LECTURE 10

REGRESSION WITH PANEL DATA: *DIFFERENCES-IN-DIFFERENCES ESTIMATION*

March 9, 2010

Outline for Today

Goal of Section II: Solving the evaluation problem in the absence of an experiment

1. Panel Data: notation and jargon
2. Differences-in-Differences estimation
3. Dynarksi study "Does Aid Matter"?

Regression with Panel Data

A ***panel dataset*** contains observations on multiple entities (individuals), where each entity is observed at more than one time point

Examples:

- Data on 420 school districts in 1999 *and* in 2000, for a total of 840 observations.
- Data on 50 U.S. states in which each state is observed in 3 years (150 observations)
- Data on 50,000 students in an urban school district over six years(300,000 observations)

Notation for panel data

A double subscript distinguishes entities and time periods

i = entity (state), n = number of entities,
so $i = 1,\ldots,n$

t = time period (year), T = number of time periods,
so $t = 1,\ldots,T$

Data: Suppose we have 1 independent variable. The data are:

$$(X_{it}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

Notation for panel data

Panel data with k independent variables:

$$(X_{1it}, X_{2it},\ldots,X_{kit}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

n = number of entities

T = number of time periods (years)

Terminology

- Panel data is also called ***longitudinal data***
- ***Balanced panel:*** no missing observations (all variables are observed for all entities and all time periods)

Example of a panel data set: Traffic deaths and alcohol taxes

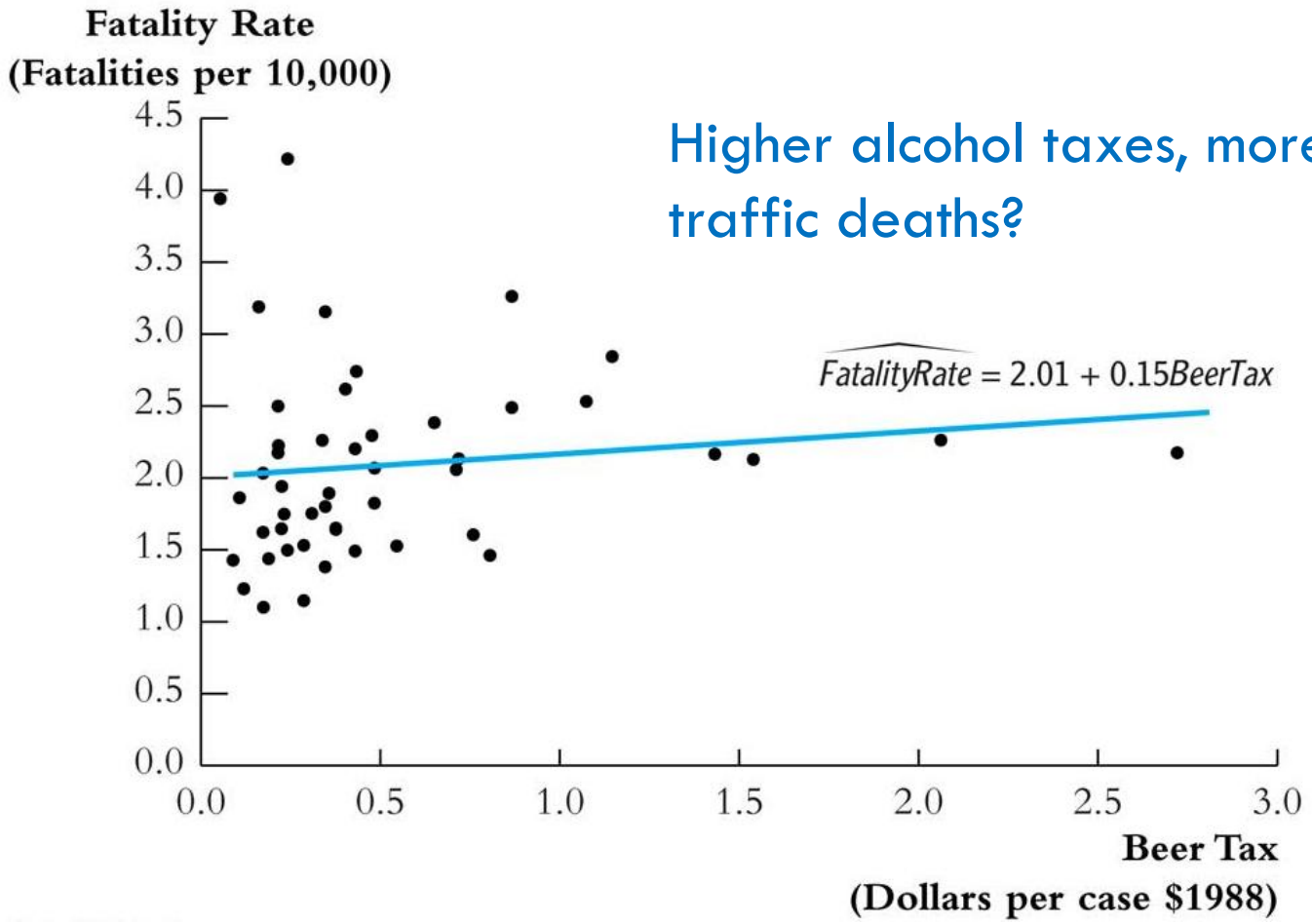
Unit of observation: a year in a U.S. state

- 48 (contiguous) U.S. states, so n = # of entities = 48
- 7 years (1982,…, 1988), so T = # of time periods = 7
- Balanced panel, so total # observations = 7·48 = 336

Variables:

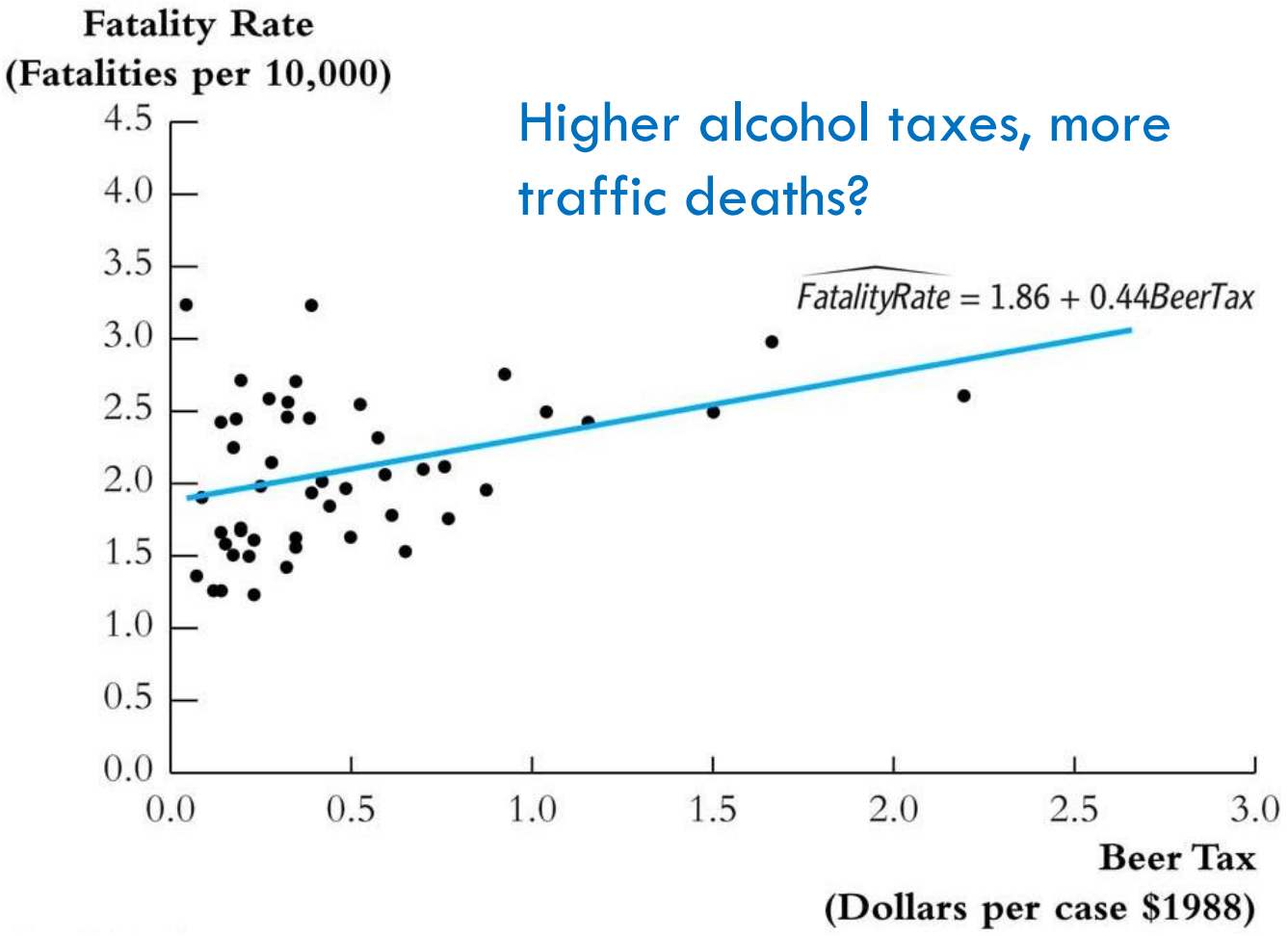
- Outcome variable: Traffic fatality rate (# traffic deaths in that state in that year, per 10,000 state residents)
- Independent variable: Tax on a case of beer

U.S. traffic death data for 1982



Higher alcohol taxes, more traffic deaths?

U.S. traffic death data for 1988



Higher alcohol taxes, more traffic deaths?

Why would there be *more* traffic deaths in states that have higher beer taxes?

- Potential omitted variables:

- Potential endogeneity:
  - States may adopt higher beer taxes in response to a large number of accidents involving drunk driving

Panel Data with Two Time Periods: Diffs-in-Diffs Estimation

- Consider the panel data model:

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

- where...
  - Z_i is a factor that does not change significantly over time (e.g. condition of roads), at least during the years in our data.
  - Z_i is unobserved, so it could cause omitted variable bias.

- Key insight: we can eliminate the effect of Z_i by "differencing" the regressions estimated for data in two different years.

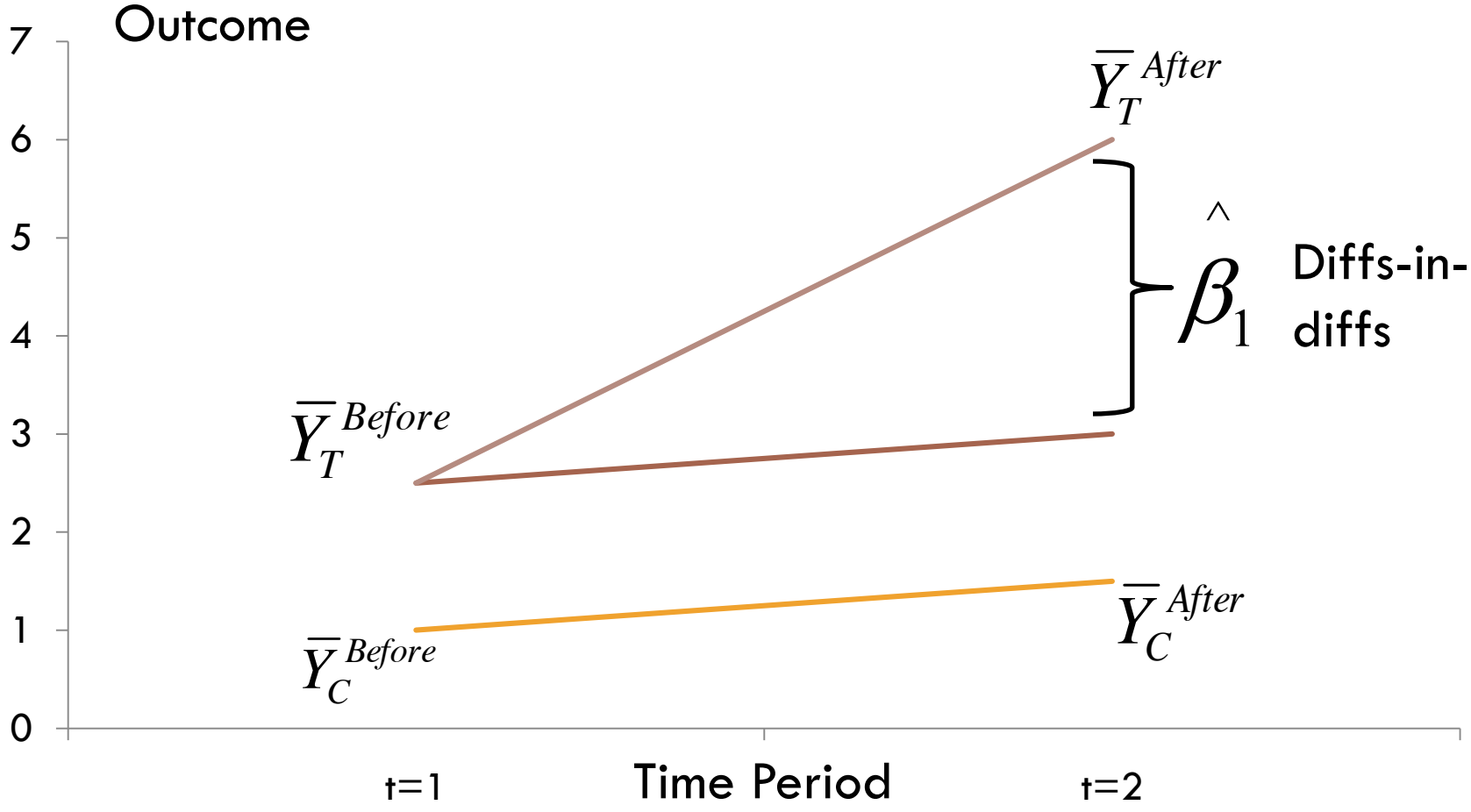
Difference-in-Differences Estimator

- We want to find the effect of $\bar{Y}_T^{After} - \bar{Y}_T^{Before}$
- Some of this effect will be because of the program, but some will occur naturally over time
- The effect of time can be measured using the control group $\bar{Y}_C^{After} - \bar{Y}_C^{Before}$

$$\text{Impact of program} = \left(\bar{Y}_T^{After} - \bar{Y}_T^{Before}\right) - \left(\bar{Y}_C^{After} - \bar{Y}_C^{Before}\right)$$

$$(treatment + time) - (time) = treatment$$

Difference-in-Differences Estimator



Fatality Rate and Beer Tax Example

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$

$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

so...

$$FatalityRate_{i1988} - FatalityRate_{i1982} =$$

$$\beta_3(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982})$$

- The diffs-in-diffs equation can be estimated by OLS, even though Z_i isn't observed.

Difference-in-Difference Estimation

- 1982 regression

$$FatalityRate_{i1982} = 2.01 + 0.15BeerTax \quad (n = 48)$$
$$(.15) \quad (.13)$$

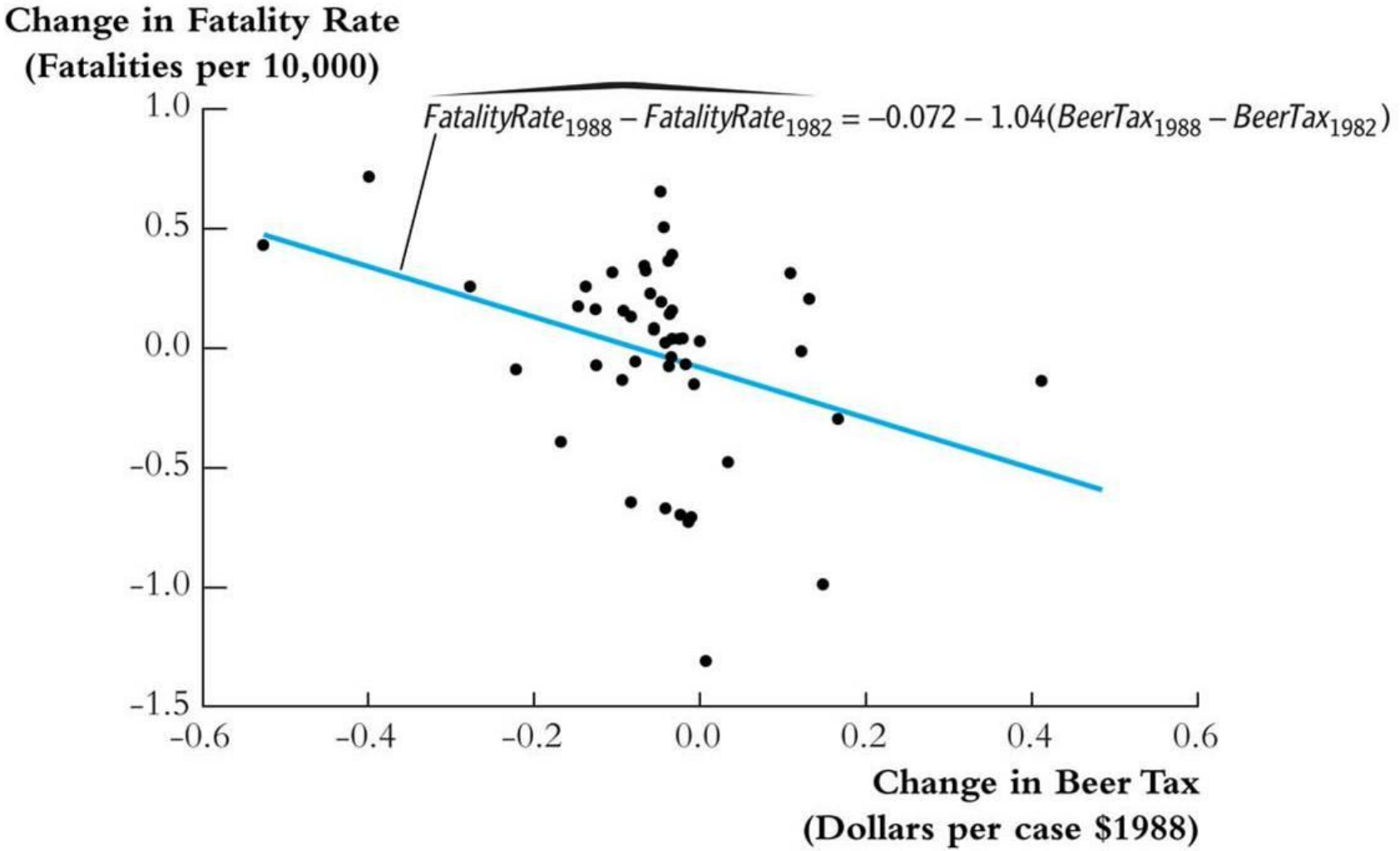
- 1988 regression

$$FatalityRate_{i1988} = 1.86 + 0.44BeerTax \quad (n = 48)$$
$$(.11) \quad (.13)$$

- Diffs-in-diffs regression $(n = 48)$

$$FR_{i1988} - FR_{i1982} = -.072 - 1.04(BeerTax1988 - BeerTax1982)$$
$$(.065) \quad (.36)$$

$\Delta FatalityRate$ v. $\Delta BeerTax$



Causal Claims Depend on Exogenous Assignment

True Experimental Solution?

- Randomly assign financial aid to eligible HS seniors
- Compare the college-going behavior of treatment and control students
- The difference in the values of the outcome between these two groups equals the causal impact of financial aid availability on college-going.
- Issues of ethics and feasibility (cost)?

Dynarski's Natural Experiment

- Seek a source of "plausibly exogenous" variation in the assignment of financial aid, such as a policy change in the availability of Federal financial aid funds
- Compare the college-going behavior of students affected by the policy change vs. those that were not

Dynarksi Paper: Policy Change

- **Social Security Benefit Program** (SSSB):
  - Eligible participants were 18-22 year old children of deceased, disabled, or retired social security beneficiaries
  - Each received an annual payment of $6700 to enroll fulltime in college
  - At program's peak, 12% of fulltime college students received SSSB benefits
- SSSB was cancelled by Congress in budget cut-backs in 1981.

Data for "Does Aid Matter?"

- National Longitudinal Survey of Youth
  - 12,686 young men and women age14-22 years old in1979 followed annually through 1994.
- Dynarkski determined which HS Seniors were eligible for program benefits before and after the policy change by identifying those whose fathers had died:
  - Before the SSSB policy change students were recipients of the "voucher"
  - After the policy change= "no voucher."

Treatment and Control Group

	High-school seniors 1979–1981		High-school seniors 1982–1983	
	Father not deceased	Father deceased	Father not deceased	Father deceased
Attend college by 23	0.502	0.560	0.476	0.352
	(0.010)	(0.043)	(0.015)	(0.066)
Complete any college by 23	0.487	0.560	0.459	0.361
	(0.010)	(0.043)	(0.015)	(0.066)
Years of schooling at 23	13.41	13.44	13.25	12.90
	(0.03)	(0.13)	(0.05)	(0.20)
Number of observations	2,745	137	1,050	54

"Treatment" Group

"Control" Group

Comparing "treatment" and "control" students

	High-school seniors 1979–1981		High-school seniors 1982–1983	
	Father not deceased	Father deceased	Father not deceased	Father deceased
Attend college by 23	0.502	0.560	0.476	0.352
	(0.010)	(0.043)	(0.015)	(0.066)
Complete any college by 23	0.487	0.560	0.459	0.361
	(0.010)		(0.01	
	3.41		13.2	
	(0.03)		(0.05	
	,745		1,05	

If the assignment to T and C was exogenous, then the mean difference in outcome between groups gives an unbiased estimate of treatment effect

Estimated probability that a student in the treatment group will attend college by age 23 is 0.560

Estimated Probability that a student in the control group will attend college by age 23 is 0.352

$$\hat{\Delta} = \bar{Y}_T - \bar{Y}_C = II - IV$$
$$= 0.560 - 0.352$$
$$= 0.208^{***}$$

Potential Problems

- The treatment and control groups were formed in different years
- It's possible that a secular trend over time affected all potential students causing a general decline in college going, among those who were eligible for SSSB-related financial aid and those who were not

Solution: Differences-in-Differences

Secular trends can often be removed by a **differences-in-differences** approach

	High-school seniors 1979–1981		High-school seniors 1982–1983		
	I	II	III	IV	
	Father not deceased	Father deceased	Father not deceased	Father deceased	Difference-in-differences
Attend college by 23	0.502 (0.010)	0.560 (0.043)	0.476 (0.015)	0.352 (0.066)	0.182 (0.096)

$$\hat{\Delta} = [\text{"}\,Secular Trend + Treatment Effect\,\text{"}] - [\text{"}\,Secular Trend\,\text{"}]$$

$$= \begin{bmatrix} Treatment/Control\ Difference \\ among\ eligible students \end{bmatrix} - \begin{bmatrix} Pre/Post\ Difference \\ among\ in\text{-}eligible students \end{bmatrix}$$

$$= [(II) - (IV)] - [(I) - (III)]$$

$$= [0.560 - 0.352] - [0.502 - 0.476]$$

$$= [0.208] - [0.026]$$

$$= 0.182^{*}$$

About 18% more HS seniors went to college by age-23 when aid was available, accounting for a secular trend in college-going.

Secular trends can often be removed by a **differences-in-differences** approach

	High-school seniors 1979–1981		High-school seniors 1982–1983		
	I	II	III	IV	
	Father not deceased	Father deceased	Father not deceased	Father deceased	Difference-in-differences
Attend college by 23	0.502	0.560	0.476	0.352	0.182
	(0.010)	(0.043)	(0.015)	(0.066)	(0.096)

$$s.e.\left[\left(\bar{Y}_2-\bar{Y}_4\right)-\left(\bar{Y}_1-\bar{Y}_3\right)\right]=\sqrt{\left(\left[\frac{s_2^2}{n_2}\right]+\left[\frac{s_4^2}{n_4}\right]\right)+\left(\left[\frac{s_1^2}{n_1}\right]+\left[\frac{s_3^2}{n_3}\right]\right)}$$

$$=\sqrt{\left(\left[s.e.\left(\bar{Y}_2\right)\right]^2+\left[s.e.\left(\bar{Y}_4\right)\right]^2\right)+\left(\left[s.e.\left(\bar{Y}_1\right)\right]^2+\left[s.e.\left(\bar{Y}_3\right)\right]^2\right)}$$

$$=\sqrt{\left(\left[0.043\right]^2+\left[0.066\right]^2\right)+\left(\left[0.010\right]^2+\left[0.015\right]^2\right)}$$

$$=0.081$$

Regression Analysis

Y=some aspect of college-going behavior) for the i^{th} HS Senior

ELIG=whether the i^{th} individual was eligible for the SSSB aid because the father was dead (1 = eligible; 0 = ineligible).

$$Y_i = \beta_0 + \beta_1 PRE_i + \beta_2 ELIG_i + \beta_3 (PRE_i \times ELIG_i) + \varepsilon_i$$

PRE =whether the i^{th} individual was a HS Senior pre- or post- the SSSB policy change (1 = pre; 0 = post)

PRE xELIG= differences-in-differences parameter

Threats to Validity

- **More pervasive in a natural experiment than in a true experiment**
- The mean outcome difference between treatment and control groups will be an unbiased estimate of the impact of treatment, ONLY if:
  - ✓ The focal policy change has truly provided exogenous treatment assignment
  - ✓ Treatment and control groups are truly identical on average, in all respects other than their access to treatment.

Summary of Panel Data Estimation

- Advantages:
  - Can eliminate unobserved variables that…
    - vary across units but not within units over time and/or…
    - vary over time but not across units
  - Different (and usually better) source of variation than cross-sectional analysis
  - Straightforward estimation via bivariate or multivariate OLS
- Limitations:
  - Need variation in treatment variable w/in units
  - Standard errors usually also need to allow for "clustering"

Thursday Preview

1. Fixed Effects
    1. Murnane & Willett, Ch. 7
    2. Hanushek & Raymond, 2003