Adding Tennessee breaks the 2×2 setup:

- More than one treated state, treated at different times
- No single "Post" period anymore
- No single binary treatment status

Need a regression that handles all this

Step 1: replace the interaction with a policy indicator

With Alabama vs. Arkansas alone, $Treat_s \times Post_t = \mathbb{1}[\text{MLDA} < 21]_{st}$

Replace the interaction with the policy indicator directly:

$$Y_{st} = \alpha + \beta \cdot \mathbb{1}[\text{Treat}_s] + \gamma \cdot \mathbb{1}[\text{Post}_t] + \delta \cdot \mathbb{1}[\text{MLDA} < 21]_{st} + e_{st}$$

- Same coefficient as before
- But now generalisable: the policy indicator can switch on at different times for different states, and switch off again

Step 2: many time periods, many treated units

Replace $Post_t$ with a full set of year dummies, $Treat_s$ with a full set of state dummies

$$Y_{st} = \alpha + \sum_{\ell \neq 1} \beta_{\ell} \mathbb{1}[\text{state} = \ell] + \sum_{y \neq 1970} \gamma_y \mathbb{1}[\text{year} = y] + \delta \cdot \mathbb{1}[\text{MLDA} < 21]_{st} + e_{st}$$

Cleaner with fixed-effects notation:

$$Y_{st} = \alpha_s + \gamma_t + \delta \cdot \mathbb{1}[\text{MLDA} < 21]_{st} + e_{st}$$

- This is the two-way fixed effects (TWFE) regression
- α_s absorbs anything constant about state s
- γ_t absorbs anything common to all states in year t
- δ identified from within-state changes in the policy

What TWFE does, intuitively

There are no fixed treated/control states anymore

- Each state is its own control during years it doesn't change policy
- Each state is treated during years immediately after it changes policy
- Ditto for years

Pools many small “natural experiments” into one estimate

Identifying assumption is still parallel trends:

- Absent the policy change, treated and untreated state-years would have moved similarly

Stata for TWFE

* Set up the panel

```
xtset state year
```

* TWFE with reghdfe (preferred)

```
reghdfe mrate under21, absorb(state year) cluster(state)
```

* Same thing with areg (slower for many FE)

```
areg mrate under21 i.year, absorb(state) cluster(state)
```

* Same thing with raw dummies (don't do this with millions of units)

```
reg mrate under21 i.state i.year, cluster(state)
```

All three give the same point estimate

- Always cluster (next section)

Why cluster?

Default OLS standard errors assume residuals are independent across observations

In panels, this is almost never true:

- Mortality rates in Alabama in 1975 and 1976 are correlated
- Anything making 1975 a bad year persists into 1976

If you ignore this, your standard errors are too small and t-stats are too big

- Particularly bad with policy variables that vary at the state level: treatment status doesn't change much within a state, so your effective N is much smaller than 51×14

Cluster at the unit level

Bertrand, Duflo, Mullainathan (QJE 2004): clustering at the state level fixes the problem

- Allows arbitrary serial correlation within state
- Assumes independence across states

Rule of thumb: need at least ≈ 30 clusters for the asymptotics to work

- With 50 US states you're fine
- With 10 Canadian provinces you are not. Use wild cluster bootstrap (`boottest`)

Cluster on the unit even when the unit FE isn't in the regression

- In Stata: `vce(cluster state)` or `cluster(state)`

Common mistakes

Clustering at the wrong level

- Treatment varies at the state level \Rightarrow cluster at state, not state-year

Clustering with too few clusters and trusting the SEs

- Wild bootstrap (`boottest`) is your friend

Reporting only robust (non-clustered) SEs in a panel setting

- Almost always wrong, and a referee will catch it

Recap

Fixed effects let us absorb anything fixed about a unit, weakening the OLS exogeneity assumption

Diff-in-diff compares pre-to-post changes in a treated group to pre-to-post changes in a control group

- Identification rests on parallel trends

Two-way fixed effects extends this to many units and many periods, pooling lots of small natural experiments

Always cluster your standard errors at the unit level

Coming up

Tuesday: Event study designs (no class Monday – Victoria Day)

- Dynamic version of DD: how does the effect evolve over time?
- How to test the parallel trends assumption (sort of)
- The recent literature on staggered DD and what to do about it

Wednesday: Tables and Figures