8. Time series, interrupted time series and difference-in-differences

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Regression discontinuity designs

- When a treatment is assigned using a precise rule based on a continuous characteristic (a "running variable"). Treatment can be "sharp" or "fuzzy."
- Under certain assumptions, the jump in mean outcomes at the cutpoint can be interpreted as the causal effect of the treatment.
- The aim is to use regression to estimate that jump. We looked at different ways of modeling and estimating this discontinuity.
- The RD assumptions suggest a number of validity checks:
  - Test for discontinuities in baseline covariates or "placebo" outcomes.
  - Test for manipulation in the running variable.

Tonight

- Different data structure: cross-section, time series, panel
- Time series models used for forecasting
- Interrupted time series research designs
- Difference-in-differences

Tonight's sample datasets

We will refer to three datasets tonight (on Github):

1. `State_school_finance_panel_1990_2010.dta`: state-level school finance and related data, 1990-2010.
2. `NYCbkfastlunch.dta`: school-level data from NYC on breakfast and lunch program participation.
3. `dynarski.dta`: data from Dynarski (2003) on Social Security survivor's benefits and college enrollment.

Types of data

Types of data

Most of the analyses we've seen thus far have been with **cross-sectional** data: typically, many units (i) at one point in time.

	state_name	year	expp_co2	personal_i~c
1	Alaska	2000	11482.09	30508
2	Alabama	2000	7331.784	24067
3	Arkansas	2000	6936.413	22574
4	Arizona	2000	6878.881	26293
5	California	2000	8705.413	33404
6	Colorado	2000	8182.125	33986
7	Connecticut	2000	12616.86	41920
8	District of Columbia	2000	15008.14	40462
9	Delaware	2000	11159.94	31009
0	Florida	2000	7687.295	29079
1	Georgia	2000	8632.021	28541
2	Hawaii	2000	8217.213	29024
3	Iowa	2000	8633.758	27285
4	Idaho	2000	7132.013	24685
5	Illinois	2000	9507.251	32645
6	Indiana	2000	9506.387	27459

Types of data

Variables in our regression equations were often subscripted i to represent the cross-sectional units:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Types of data

Some analyses call for **time series** data: one unit observed at multiple points in time (t).

state_name	year	expp_co2	personal_i~c
Tennessee	1990	5779.153	16574
Tennessee	1991	5403.432	17242
Tennessee	1992	5612.848	18527
Tennessee	1993	5663.499	19331
Tennessee	1994	5815.078	20283
Tennessee	1995	5872.513	21339
Tennessee	1996	6264.442	22136
Tennessee	1997	6599.753	23031
Tennessee	1998	6743.191	24462
Tennessee	1999	6931.634	25370
Tennessee	2000	7085.176	26689
Tennessee	2001	7204.988	27551
Tennessee	2002	7292.418	28162
Tennessee	2003	7539.207	29041
Tennessee	2004	7779.926	30285
Tennessee	2005	7694.059	31327
Tennessee	2006	7586.821	32885

Types of data

Variables in a time series regression equations are typically subscripted t to represent the time periods:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

Types of data

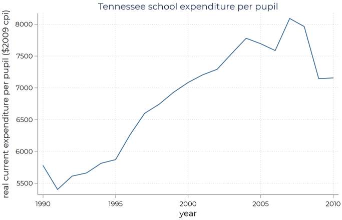
Panel data is many cross-sectional units (i) observed at multiple points in time (t).

state_name	year	expp_co2	personal_i~c
Alaska	2005	12606.64	36911
Alaska	2006	13115.29	38951
Alaska	2007	15148.9	41316
Alabama	2005	8439.994	29681
Alabama	2006	8936.927	31208
Alabama	2007	9491.025	32528
Arkansas	2005	8821.27	27858
Arkansas	2006	8929.807	29385
Arkansas	2007	8977.985	31353
Arizona	2005	7156.699	32223
Arizona	2006	7785.889	34326
Arizona	2007	7995.629	35441
California	2005	9118.798	38731
California	2006	9526.36	41518
California	2007	10042.94	43211
Colorado	2005	8970.674	38795
Colorado	2006	8817.396	41181
Colorado	2007	9469.688	42724
Connecticut	2005	14359.31	48134

Time series applications

We will be looking at research designs with both time series and panel data. Why might time series data be useful?

- **Forecasting**: predicting the future using information from the past
- **Impact analysis**: in <u>some</u> cases, may be able to infer the effects of a policy change or other event



Time series applications

These are very different approaches!

- With forecasting, we're not estimating any causal effects—just trying to get the best out-of-sample prediction—so omitted variables bias is not really an issue.
- With impact analysis, we <u>are</u> interested in causal inference, so we must be very attentive to omitted variables bias.

Time series data also introduces other issues we must pay attention to, like correlation over time in observed values.

Forecasting

Forecasting

Forecasting models can get very complex—stock traders use very sophisticated ones—but the idea is to use data from past periods to predict observations in the future.

- **Autoregression**: predicting future values of Y using past values of Y.
- Can also use linear and nonlinear fits to past values to predict future values.

Autoregression

Models generally take the form:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-1} + u_t$$

In this type of model Y in period t is usually strongly correlated with Y in period $t-1$. This is called **autocorrelation** or **serial correlation**. It is important to account for this in calculating standard errors.

Estimating time trends

Another approach is to fit a linear (or nonlinear) function of time. For example, with annual data, choose a base year where $t = 0$. A linear regression on t is:

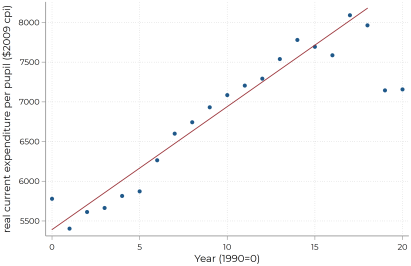
$$Y_t = \beta_0 + \beta_1 t + u_t$$

A quadratic function of t:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + u_t$$

Estimating time trends

Predicting school expenditures per pupil in Tennessee using 1990-2008 data and a linear time trend:



Estimating time trends

Estimated regression coefficients (uses 1990-2008):

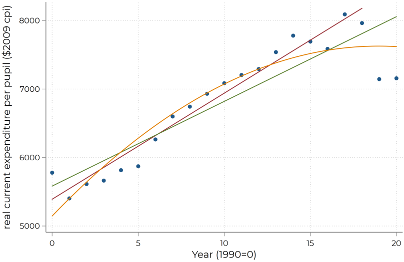
```
. reg expp_co2 yeart if year<=2008

      Source |       SS           df       MS      Number of obs   =        19
-------------+----------------------------------   F(1, 17)        =    371.79
       Model |  13676317.7         1  13676317.7   Prob > F        =    0.0000
    Residual |  625346.734        17  36785.102    R-squared       =    0.9563
-------------+----------------------------------   Adj R-squared   =    0.9537
       Total |  14301664.4        18  794536.91    Root MSE        =    191.79

------------------------------------------------------------------------------
    expp_co2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       yeart |   154.8985   8.033385    19.28   0.000     137.9495    171.8474
       _cons |   5391.339   84.63696    63.70   0.000     5212.771    5569.908
------------------------------------------------------------------------------
```

Estimating time trends

Things look different—and less useful—if you include 2009 and 2010 with a linear (or quadratic) fit:



Something clearly happened in 2009.

Interrupted time series

Interrupted time series

In an **interrupted time series** (ITS), there is some kind of treatment or intervention that happens at a specific point in time. (It "interrupts" the time series). The treatment might be best represented by an intercept shift (where $Post_t = 1$ if t is $\geq$ the treatment date):

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Post_t + u_t$$

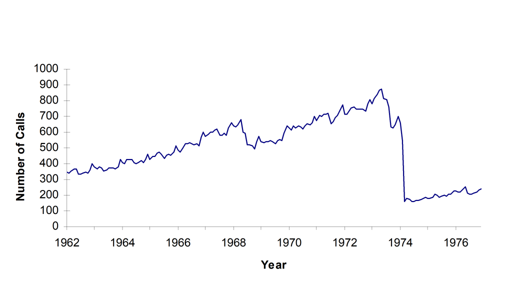
Or, it might be represented by an intercept shift *and* and slope change:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Post_t + \beta_3 Post_t \times t + u_t$$

Here there is an interaction between t and "Post".

Interrupted time series

ITS models are sometimes used to make causal inferences about the effect of a treatment, although we should generally be cautious doing so. Example: effect of charging for directory assistance in Cincinnati:



Interrupted time series

The above example can make a strong claim to causal inference:

- There is a very clear intervention time point (1974).
- The effect is instantaneous and very large.
- There is no ambiguity about the functional form since we have a long time series (since 1962) and a pretty predictable trend.
- There are no known alternative explanations for the large and sudden drop in directory assistance calls at the same point in time.

Interrupted time series in education

It can be challenging to meet these conditions in education policy research. Why?

- Long time series are often not available, which makes it difficult to find a functional form to establish the trend.
- Interventions are often not instantaneous, but take time to implement and observe the effects of.
- Effect sizes are usually small—nothing like the Cincinnati example.

Consequently, we need better methods to discern the counterfactual. What would have happened to the treated group (students, schools, districts, state) had they not been treated.

Internal validity in ITS: history

The most common threat to internal validity in a simple ITS is **history**—that some other event occurred around the same time as the intervention and could have produced the same effect. Possible solutions:

- Add a control group (exposed to the same history).
- Add a nonequivalent dependent variable.
- The narrower the intervals measured (e.g., monthly rather than yearly), the fewer the historical events that can explain the findings within that interval.

Note: stock analysts often work with high-frequency data (e.g., minutes) giving them the opportunity to estimate the effect of "events" on a stock price, such as announcements (like a new CEO).

Internal validity in ITS: instrumentation

Another threat is **instrumentation**, where the way an outcome was measured changes at the same time that the intervention occurred. This can happen in education when news tests are adopted or tests are re-scaled.

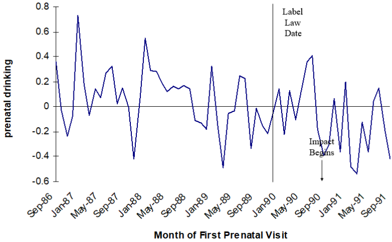
- In Chicago, when Orlando Wilson took over the CPD (1960), he changed reporting requirements, making reporting more accurate. The result appeared to be an increase in crime when he took office.

It is important to have a deep understanding of the outcome measure over time and ask about any changes that may have been made (and when).

Internal validity in ITS: diffusion

Another threat to ITS validity is **diffusion**, when there is ambiguity about when the intervention took effect.

- For example, legislation may pass in year t but not take effect or be fully implemented until year $t+1$ or $t+2$



Difference-in-differences

Difference-in-differences

Difference-in-differences is a design that—in its most common (but not only) application—contrasts *changes over time* for treated and untreated groups. DD is often used with **natural experiments**, settings in which an external force "naturally" assigns units into treatment and control groups.

DD models are typically estimated with *panel* or *repeated cross-section* data. But they can also work with other data structures.

High-stakes testing in Chicago

Do test-based "high-stakes" accountability policies improve student academic performance?

- A potential "natural experiment": in Chicago, the Iowa Test of Basic Skills (ITBS) became "high stakes" for students and schools in 1997. The test was administered—but was "low stakes"—prior to that year. The test is given in grades 3, 6, and 8.
- Many other districts in Illinois also regularly administered the ITBS to these grades, but the test was low stakes.

Note: this is a simplified example inspired by Jacob (2005).

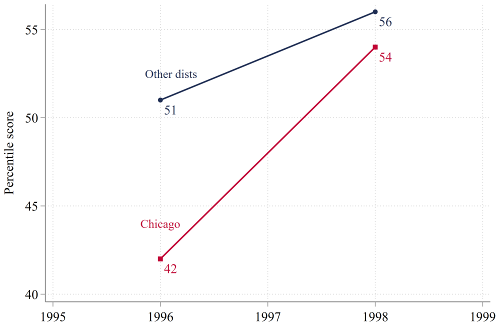
High-stakes testing in Chicago

Consider two comparisons:

- "Cross-sectional": the mean scores of Chicago 6th graders in 1998 (treated) vs. other Illinois 6th graders in 1998 (untreated).
- "Interrupted time series (ITS)": the pre-to-post change in mean scores of Chicago 6th graders between 1996 and 1998.

Clearly, a better ITS design would have more data points than two—to better establish a trend—but this is just an example!

High-stakes testing in Chicago



High-stakes testing in Chicago

The cross sectional comparison in 1998 suggests *worse* outcomes for Chicago:

$$Y_{Chicago,1998} - Y_{Other,1998} = 54 - 56 = -2$$

The **first difference** for Chicago suggests a large *improvement*:

$$Y_{Chicago,1998} - Y_{Chicago,1996} = 54 - 42 = +12$$

Conflicting conclusions!

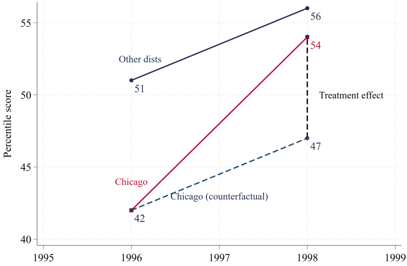
High-stakes testing in Chicago

Problems:

- The cross sectional comparison fails to recognize that Chicago 6th graders performed worse in 1996 than 6th graders in other districts did (i.e., baseline differences between treated and untreated—*an omitted variables bias problem*).
- The first difference is unable to differentiate between a treatment effect for Chicago (if any) and gains between 1996 and 1998 common to all cohorts (accounting for "history").

High-stakes testing in Chicago

Under the assumption that the change over time in other (untreated) districts represents what *would have happened* in Chicago (treated) in the absence of treatment, we can contrast *changes* in the two, or the **difference-in-differences**:



High-stakes testing in Chicago

The difference-in-differences:

$$\delta_{DD} = \underbrace{(Y_{Chicago,1998} - Y_{Chicago,1996})}_{\text{Change in Chicago}} - \underbrace{(Y_{Other,1998} - Y_{Other,1996})}_{\text{Change in other districts}}$$

$$\delta_{DD} = (54 - 42) - (56 - 51) = +7$$

The differencing of the two "first differences" represents the **second difference**. There was a "counterfactual" gain of 5 implied by the other districts.

High-stakes testing in Chicago

An equivalent way to write δ_{DD}:

$$\delta_{DD} = \underbrace{(Y_{Chicago,1998} - Y_{Other,1998})}_{\text{Difference "post"}} - \underbrace{(Y_{Chicago,1996} - Y_{Other,1996})}_{\text{Difference "pre"}}$$

Writing δ_{DD} this way makes it clear we are "netting out" pre-existing differences between the two groups.

Note in this example δ_{DD} was calculated using only four numbers (mean scores in Chicago and other districts for 1996 and 1998).

Card & Krueger (1994)

A classic DD study of the impact of the minimum wage on fast food employment (an industry likely to be affected by the minimum wage).

- NJ increased its minimum wage in April 1992, PA did not.
- Card & Krueger collected data on employment at fast food restaurants in NJ and Eastern PA before and after the minimum wage increase.

Next figure: the minimum wage increase seemed to be "binding." That is, it really did lead to higher starting wages in NJ. (This is important—if the minimum wage were not binding, it wouldn't make for a very interesting study).

Card & Krueger (1994)

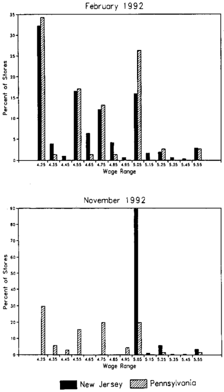


FIGURE 1. DISTRIBUTION OF STARTING WAGE RATES

Card & Krueger (1994)

Main result (portion of Table 3 in C&K):

	Stores by State		
	PA	*NJ*	*NJ* − *PA*
FTE before	23.3 (1.35)	20.44 (-0.51)	-2.89 (1.44)
FTE after	21.15 (0.94)	21.03 (0.52)	-0.14 (1.07)
Change in mean FTE	-2.16 (1.25)	+0.59 (0.54)	**2.76** **(1.36)**

Standard errors in parentheses. FTE=full time equivalent employees.

Mean employment fell in PA and *rose* in NJ, for $\delta_{DD} = 2.76$. A surprising result to many economists who expected to see a reduction in employment following an increase in the minimum wage.

2x2 difference-in-differences

The two examples thus far are the simplest form of difference-in-differences:

- Two groups: treated and an untreated comparison
- Two time periods: pre and post, before and after treatment occurs
- Treated units are all treated at the same time

DD can accommodate much more complicated designs (e.g., staggered treatment, multiple groups).

Causal interpretation of difference-in-differences

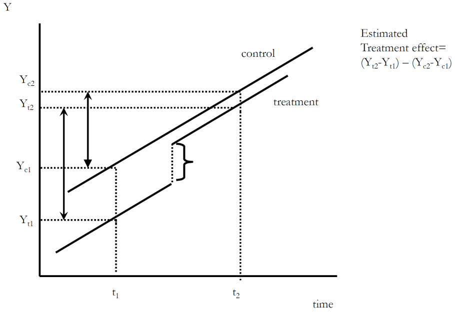
Under what conditions might the difference-in-differences design estimate a *causal effect*?

The key condition is the **parallel trends assumption**: the change over time for the control group represents what would have happened to the treatment group had they not been treated.

- We often "test" the parallel trends assumption by looking at trends in the two groups prior to treatment. If trends were similar *before* treatment, we might expect them to remain similar in the absence of treatment.

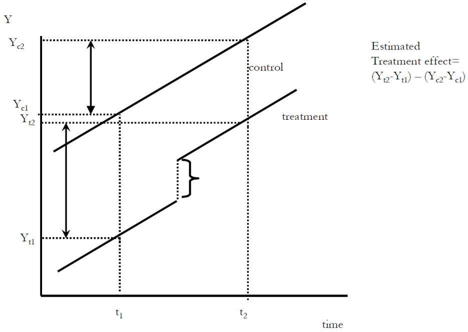
Parallel trends assumption

The key assumption in DD is parallel trends: that the time trend in the absence of treatment would be the same in both groups.

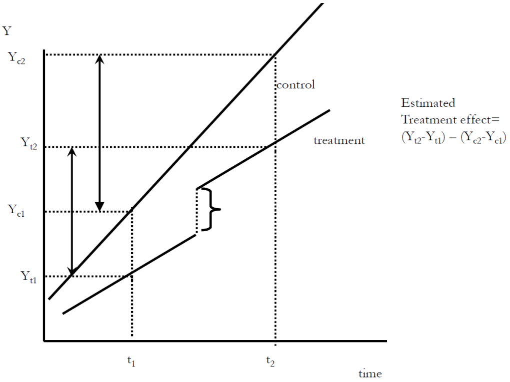


Parallel trends assumption

Size of baseline difference in treated and untreated groups doesn't matter.



Violation of parallel trends assumption



Regression difference-in-differences (2x2)

With many units (i) and two groups ($D_i = 0$ or $D_i = 1$) observed in "pre" and "post" periods, we can use regression to estimate δ_{DD}:

$$Y_{it} = \alpha + \beta D_i + \lambda POST_t + \delta(D_i \times POST_t) + u_{it}$$

where $D_i = 1$ for units i who are ultimately treated, and $POST_t = 1$ for observations in the "post" period. The post period is the same for all units.

Very easy to implement in Stata, especially with factor variable notation:

```
reg y i.evertreated##i.post
```

Regression difference-in-differences (2x2)

$$Y_{it} = \alpha + \beta D_i + \lambda POST_t + \delta(D_i \times POST_t) + u_{it}$$

α is the pre-period mean for the $D_i = 0$ group

$\alpha + \beta$ is the pre-period mean for the $D_i = 1$ group

β is the baseline mean *difference* between the $D_i = 0$ and $D_i = 1$

$\alpha + \lambda$ is the *post*-period mean for the $D_i = 0$ group

λ is the change over time for the $D_i = 0$ group

$\alpha + \beta + \lambda + \delta$ is the *post*-period mean for the $D_i = 1$ group

$\lambda + \delta$ is the change over time for the $D_i = 1$ group

δ is the *differential* change over time for the $D_i = 1$ group **(DD)**

Regression difference-in-differences (2x2)

Example: some NYC schools adopted a breakfast in the classroom program in 2010. What was the impact of this program on average daily participation in breakfast?

```
. reg bkfast_part i.everbic##i.post

      Source |       SS           df       MS      Number of obs   =     6,160
-------------+----------------------------------   F(3, 6156)      =    122.75
       Model |  6.66627777         3  2.22209259   Prob > F        =    0.0000
    Residual |  111.439598     6,156  .018102599   R-squared       =    0.0564
-------------+----------------------------------   Adj R-squared   =    0.0560
       Total |  118.105875     6,159  .019176145   Root MSE        =    .13455

------------------------------------------------------------------------------
 bkfast_part |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.everbic |   .0364431    .011215     3.25   0.001     .0144578    .0584285
      1.post |   .0004512   .0035743     0.13   0.900    -.0065557    .0074581
             |
everbic#post |
        1 1  |   .2219777   .0177852    12.48   0.000     .1871125     .256843
             |
       _cons |   .2494476   .0022566   110.54   0.000     .2450239    .2538713
------------------------------------------------------------------------------
```

Another example: financial aid eligibility

Part 1 This exercise will replicate the simple difference-in-differences result from chapter 8 of Murnane & Willett. This example comes from Dynarski (2003), who looked at the effects of Social Security survivor's benefits on college enrollment. The dataset used here is from the NLSY and consists of high school seniors in 1979-1983. The file is called *dynarski.dta* and is available via Github:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/dynarski.dta, clear
```

1. Do a cross-tabulation of *yearsr* (the year in which the student is a senior) and *offer* (= 1 if the student was a senior in a year in which Social Security survivor's benefits were available). What is the best way to define the "treatment period" here, and which students are "treated"?

2. Estimate a first difference model for the effect of survivor's benefits by limiting the analysis to the "ever treated," comparing outcomes in the treated and non-treated periods. The outcomes of interest are *coll* (whether the student was enrolled full time in college by age 23) and *hgc23* (the highest grade completed by age 23). You can do this with an OLS regression and be sure to use the sampling weights `weight=[wt88]`. Your results can be compared to Table 8.1 in Murnane & Willett.

3. Now estimate a difference-in-differences model of the effect of survivor's benefits by including the "never treated" group in the regression. Again be sure to use the sampling weights. Your results can be compared to Table 8.2 in Murnane & Willett.

NOTE: by "first difference" I mean the ITS.

Another example: financial aid eligibility

Table 8.1 "First difference" estimate of the causal impact of an offer of $6,700 in financial aid (in 2000 dollars) on whether high-school seniors whose fathers were deceased attended college by age 23 in the United States

(a) *Direct Estimate*

H.S. Senior Cohort	Number of Students	Was Student's Father Deceased	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* (standard error)	Between-Group Difference in Avg Value of *COLL*	H_0: $\mu_{OFFER} = \mu_{NO\ OFFER}$ *t*-statistic	*p*-value
1979–81	137	Yes	Yes (*Treatment Group*)	0.560 (0.053)	**0.208***	2.14	0.017†
1982–83	54	Yes	No (*Control Group*)	0.352 (0.081)			

~p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

(b) *Linear-Probability Model (OLS) Estimate*

Predictor	Estimate	Standard Error	H_0: $\beta = 0$; *t*-statistic	*p*-value
Intercept	0.352***	0.081	4.32	0.000
OFFER	0.208*	0.094	2.23	0.013†
R^2	0.036			

~p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

Another example: financial aid eligibility

Table 8.2 Direct "difference-in-differences" estimate of the impact of an offer of $6,700 in financial aid (in 2000 dollars) on whether high-school seniors whose fathers were deceased attended college by age 23, in the United States

H.S. Senior Cohort	Number of Students	Was Student's Father Deceased?	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* *(standard error)*	Between-Group Difference in Avg Value of *COLL*	"Difference in Differences" Estimate *(standard error)*	"Difference in Differences" p-value
1979–81	137	Yes	Yes *(Treatment Group)*	0.560 (0.053)	0.208 *(First Diff)*	0.182* (0.099)	0.033†
1982–83	54	Yes	No *(Control Group)*	0.352 (0.081)			
1979–81	2,745	No	No	0.502 (0.012)	0.026 *(Second Diff)*		
1982–83	1,050	No	No	0.476 (0.019)			

~ $p<0.10$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$.
†One-tailed test.

Next time

More with panel data!

- See S&W chapter 10
- Sample papers on syllabus, including Jack et al. (2023) on pandemic learning mode and student outcomes