Regression with Panel Data

SW Chapter 10

Types of data

Difference-in-differences

Two-period panel data analysis

Fixed effects

Least squares assumptions

Learning objectives

- Understand difference between types of data
- Conduct difference-in-difference regressions
- Estimate first-difference regressions with panel models
- Estimate and interpret regressions with entity-level fixed effects, time-level fixed effects, and with both entity and time fixed effects
- Understand challenges to valid estimation under fixed effect models
- Use clustered standard errors to account for autocorrelation within panel data

Types of data

Types of data

- Cross-sectional: random (independent) sampling of units at one point in time
  - What we've been using so far!
  - Relationship between hours worked and wages
  - How CEO tenure relates to compensation
- Time-series: observations over time
  - Stock-market dividends for Apple over the past 10 years
  - GDP growth over time
  - Annual infant mortality rates

Types of data

- Panel (longitudinal): Cross-sections over time
    - Track how individuals' earnings change over time
    - Crime rates by city over time
    - Very useful for policy analysis!
    - Different from a "pooled cross-section" (multiple cross sections) – like lots of waves of the ACS – different units each time

Cross-sectional data

id[1] | 1

	id	wage	educ	exper	tenure	nonwhite	female	married	numdep	smsa	northcen
1	1	3.1	11	2	0	0	1	0	2	1	0
2	2	3.2	12	22	2	0	1	1	3	1	0
3	3	3	11	2	0	0	0	0	2	0	0
4	4	6	8	44	28	0	0	1	0	1	0
5	5	5.3	12	7	2	0	0	1	1	0	0
6	6	8.8	16	9	8	0	0	1	0	1	0
7	7	11	18	15	7	0	0	0	0	1	0
8	8	5	12	5	3	0	1	0	0	1	0
9	9	3.6	12	26	4	0	1	0	2	1	0
10	10	18	17	22	21	0	0	1	0	1	0
11	11	6.3	16	8	2	0	1	0	0	1	0
12	12	8.1	13	3	0	0	1	0	0	1	0
13	13	8.8	12	15	0	0	0	1	2	1	0
14	14	5.5	12	18	3	0	0	0	0	1	0
15	15	22	12	31	15	0	0	1	1	1	0
16	16	17	16	14	0	0	0	1	1	1	0
17	17	7.5	12	10	0	0	1	1	0	1	0
18	18	11	13	16	10	0	1	0	0	1	0
19	19	3.6	12	13	0	0	1	1	3	1	0
20	20	4.5	12	36	6	0	1	1	0	1	0
21	21	6.9	12	11	4	0	1	0	0	1	0
22	22	8.5	12	29	13	0	0	1	3	1	0
23	23	6.3	16	9	9	0	1	0	0	1	0
24	24	.53	12	3	1	0	1	0	0	1	0
25	25	6	11	37	8	1	1	0	0	1	0
26	26	9.6	16	3	3	1	0	1	1	1	0

Time-series data

	year	unem	inf	inf_1	unem_1	cinf	cunem
1	1948	3.8	8.1	.	.	.	.
2	1949	5.9	-1.2	8.1	3.8	-9.3	2.1
3	1950	5.3	1.3	-1.2	5.9	2.5	-.5999999
4	1951	3.3	7.9	1.3	5.3	6.6	-2
5	1952	3	1.9	7.9	3.3	-6	-.3
6	1953	2.9	.8	1.9	3	-1.1	-.0999999
7	1954	5.5	.7	.8	2.9	-.1	2.6
8	1955	4.4	-.4	.7	5.5	-1.1	-1.1
9	1956	4.1	1.5	-.4	4.4	1.9	-.3000002
10	1957	4.3	3.3	1.5	4.1	1.8	.2000003
11	1958	6.8	2.8	3.3	4.3	-.5	2.5
12	1959	5.5	.7	2.8	6.8	-2.1	-1.3
13	1960	5.5	1.7	.7	5.5	1	0
14	1961	6.7	1	1.7	5.5	-.7	1.2
15	1962	5.5	1	1	6.7	0	-1.2
16	1963	5.7	1.3	1	5.5	.3	.1999998
17	1964	5.2	1.3	1.3	5.7	0	-.5
18	1965	4.5	1.6	1.3	5.2	.3000001	-.6999998
19	1966	3.8	2.9	1.6	4.5	1.3	-.7
20	1967	3.8	3.1	2.9	3.8	.1999998	0
21	1968	3.6	4.2	3.1	3.8	1.1	-.2
22	1969	3.5	5.5	4.2	3.6	1.3	-.0999999
23	1970	4.9	5.7	5.5	3.5	.1999998	1.4
24	1971	5.9	4.4	5.7	4.9	-1.3	1
25	1972	5.6	3.2	4.4	5.9	-1.2	-.3000002
26	1973	4.9	6.2	3.2	5.6	3	-.6999998
27	1974	5.6	11	6.2	4.9	4.8	.6999998
28	1975	8.5	9.1	11	5.6	-1.9	2.9

Vars: 7 Obs: 56

Panel data

Data Editor (Browse) - PRISON.DTA

Edit Browse Filter Variables Properties Snapshots

state[1] 1

	state	year	govelec	black	metro	unem	criv	crip	lcriv	lcrip
1	1	80	0	.256	.632	.08775	4.447868	44.47638	1.492425	3.794958
2	1	81	0	.2557	.6362	.10667	4.700944	44.24879	1.547763	3.789828
3	1	82	1	.2554	.6404	.14367	4.49758	42.05045	1.503539	3.73887
4	1	83	0	.2551	.6446	.13667	4.186833	37.08439	1.431945	3.613196
5	1	84	0	.2548	.6488	.11167	4.353239	35.04226	1.47092	3.556555
6	1	85	0	.2545	.6530001	.08908	4.630758	35.2668	1.532721	3.562942
7	1	86	1	.2542	.6572	.09833	5.665331	37.8735	1.734365	3.634252
8	1	87	0	.2539	.6614	.07775	5.686924	39.58107	1.73817	3.678351
9	1	88	0	.2536	.6656	.07208	5.728628	41.05592	1.745476	3.714935
10	1	89	0	.2533	.6698	.07025	6.036973	41.25161	1.797903	3.71969
11	1	90	1	.253	.674	.06775	7.069136	41.96889	1.955738	3.736928
12	1	91	0	.2527	.6782	.072	8.439609	45.20342	2.132936	3.811173
13	1	92	0	.2524	.6824	.073	8.712421	43.94321	2.16475	3.782898
14	1	93	0	.2521	.6866	.075	7.804156	40.98352	2.054657	3.71317
15	2	80	0	.034	.434	.09592	4.773632	57.0398	1.563107	4.043749
16	2	81	0	.0347	.4317	.09233	6.069378	58.93301	1.803256	4.076401
17	2	82	1	.0354	.4294	.09983	6.071111	54.39778	1.803542	3.996323
18	2	83	0	.0361	.4271	.10308	6.02459	53.05123	1.795849	3.971258
19	2	84	0	.0368	.4248	.1015	6.046692	53.43969	1.799511	3.978554
20	2	85	0	.0375	.4225	.09608	5.697369	51.85714	1.740004	3.948493
21	2	86	1	.0382	.4202	.10908	5.599265	55.7114	1.722635	4.020185
22	2	87	0	.0389	.4179	.108	4.435993	47.94249	1.489751	3.870002
23	2	88	0	.0396	.4156	.09225	4.948339	41.63469	1.599052	3.728934
24	2	89	0	.0403	.4133	.06742	4.795247	41.25594	1.567625	3.719795
25	2	90	1	.041	.411	.07025	5.207581	45.95126	1.650115	3.827581
26	2	91	0	.0417	.4087	.085	6.149385	50.96661	1.816352	3.931171
27	2	92	0	.0424	.4064	.091	6.593537	49.0068	1.88609	3.891959
28	2	93	0	.0431	.4041	.076	7.607679	48.07178	2.029158	3.872695

Vars: 45 Obs: 714

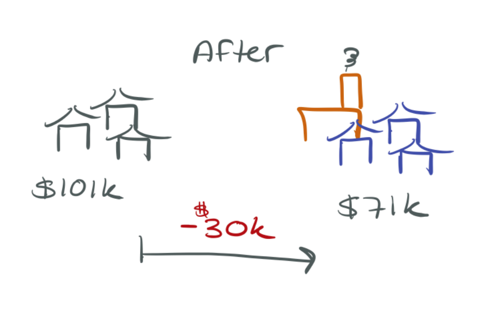
Panel vs. repeated cross-sections

- Repeated (or pooled) cross-sections:
  - Draw randomly from large population at various points in time
  - Observations are independent over time: Bob surveyed in 1990 is independent of Jane in 1991
  - Independence $\Rightarrow$ inference is pretty easy!
- Panel data
  - Track the same population at various points in time
  - Population may be independent at first draw: Bob in 1990 is independent of Jane in 1990
  - But dependence over time: Bob in 1990 is related to Bob in 1991
  - Dependence $\Rightarrow$ Detroit will probably have high crime rates next year.

Difference-in-differences

Example: effect of garbage incinerator location on housing prices

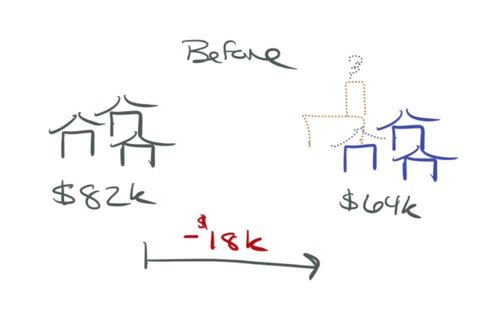
Consider effect of the location of a house on its price before and after the garbage incinerator was built:



$$\widehat{rprice} = 101307.5 - 30688.28 nearinc$$

Example: effect of garbage incinerator location on housing prices

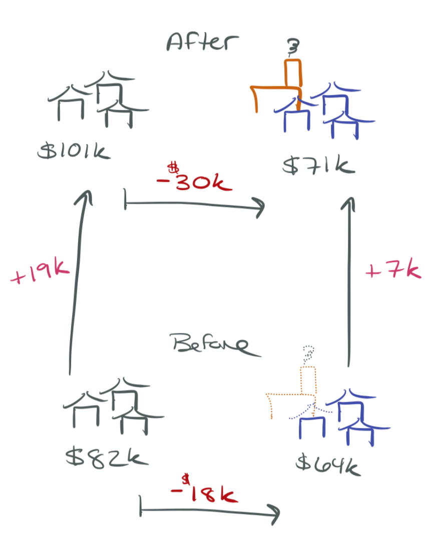
No! Look at the relationship between pricing and incinerator location *before* incinerator was built



Before incinerator was built:

$$\widehat{rprice} = 82517.23 - 18824.37 nearinc$$

Example: effect of garbage incinerator location on housing prices



Example: effect of garbage incinerator location on housing prices

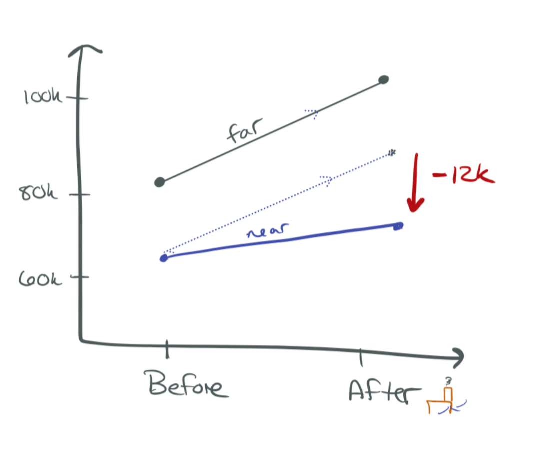
We should account for this:

$$\hat{\delta}_1 = -30688.27 - (-18824.37) = -11863.90$$

- Incinerator reduces prices by \$12k
- This is equivalent to the following:

$$\hat{\delta}_1 = (\overline{rprice}_{1,nr} - \overline{rprice}_{1,fr}) - (\overline{rprice}_{0,nr} - \overline{rprice}_{0,fr})$$

Example: effect of garbage incinerator location on housing prices



Difference-in-differences in a regression framework

We can capture this in a single regression:

$$rprice = \beta_0 + \delta_0 after + \beta_1 nearinc + \delta_1 afterXnearinc + u$$

- Easy to estimate both coefficients and standard errors
- If houses sold before and after incinerator was built were systematically different, want to control for those differences
- Doing this will reduce error variance and standard errors

Policy evaluation using difference-in-differences

We can apply this logic to evaluate the impact of policies for which we have a before and after period, and groups that were and were not affected:

$$y = \beta_0 + \delta_0 after + \beta_1 treated + \delta_1 afterXtreated + other + u$$

$$\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{1,C}) - (\bar{y}_{0,T} - \bar{y}_{0,C})$$

	Before	After	After - Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment - Control	β_1	$\beta_1 + \delta_1$	δ_1

Two-period panel data analysis

Two-period panel data analysis

- Now, assume we have the same observations over two time periods
- Consider relationship between unemployment rates and crime
  - Detroit has high unemployment rates and high crime
  - Does high unemployment cause high crime, or is there another explanation?
  - Explanatory variables could help

Effect of unemployment on city-level crime rates

- Assume that no other explanatory variables are available. Will it be possible to estimate the causal effect of unemployment on crime?
- Yes, if cities are observed for at least two periods and other factors affecting crime stay approximately constant over those periods.
- Consider a set of cities observed in two periods - 1982 and 1987:

$$crimerte_{it} = \beta_0 + \delta_0 d87_{it} + \beta_1 unemp_{it} + a_i + u_{it}$$
$$t = 1982, 1987$$

- $d87_{it}$: time dummy for second period
- a_i: unobserved time-constant factors (**fixed effects**)
- $u_{i,t}$: other unobserved factors (**idiosyncratic error**)

Effect of unemployment on city-level crime rates

$$crmrte_{i,1987} = \beta_0 + \delta_0 1 + \beta_1 unemp_{i,1987} + a_i + u_{i,1987}$$
$$crmrte_{i,1982} = \beta_0 + \delta_0 1 + \beta_1 unemp_{i,1982} + a_i + u_{i,1982}$$

$$\Delta crmrte_i = \delta_0 + \beta_1 Deltaunemp_i + \delta u_i$$

See how the fixed effect drops out!

You can estimate a first-differenced equation using OLS

$$\Delta \widehat{crmrte} = 15.4 + 2.22 \Delta unemp$$

A 1pp increase in unemployment rate $\Rightarrow$ 2.22 more times per 1,000 people

Discussion of first-difference panel estimator

- The first-differenced panel estimator lets us causal effects in the presence of time-invariant endogeneity
- Will not solve time-variant endogeneity!
- However, first-differenced estimates will be imprecise if explanatory variables vary only little over time (no estimate possible if time-invariant)

Fixed effects

Fixed effects - a wonder!

- What if, instead of differencing out individual-level characteristics? What if we control for them?
- For example, if we look at a panel of wages over time, there might be omitted Bob-specific characteristics:
    - Bob is hard working (+),
    - Bob is a competitive negotiator (+)
    - Bob smells funny (-)
    - We can never control for all of them.

Fixed effects - a wonder!

- We can never control for all of Bob's unique quirks and features.
- But, we can "control for" Bob.
- We will give Bob his very own **fixed effect**: the unique "stuff" that Bob brings to the table.
- Only covers stuff that stays constant over time.

Fixed effects estimation

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$$
$$\bar{y}_{it} = \beta_1 \bar{x}_{it} + \ldots + \beta_k \bar{x}_{ik} + \bar{a}_i + \bar{u}_i$$
$$[y_{it} - \bar{y}_i] = \beta_1 [x_{it1} - \bar{x}_{i1}] + \ldots + \beta_k [x_{itk} - \bar{x}_{ik}] + [u_{it} - \bar{u}_i]$$

Because $a_i - \bar{a}_i = 0$, the fixed effect drops out

- Estimate time-demeaned equations by OLS
- Uses time variation within cross-sectional units: **within estimation**
- Functionally equivalent to including a dummy variable for ever i (fixed effect)

Estimating fixed-effect models in practice

1. Hard way:
   - Demean your data by hand
   - Estimate OLS on demeaned data
2. Easier way:
   - Estimate OLS, including fixed effects by using dummy variables
   - The command `areg` will let you absorb, or "eat up," one set of fixed effects
3. Powerful way:
   - Tell Stata you have panel data using xtset
   - Estimate using `xtreg` (still OLS)

Least squares assumptions

Least squares assumptions for panel data

Consider a single X:

$$Y_t = \beta_1 X_{it} + \alpha_i + u_{it}, i = 1, ..., n; t = 1, ...T$$

1. $E[u_{it}|X_{i1}, ...X_{iT}, \alpha_i) = 0$
2. $(X_{i1}, ...X_T, u_{i1}, ..., u_{iT}), i = 1, ..., n$ are i.i.d. draws from their joint distribution
3. (X_{it}, u_{it}) have finite fourth moments
4. No perfect multicollinearity

- u_{it} cannot be correlated with any **present, past, or future** values of X!
- However, only have to be independent draws *across* entities

If these hold, estimates are consistent and normally distributed for large n!

Serial correlation

$$Y_{it} = \alpha_i + \beta_1 X_{it} + \omega_{it}$$

- Time-varying omitted variable is ω_{it}
- ω_{it} reflects factors changing over time that affect outcome
- If Y_{it} is traffic fatalities, maybe ω_{it} includes road repairs in state i
- Likely quality of roads last period in state i are similar to quality of roads in this period
- Roads yesterday are about the same as roads today

Autocorrelation

- Previously, we assumed that different observations were independent
  - No twins, no school districts in the same county
- Implausible for state data over time
- Data on the same entity over time is likely to suffer from autocorrelation
- Without correction, we will estimate incorrect standard errors, inference will be wrong

Autocorrelation: clustering

- Clustered standard errors correct for autocorrelation
- Otherwise, confidence intervals will not have 95
- Suppose entity is a U.S. state
  - Tell Stata to allow for omitted variables ω_{it} for different time periods from same state to be correlated
  - Add option `cluster(state)` to `regress` or add option `vce(cluster)` to `xtreg`

Miscellaneous notes

- Covariates: The fixed effect covers all time-invariant factors! (So you don't need them in your model)
- Interpreting fixed effects models:
  - We don't usually look at the coefficients on the fixed effects themselves, though they can be informative.
  - Which fixed effect is "omitted" - avoiding the dummy variable trap – doesn't really matter, because we're not interpreting them!
  - Not reported in regression estimates
- Changes interpretation of models
  - Only look at effect of variables that change over time

Summary

Interesting ways to work with panel data

Difference-in-differences specifications

- Repeated cross-sectional data at two points in time
- Often a “before” vs. “after” and a “treatment” vs. “control”
- GM assumptions apply for OLS to be BLUE
- **Assumptions we need:**
  - Parallel trends assumption: That in absence of “treatment,” gap between two groups would have stayed parallel
- **Assumptions we don’t need:**
  - There can be unobserved factors that happened at one point of time, if they affect the two groups equally
  - The two groups can be very different from each other, so long as they are on the same trajectories

Interesting ways to work with panel data

First-difference models

- Panel data
- Regress the change in Y on change in X
- **Assumptions we need:**
  - Still have to be careful of any omitted characteristics that vary over time
- **Assumptions we don't need:**
  - Any time-invariant characteristics are differenced out! $\Rightarrow$ no time-invariant omitted variable bias!

Interesting ways to work with panel data

Fixed effects

- Panel data
- Regress Y on X , include entity-specific and/or time-specific fixed effects
- **Assumptions we need:**
  - With entity-effects only: Still have to be careful of any omitted characteristics that vary over time
  - With time-effects only: Still have to be careful of any omitted characteristics that vary across entities at a particular period in time
  - With both: Careful of omitted characteristics that vary across entities *and* time
- **Assumptions we don't need:**
  - Any time-invariant characteristics are controlled! $\Rightarrow$ no time-invariant omitted variable bias!
  - Any entity-invariant characteristics are controlled! $\Rightarrow$ no omitted variable bias for aggregate time shocks!

Conclusion

Types of data

Difference-in-differences

Two-period panel data analysis

Fixed effects

Least squares assumptions