Lecture 7: Introduction to Panel Data

Applied Micro Econometrics,Fall 2021

Zhaopeng Qu

Nanjing University Business School

December 09 2021

Panel Data: What and Why

Introduction

What is Panel Data

- So far, we have only focused on data cross entities.Now it is the time to add time, which leads us to use **Panel Data**.
- **Panel data** refers to data with observations on *multiple entities, where each entity is observed at two or more points in time*.
- If the data set contains observations on the variables X and Y,then the data are denoted

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

  - the first subscript,i refers to the entity being observed
  - the second subscript,t refers to the date at which it is observed
- Extension: not necessarily involves time dimension
  - outcome of employee i in firm m
    $(X_{im}, Y_{im})\ i = 1, ...n \ and \ m = 1, ..., M$

Introduction: Data Structure

- **Balanced** v.s **Unbalanced**
  - Balanced panel: *each unit of observation i is observed the same number of time periods, T. Thus, the total sample size is NT.*
  - Unbalanced panel: *each unit of observation i is observed an unequal number of time periods, T_i, commonly some missing values for some entities at some periods*.
- **Micro** v.s **Macro**
  - Micro: large N, and small T,more similar to cross-section data
  - Macro: small N, and large T,more similar to time series data
- In our class, we focus on **balanced** and **micro** panel data.

Example: Traffic Deaths and Alcohol Taxes

	state	year	beertax	fatal	pop	fa_rate
1	al	1982	1.53937948	839	3942002.2	2.12836
2	al	1985	1.65254235	882	4021007.8	2.19348
3	al	1984	1.71428561	932	3988991.8	2.33643
4	al	1983	1.78899074	930	3960008.0	2.34848
5	al	1988	1.50144362	1023	4101992.2	2.49391
6	al	1986	1.60990703	1081	4049993.8	2.66914
7	al	1987	1.55999994	1110	4082999.0	2.71859
8	az	1983	0.20642203	675	2977004.2	2.26738
9	az	1982	0.21479714	724	2896996.5	2.49914
10	az	1988	0.34648702	944	3488995.0	2.70565
11	az	1987	0.36000001	937	3385996.2	2.76728
12	az	1985	0.38135594	893	3186998.0	2.80201
13	az	1984	0.29670331	869	3071995.8	2.82878
14	az	1986	0.37151703	1007	3278998.0	3.07106
15	ar	1984	0.59890109	525	2346001.8	2.23785
16	ar	1985	0.57733053	534	2359001.0	2.26367
17	ar	1982	0.65035802	550	2306998.5	2.38405
18	ar	1983	0.67545873	557	2324999.0	2.39570
19	ar	1986	0.56243551	603	2371000.5	2.54323
20	ar	1988	0.52454287	610	2395002.8	2.54697
21	ar	1987	0.54500002	639	2387999.5	2.67588
22	ca	1983	0.10321102	4573	25311062.0	1.80672
23	ca	1982	0.10739857	4615	24785976.0	1.86194
24	ca	1985	0.09533899	4960	26365028.0	1.88128
25	ca	1988	0.08662175	5390	28314028.0	1.90365

Example: Traffic deaths and alcohol taxes

- Observational unit: *one* year in *one* U.S. state
  - Total 48 U.S. states, so N = the number of entities = 48
  - 7 years (1982,..., 1988),so T = the number of time periods = 7.
- Balanced panel, so total number of observations

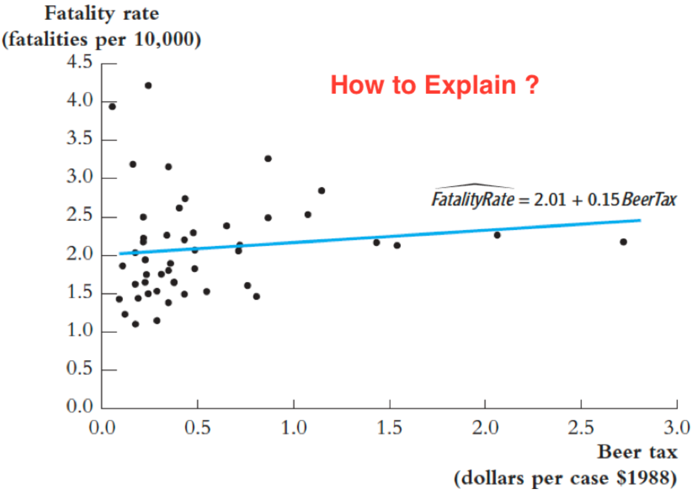
$$NT = 7 \times 48 = 336$$

- Variables:
  - Dependent 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**
  - Other Controls (legal driving age, drunk driving laws, etc.)
- A simple OLS regression model with $t = 1982, 1988$

$$FatalityRate_{it} = \beta_{0t} + \beta_{1t} BeerTax_{it} + u_{it}$$

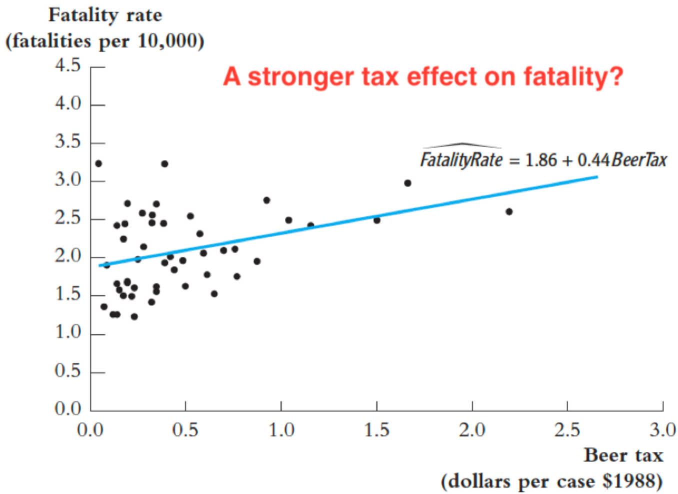
U.S. traffic death data for 1982

- Higher alcohol taxes, more traffic deaths



U.S. traffic death data for 1988

- Still higher alcohol taxes, more traffic deaths



(b) 1988 data

Pooled Cross-Sectional Data(1982-1988)

- The positive relationship between alcohol taxes and traffic deaths might be due to using only two years data.Therefore,we run the following regression using full years data

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

- This is a simple OLS, only now sample size is $NT = 7 \times 48 = 336$
- If you we would like to control the time, in other words, we would like to strict our regression within every years and then make an average, then we should run

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \lambda T_t + u_{it}$$

Pooled Cross-Sectional Data(1982-1988)

- Still higher alcohol taxes, more traffic deaths(though some nonlinear pattern)

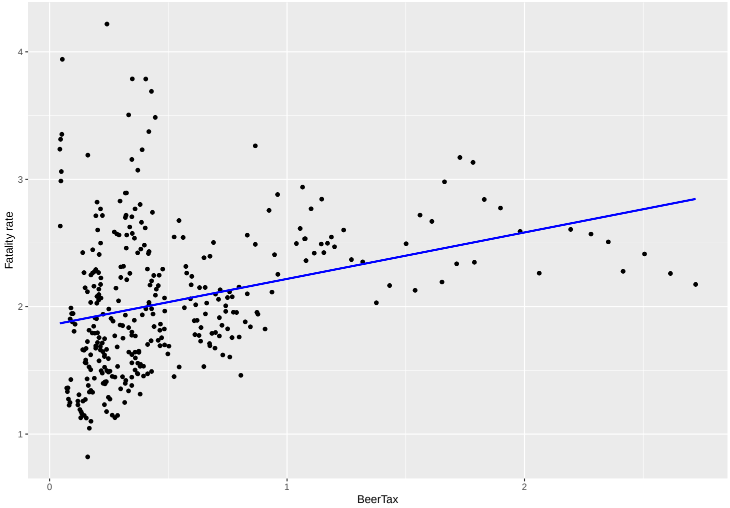


Table 1:

	Dependent Variable: Fatality Rate	
	Pooled OLS	Pooled OLS with Time
	(1)	(2)
beertax	0.365***	0.366***
	(0.053)	(0.053)
year_1983		−0.082
		(0.128)
year_1984		−0.072
		(0.121)
year_1985		−0.111
		(0.120)
year_1986		−0.016
		(0.121)
year_1987		−0.016
		(0.122)
year_1988		−0.001
		(0.119)
Constant	1.853***	1.895***
	(0.047)	(0.105)
Observations	336	336
Adjusted R^2	0.091	0.079

Pooled Cross-Sectional Data(1982-1988)

- Could we are safety to make a conclusion:
  Higher beer tax cannot make less but more fatalities
- In other words : does the regression satisfy **OLS Assumption 1-4** to obtain an *unbiased* and *consistent* estimation for the conclusion?
- **Question**: are there some threatens to the internal validity of the estimate?

Pooled Cross-Sectional Data(1982-1988)

- **Assumption 1**, $E(u_i|X_i) = 0$ may not satisfied for some unobservables(**OVB**).
  - Some unobservable factors that determines the fatality rate may be correlated with $BeerTax$, such as **local cultural attitude** toward drinking and driving.
- **Assumption 2** random sampling is not satisfied for **serial correlation** of important variables.
  - Both *Beertax* and *Fatality rate* might be serial correlated between different periods.

Before-After Model

Simple Case: Panel Data with Two Time Periods

- Firstly let adjust our model with some unobservables

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

where u_{it} is the error term and $i = 1, ...n \; and \; t = 1, ..., T$

- Z_i is the **unobservable factor** that determines the fatality rate in the i state but **does not change over time**.
- The omission of Z_i might cause omitted variable bias(**OVB**) but we don't have data on Z_i.
- The key idea: Any change in the fatality rate from 1982 to 1988 cannot be caused by Z_i, because Z_i (by assumption) does not change between 1982 and 1988.

Panel Data with Two Time Periods

- Consider the regressions for 1982 and 1988...

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$
$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

- Then make a difference

$$FatalityRate_{i1988} - FatalityRate_{i1982} =$$
$$\beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982})$$

Panel Data with Two Time Periods

- Assumption: if $E(u_{it}|BeerTax_{it}, Z_{it}) = 0$,then $(u_{i1988} - u_{i1982})$ is uncorrelated with $(BeerTax_{i1988} - BeerTax_{i1982})$
- Then this "difference" equation can be estimated by OLS, even though Z_i isn't observed.
- Intuition: because the omitted variable Z_i doesn't change, it cannot be a determinant of the change in Y.

Case: Traffic deaths and beer taxes

1982 data:

$$\widehat{FatalityRate} = \underset{(.11)}{1.86} + \underset{(.13)}{0.44}BeerTax \qquad (n = 48)$$

1988 data:

$$\widehat{FatalityRate} = \underset{(.15)}{2.01} + \underset{(.13)}{0.15}BeerTax \qquad (n = 48)$$

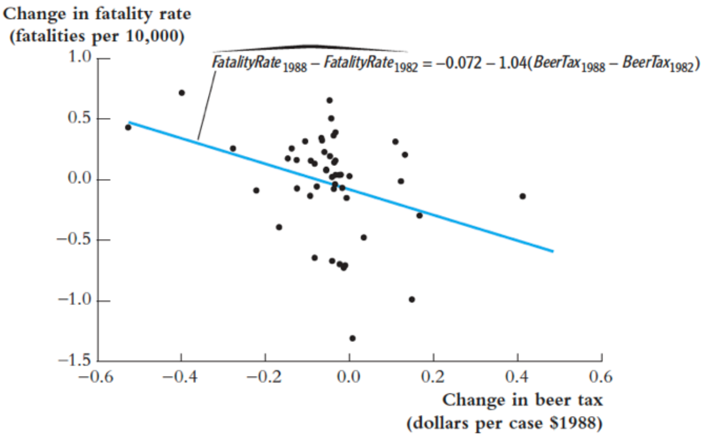
Difference regression $(n = 48)$

$$\widehat{FR_{1988} - FR_{1982}} = \underset{(.065)}{-.072} - \underset{(.36)}{1.04}(BeerTax_{1988} - BeerTax_{1982})$$

Change in traffic deaths and change in beer taxes

FIGURE 10.2 Changes in Fatality Rates and Beer Taxes, 1982–1988

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in real beer taxes between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.



Wrap up

- In contrast to the cross-sectional regression results, the estimated effect of a change in the real beer tax is **negative**, as predicted by economic theory.
- By examining changes in the fatality rate over time, the regression controls for some **unobservable but fixed factors** such as cultural attitudes toward drinking and driving.
- But there are many factors that influence traffic safety, and if they change over time and are correlated with the real beer tax, then their omission will still produce omitted variable bias(OVB).

Wrap up

- This “before and after” analysis works *when the data are observed in **two** different years*.
- Our data set, however, contains observations for **seven** different years,and it seems foolish to discard those potentially useful additional data.
- But the “before and after” method does not apply directly when $T > 2$. To analyze all the observations in our panel data set, we use a more general regression setting: **fixed effects**

Fixed Effects Model

Introduction

Introduction

- Fixed effects regression is a method for controlling for omitted variables in panel data when *the omitted variables vary across entities (states) but do not change over time*.
- Unlike the "before and after" comparisons,fixed effects regression can be used when there are **two or more time** observations for each entity.

Fixed Effects Regression Model

- The **dependent variable** (FatalityRate) and **independent variable** (BeerTax) denoted as Y_{it} and X_{it}, respectively. Then our model is

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it} \tag{11.1}$$

- Where Z_i is an **unobserved variable** that varies from one state to the next but **does not change over time**
  - eg. Z_i can still represent cultural attitudes toward drinking and driving.
- We want to estimate β_1, the effect on Y of X holding constant the unobserved state characteristics Z.

Fixed Effects Regression Model

- Because Z_i varies from one state to the next but is constant over time,then let $\alpha_i = \beta_0 + \beta_2 Z_i$,the Equation becomes

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{11.2}$$

- This is the **fixed effects regression model**, in which α_i are treated as *unknown intercepts* to be estimated, one for each state. The interpretation of α_i as a *state-specific intercept* in Equation (11.2).
- Because the intercept α_i can be thought of as the "effect" of being in entity i (in the current application, entities are states),the terms α_i,known as **entity fixed effects**.
- The variation in the entity fixed effects comes from omitted variables that, like Z_i in Equation (11.1), vary across entities but not over time.

Alternative : Fixed Effects by using binary variables

- How to estimate these parameters α_i.
- To develop the fixed effects regression model using binary variables, let $D1_i$ be a binary variable that equals 1 when i $=$ 1 and equals 0 otherwise, let $D2_i$ equal 1 when i $=$ 2 and equal 0 otherwise, and so on.
- Arbitrarily omit the binary variable $D1_i$ for the first group. Accordingly, the fixed effects regression model in Equation (7.2) can be written equivalently as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i + u_{it} \qquad (7.3)$$

- Thus there are two equivalent ways to write the fixed effects regression model, Equations (7.2) and (7.3).
- In both formulations, the slope coefficient on X is the same from one state to the next.

Estimation and Inference

Estimation: Introduction

- In principle the binary variable specification of the fixed effects regression model can be estimated by OLS.
- But it is tedious to estimate so many fixed effects.If $n = 1000$, then you have to estimate $1000 - 1 = 999$ fixed effects.
- There are some special routines, which are equivalent to using OLS on the full binary variable regression, are *faster* because they employ some *mathematical simplifications* that arise in the algebra of fixed effects regression.

Estimation: The "entity-demeaned"

- Computes the OLS fixed effects estimator in two steps
- The **first** step:
  - take the average across times t of both sides of Equation (7.2);

$$\bar{Y}_i = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_t \tag{7.4}$$

  - demeaned: let Equation(7.2) minus (7.4)

$$Y_{it} - \bar{Y}_i = \beta_1 X_{it} - \bar{X}_i + (\alpha_i - \alpha_i) + u_{it} - \bar{u}_i$$

Estimation: The "entity-demeaned"

- Let

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$$
$$\tilde{X}_{it} = X_{it} - \bar{X}_i$$
$$\tilde{u}_{it} = u_{it} - \bar{u}_i$$

- Then the **second** step: accordingly,estimate

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \tag{7.5}$$

- Then the estimator is known as the **within estimator**. Because it matters not if a unit has consistently high or low values of Y and X. All that matters is how the variations around those mean values are correlated.
- In fact, this estimator is identical to the OLS estimator of β_1 without intercept obtained by estimation of the fixed effects model in Equation (7.3)

OLS estimator without intercept

- OLS estimator without intercept

$$Y_i = \beta_1 X_i + u_i$$

- The least squared term

$$\min_{b_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - b_1 X_i)^2$$

- F.O.C, thus differentiating with respect to β_1, we get

$$\sum_{i=1}^{n} 2(Y_i - b_1 X_i) X_i = 0$$

- At last,

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

Fixed effects estimator(I)

- The second step:

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \tag{11.4}$$

- Then the fixed effects estimator can be obtained based on OLS estimator without intercept

$$\hat{\beta}_{demean} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{Y}_{it} \tilde{X}_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it}^2}$$

Fixed effect estimator(II)

- The fixed effects model is

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{7.2}$$

- Equivalence to

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i + u_{it} \tag{7.3}$$

- Then we can think of α_i as fixed effects or "nuisance parameters" to be estimated,thus yields

$$(\widehat{\beta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_n) = \underset{b, a_1, \ldots, a_n}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{t=1}^{T} (Y_{it} - bX_{it} - a_i)^2$$

this amounts to including $n = n + 1 - 1$ dummies in regression of Y_{it} on X_{it}

Fixed effect estimator(II)

- The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\widehat{\beta}X_{it}-\widehat{\alpha}_i)X_{it}=0$$

- And

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\widehat{\beta}X_{it}-\widehat{\alpha}_i)=0$$

Fixed effect estimator(II)

- Therefore, for $i = 1, \ldots, N$,

$$\widehat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \widehat{\beta}X_{it}) = \overline{Y}_i - \overline{X}_i\widehat{\beta},$$

where

$$\bar{X}_i \equiv \frac{1}{T}\sum_{t=1}^{T}X_{it}; \bar{Y}_i \equiv \frac{1}{T}\sum_{t=1}^{T}Y_{it}$$

Fixed effect estimator(II)

- Plug this result into the first FOC to obtain:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\widehat{\beta}X_{it}-\widehat{\alpha}_i)X_{it} &= \sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-X_{it}\hat{\beta}-\overline{Y}_i+\overline{X}_i\hat{\beta})X_{it}\\ &= \left(\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\overline{Y}_i)X_{it}\right)\\ &-\hat{\beta}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it}-\bar{X}_i)X_{it}\right)=0\end{aligned}$$

Fixed effect estimator(II)

- Then we could obtain

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it}-\bar{X}_i)(X_{it}-\bar{X}_i)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\bar{Y})(X_{it}-\bar{X}_i)} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^{2}}$$

  with time-demeaned variables $\tilde{X}_{it} \equiv X_{it} - \bar{X}, \tilde{Y}_{it} \equiv Y_{it} - \bar{Y}_i$
- which is same as we obtained in demeaned method.

Fixed effect estimator(III): first-differencing

- The fixed effects model is

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{11.2}$$

- Then implies

$$\begin{aligned} Y_{i1} &= \beta_1 X_{i1} + \alpha_i + u_{i1} \\ Y_{i2} &= \beta_1 X_{i2} + \alpha_i + u_{i2} \\ \vdots &= \vdots \\ Y_{iT} &= \beta_1 X_{iT} + \alpha_i + u_{iT} \end{aligned}$$

- Taking the differences between consecutive years

$$\begin{aligned} Y_{i2} - Y_{i1} &= \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1}) \\ Y_{i3} - Y_{i2} &= \beta_1 (X_{i3} - X_{i2}) + (u_{i3} - u_{i2}) \\ \vdots &= \vdots \\ Y_{iT} - Y_{i,T-1} &= \beta_1 (X_{iT} - X_{i,T-1}) + (u_{iT} - u_{i,T-1}) \end{aligned}$$

Fixed effect estimator(III): first-differencing

- New notation,we use Δ represents the change from the preceding year,then

$$\begin{aligned}\Delta Y_{i2} &= \beta_1 \Delta X_{i2} + \Delta u_{i2} \\ \Delta Y_{i3} &= \beta_1 \Delta X_{i3} + \Delta u_{i3} \\ \vdots &= \vdots \\ \Delta Y_{iT} &= \beta_1 \Delta X_{iT} + \Delta u_{iT}\end{aligned}$$

- The first-difference fixed effect model is

$$\Delta Y_{it} = \beta_1 \Delta X_{it} + \Delta u_{it} \; i = 1, ..., N, ; t = 2, ..., T \tag{11.5}$$

- Then first-difference estimator is

$$\hat{\beta}_{fd} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta Y_{it} \Delta X_{it}}{\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta X_{it}^2}$$

The Fixed Effects Regression Assumptions

- The simple fixed effect model

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

1. **Assumption 1**: u_{it} has conditional mean zero with X_{it}, or X_i at any time t and α_i

$$E(u_{it}|X_{i1}, X_{i2}, ..., X_{iT}, \alpha_i) = 0$$

2. **Assumption 2**: $(X_{i1}, X_{i2}, ..., X_{iT}, u_{i1}, u_{i2}, ..., u_{iT}), i = 1, 2, ..., n$ are $i.i.d.$
3. **Assumption 3**: Large outliers are unlikely.
4. **Assumption 4**: There is no perfect multicollinearity.

The Fixed Effects Regression Assumptions

- **Assumption 1**: u_{it} has conditional mean zero with X_{it}, or X_i at any time t and α_i,thus

$$E(u_{it}|X_{i1}, X_{i2}, ..., X_{iT}, \alpha_i) = 0$$

- u_{it} has mean zero, given the state fixed effect and the entire history of the Xs for that state.
- No feedback effect from u to future X
  - *Whether a state has a particularly high fatality rate this year does not subsequently affect whether it increases the beer tax.*

Fixed effect estimator(III): first-differencing

- When $T = 2$, FD and demean estimators and all test statistics are *identical*.
- When $T = 3$, FD and demean estimators are not the same, while both are consistent(T fixed as $N \to \infty$) if certain assumptions are satisfied.
- But if the strict exogenous assumption is not satisfied, then the demean estimator has more advantages over the FD estimator for having substantial less bias.

Statistical Properties of Fixed Effects Model

- Unbiasedness and Consistency

$$
\begin{aligned}
\widehat{\beta}_{fe-demean} &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2} \\
&= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}(\beta_1\tilde{X}_{it}+\tilde{u}_{it})}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2}
\end{aligned}
$$

Statistical Properties

- Unbiasedness and Consistency

$$\begin{aligned}\hat{\beta}_{fd} &= \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\Delta Y_{it}\Delta X_{it}}{\sum_{i=1}^{n}\sum_{t=2}^{T}\Delta X_{it}^2} \\ &= \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\Delta X_{it}(\beta_1\Delta X_{it}+\Delta u_{it})}{\sum_{i=1}^{n}\sum_{t=1}^{T}\Delta X_{it}^2} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\Delta X_{it}\Delta u_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\Delta X_{it}^2}\end{aligned}$$

- It is very familiar: paralleling the derivation of OLS estimator, we could prove the estimator of fixed effects model is **unbiased** and **consistent**.

Statistical Properties

- Similarly, in panel data, if the fixed effects regression assumptions—holds, then the sampling distribution of the fixed effects OLS estimator is normal in large samples.
- Then the variance of that distribution can be estimated from the data, the square root of that estimator is the standard error,
- And the standard error can be used to construct t-statistics and confidence intervals.
- Statistical inference—testing hypotheses (including joint hypotheses using F-statistics) and constructing confidence intervals—proceeds in exactly the same way as in multiple regression with cross-sectional data.

Fixed Effects: goodness of fit

- Three measures of goodness of fit are commonly reported
  - *Within* R^2: demeaned Y_{it} and demeaned predicted $\hat{Y}_{it}$ using demeaned X_{it} and estimate coefficient $\hat{\beta}$
  - *Between* R^2: average Y_i and average predicted $\hat{\bar{Y}}_i$ using average $\bar{X}_i$ and estimate coefficient $\hat{\beta}$
  - *Overall* R^2: Y_{it} and predicted $\hat{Y}_{it}$

Fixed Effects: Extension to multiple X's.

- The multiple fixed effects regression model is

$$Y_{it} = \beta_1 X_{1,it} + ... + \beta_k X_{k,it} + \alpha_i + u_{it}$$

- Equivalently, the fixed effects regression can be expressed in terms of a common intercept

$$\begin{aligned} Y_{it} =& \beta_0 + \beta_1 X_{1,it} + ... + \beta_k X_{k,it} \\ &+ \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i + u_{it} \end{aligned}$$

Application to Traffic Deaths

- The OLS estimate of the fixed effects regression based on all 7 years of data (336 observations), is

$$\widehat{FatalityRate} = \underset{(0.29)}{-0.66}BeerTax + StateFixedEffects$$

- The estimated state fixed intercepts are not listed to save space and because they are not of primary interest.
- As predicted by economic theory,higher real beer taxes are associated with fewer traffic deaths, which is the opposite of what we found in the initial cross-sectional regressions.

Application to Traffic Deaths

- Recall: The result in Before-After Model is

Difference regression ($n = 48$)

$$\widehat{FR_{1988} - FR_{1982}} = -.072 - 1.04(BeerTax_{1988} - BeerTax_{1982})$$
$$(.065) \quad (.36)$$

- The magnitudes of estimate coefficients are not identical,because they use different data.
- And because of the additional observations, the standard error now is also smaller than before-after model.

Extension: Both Entity and Time Fixed Effects

Regression with Time Fixed Effects

- Just as fixed effects for each entity can control for variables that are constant over time but differ across entities, so can **time fixed effects** control for variables that *are constant across entities but evolve over time*.
    - Like *safety improvements in new cars* as an **omitted variable** that changes over time but has the same value for all states.
- Now our regression model with **time fixed effects**

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$$

- where S_t is *unobserved* and where the single t subscript emphasizes that safety changes over time but is constant across states. Because $\beta_3 S_3$ represents variables that determine Y_{it}, if S_t is correlated with X_{it}, then omitting S_t from the regression leads to omitted variable bias.

Time Effects Only

- Although S_t is unobserved, its influence can be eliminated because it varies over time but not across states, just as it is possible to eliminate the effect of Z_i, which varies across states but not over time.
- Similarly,the presence of S_t leads to a regression model in which each time period has its own intercept,thus

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

- This model has a different intercept, λ_t, for each time period, which are known as **time fixed effects**.The variation in the time fixed effects comes from omitted variables that vary over time but not across entities.

Time Effects Only

- Just as the **entity fixed effects** regression model can be represented using $n-1$ binary indicators, the time fixed effects regression model be represented using $T-1$ binary indicators too:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \delta_2 B2_t + ... + \delta_T BT_t + \alpha_i + u_{it} \tag{11.18}$$

  - where $\delta_2, \delta_3, ..., \delta_T$ are unknown coefficients
  - where $B2_t = 1$ if $t = 2$ and $B2_t = 0$ otherwise and so forth.
- Nothing new, just a another form of Fixed Effects model with another explanation.

Time Effects Only

Table 2:

	Dependent Variable: Fatality Rate	
	Pooled OLS	Pooled OLS with Time
	(1)	(2)
beertax	0.365***	0.366***
	(0.053)	(0.053)
year_1983		−0.082
		(0.128)
year_1984		−0.072
		(0.121)
year_1985		−0.111
		(0.120)
year_1986		−0.016
		(0.121)
year_1987		−0.016
		(0.122)
year_1988		−0.001
		(0.119)
Constant	1.853***	1.895***
	(0.047)	(0.105)

Both Entity and Time Fixed Effects

- If some omitted variables are constant over time but vary across states (such as cultural norms) while others are constant across states but vary over time (such as national safety standards)
- Then, combined entity and time fixed effects regression model is

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

  - where α_i is the **entity fixed effect** and λ_t is the **time fixed effect**.
- This model can equivalently be represented as follows

$$\begin{aligned} Y_{it} =& \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i \\ &+ \delta_2 B2_t + \delta_3 B3_t + ... + \delta_T BT_i + u_{it} \end{aligned}$$

Both Entity and Time Fixed Effects: Estimation

- The time fixed effects model and the entity and time fixed effects model are both **variants** of *the multiple regression model*.
- Thus their coefficients can be estimated by OLS by including the additional time and entity binary variables.
- Alternatively,first deviating Y and the X's from their entity and time-period means and then by estimating the multiple regression equation of deviated Y on the deviated X's.

Application to traffic deaths

- This specification includes the beer tax, 47 state binary variables (state fixed effects), 6 single-year binary variables (time fixed effects), and an intercept, so this regression actually has $1 + 47 + 6 + 1 = 55$ right-hand variables!

$$\widehat{FatalityRate} = \underset{(0.36)}{-0.64} BeerTax + StateFixedEffects + TimeFixedEffects. \quad (10.21)$$

- When time effects are included,this coefficient is less precisely estimated, it is still significant only at the 10%, but not the 5%.
- This estimated relationship between the real beer tax and traffic fatalities is immune to omitted variable bias from variables that are constant either over time or across states.

Measurement error in FE

Recall: Classical measurement error of X

- The true model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  with $E[u_i|X_i] = 0$

- Due to the **classical measurement error**,we only have X_i^* thus

$$X_i^* = X_i + w_i$$

  with $E[w_i|X_i] = 0$

- Then we have to estimate the model is

$$Y_i = \beta_0 + \beta_1 X_i^* + e_i$$

  where $e_i = -\beta_1 w_i + u_i$

Recall: Classical measurement error of X

- Similar to OVB bias in simple OLS model, we had derived that

$$plim(\hat{\beta}_1) = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$$

- Then we have

$$plim(\hat{\beta}_1) = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \leq \beta_1$$

- The classical measurement error β_1 is biased towards 0, which is also called **attenuation bias**

Measurement error in FE

- Suppose we will estimate a fixed effect model

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

- Unfortunately, our measurement of X is not accurate, suppose it satisfies the classical measurement error, thus

$$X^*_{it} = X_{it} + w_{it}$$

with $E[w_{it}|X_{it}] = 0$

- Then we estimate

$$Y_{it} = \beta_1 X^*_{it} + \alpha_i + e_{it}$$

with $e_{it} = -\beta_1 w_{it} + u_{it}$

Measurement error in FE

- First difference estimator for fixed effect

$$\Delta Y_{it} = \beta_1 \Delta X^*_{it} + \Delta e_{it}$$

with $\Delta e_{it} = -\beta_1 \Delta w_{it} + \Delta u_{it}$

- Following the formula of ME in Simple OLS regression,we have

$$plim(\hat{\beta}_1) = \beta_1 \frac{\sigma^2_{\Delta X}}{\sigma^2_{\Delta X} + \sigma^2_{\Delta w}}$$

- Assume that time series X_t is stationary, which means that the expectation and variance are both constant.

$$\begin{aligned}\sigma^2_{\Delta X} &= Var(X_{it}) - 2Cov(X_{it}, X_{i,t-1}) + Var(X_{i,t-1}) \\ &= 2\sigma^2_X - 2\rho\sigma^2_X \\ &= 2\sigma^2_X(1-\rho)\end{aligned}$$

Measurement error in FE

- Similarly, define r to be the autocorrelation coefficient in w_{it},then the **attenuation bias** in fixed effect model is

$$plim(\hat{\beta}) = \beta \frac{\sigma_X^2(1-\rho)}{\sigma_X^2(1-\rho) + \sigma_w^2(1-r)}$$

- If both X_{it} and w_{it} are uncorrelated over time(t), then $\rho = 0$ and $r = 0$, the bias equals to the one in simple OLS case.
- If measurement error is uncorrelated over time, but X_{it} are correlated over time, thus $\rho \neq 0$ and $r = 0$.Then we have

$$plim(\hat{\beta}) = \beta \frac{\sigma_X^2(1-\rho)}{\sigma_X^2(1-\rho) + \sigma_w^2} < \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$$

- It means that *attenuation bias in fixed-effect model* will be **larger** than the bias in OLS. In other words, measurement error will be magnified in a FE model.

Autocorrelation

Autocorrelated in Panel Data

- An important difference for a key assumption
  - **Cross-Section**: Assumption 2 holds: i.i.d sample.
  - **Panel data**: independent across entities but no such restricition **within** an entity.
- Like X_{it} can be correlated over time within an entity, thus

$$Cov(X_t, X_s) \neq 0 : for\, t \neq s$$

then the X_t is said to be **autocorrelated or serially correlated**.

Autocorrelated in Panel Data

- In the traffic fatality example, X_{it}, the beer tax in state i in year t,is autocorrelated:
  - Most of the time, the legislature does not change the beer tax, so if it is high one year relative to its mean value for state i,it will tend to be high the next year,too.

Autocorrelated in Panel Data

- Similarly, u_{it} would be also autocorrelated. It consists of time-varying factors that are determinants of Y_{it} but are not included as regressors, and some of these omitted factors might be autocorrelated. It can formally be expressed as

$$Cov(u_{it}, u_{is}|X_{it}, X_{is}, \alpha_i) \neq 0 \; for \; t \neq s$$

  - eg. a downturn in the local economy and a road improvement project.

Autocorrelated in Panel Data

- If the regression errors are autocorrelated, then the usual heteroskedasticity-robust standard error formula for cross-section regression is not valid.
- The result: an analogy of **heteroskedasticity**.
- OLS panel data estimators of β are unbiased and consistent but the standard errors will be wrong
    - usually the OLS standard errors understate the true uncertainty
- This problem can be solved by using **"heteroskedasticity and autocorrelation-consistent(HAC) standard errors"**

Standard Errors for Fixed Effects Regression

- The standard errors used are one type of HAC standard errors, **clustered standard errors**.
- The term **clustered** arises because these standard errors allow the regression errors to have an arbitrary correlation within a cluster, or grouping, but assume that the regression errors are uncorrelated across clusters.
- In the context of panel data,each cluster consists of an entity.Thus **clustered standard errors** allow for heteroskedasticity and for arbitrary autocorrelation *within an entity*, but treat the errors as *uncorrelated across entities*.
- Like **heteroskedasticity-robust standard errors** in regression with cross-sectional data, **clustered standard errors** are valid whether or not there is heteroskedasticity,autocorrelation,or both.

Application: Drunk Driving Laws and Traffic Deaths

Application: Drunk Driving Laws and Traffic Deaths

- Two ways to cracks down on Drunk Driving
    1. toughening driving laws
    2. raising taxes
- Both driving laws and economic conditions could be omitted variables,it is better to put them into the regression as covariates.
- Besides, In two way fixed effect model, controlling both unobservable variables simultaneously that
    - do not change over time
    - do not vary across states

Application: Drunk Driving Laws and Traffic Deaths

TABLE 10.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: Traffic fatality rate (deaths per 10,000).

OLS: (1); Only State Fixed: (2); Both State and Time Fixed Effects: (3)–(7)

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36** (0.05)	−0.66* (0.29)	−0.64+ (0.36)	−0.45 (0.30)	−0.69* (0.35)	−0.46 (0.31)	−0.93** (0.34)
Drinking age 18				0.028 (0.070)	−0.010 (0.083)		0.037 (0.102)
Drinking age 19				−0.018 (0.050)	−0.076 (0.068)		−0.065 (0.099)
Drinking age 20				0.032 (0.051)	−0.100+ (0.056)		−0.113 (0.125)
Drinking age						−0.002 (0.021)	
Mandatory jail or community service?				0.038 (0.103)	0.085 (0.112)	0.039 (0.103)	0.089 (0.164)
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063** (0.013)		−0.063** (0.013)	−0.091** (0.021)
Real income per capita (logarithm)				1.82** (0.64)		1.79** (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes

Summary

Wrap up

- We've showed that how panel data can be used to control for **unobserved omitted variables** that differ across entities but are **constant** over time.
- The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics.
- Double fixed Effects model, thus both entity and time fixed effects can be included in the regression to control for variables that vary across entities but are constant over time and for variables that vary over time but are constant across entities.

Wrap up

- Despite these virtues, one shortcoming of fixed effect model is that it will **exaggerate the attenuation bias** as when X is measured with some errors.
- Second,fixed effect model eliminate the OVB bias with demean or differences. But in the mean time, it also **diminishes the variations of Xs** significantly, which will make the estimate less precise.
  - If the treatment variable of the interest is also constant, then it will gone when you use fixed effect model.
- Last but not least, entity and time fixed effects regression *cannot* control for *omitted variables* that *vary both across entities and over time*.There remains a need for new methods that can eliminate the influence of unobserved omitted variables.

Difference in Differences

Introduction

Difference in Differences: Introduction

- DD(or DID) is a special case for "twoway fixed effects" under certain assumption, which is one of most popular research designs in applied microeconomics.
- It was introduced into economics via Orley Ashenfelter in the late 1970s and then popularized through his student David Card (with Alan Krueger) in the 1990s.

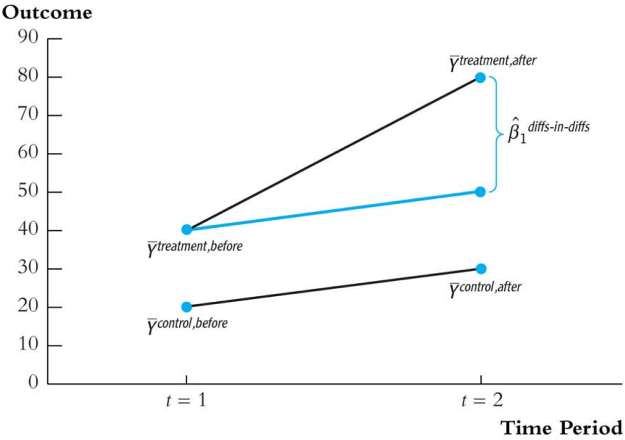
RCT and Difference in Differences

- A typical RCT design requires a causal studies to do as follow
  1. Randomly assignment of treatment to divide the population into a "treatment" group and a "control" group.
  2. Collecting the data at the time of post-treatment then comparing them.
- It works because *treatment* and *control* are randomized.
- What if we have the treatment group and the control group, but they are not fully randomized?
- If we have observations across two times at least with one before treatment and the other after treatment, then an easy way to make causal inference is **Difference in Differences(DID)** method.

DID estimator

- The DID estimator is

$$\hat{\beta}_{DID} = (\bar{Y}_{treat,post} - \bar{Y}_{treat,pre}) - (\bar{Y}_{control,post} - \bar{Y}_{control,pre})$$



Card and Krueger(1994): Minimum Wage on Employment

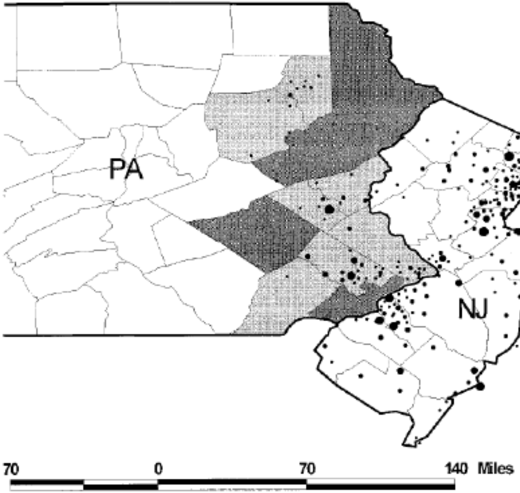
Introduction

- Theoretically,in competitive labor market, increasing binding minimum wage decreases employment.But what about the reality?
- Ideal experiment: randomly assign labor markets to a control group (minimum wage kept constant) and treatment group (minimum wage increased), compare outcomes.
- Policy changes affecting some areas and not others create natural experiments.
  - Unlike ideal experiment, control and treatment groups here are not randomly assigned.

Card and Krueger(1994): Backgroud

- Policy Change: in April 1992
  - Minimum wage in New Jersey from \$4.25 to \$5.05
  - Minimum wage in Pennsylvania constant at \$4.25
- Research Design:
  - Collecting the data on employment at 400 fast food restaurants in NJ(treatment group) in Feb.1992 (before treatment)and again November 1992(after treatment).
  - Also collecting the data from the same type of restaurants in eastern Pennsylvania(PA) as control group where the minimum wage stayed at \$4.25 throughout this period.

Card & Krueger(1994): Geographic Background



Card & Krueger(1994): Model Graph

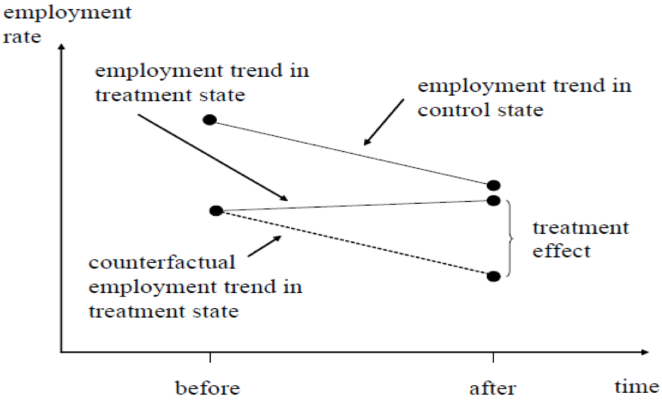


Figure 5.2.1: Causal effects in the differences-in-differences model

Card & Krueger(1994):Result

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before, all available observations	23.33	20.44	-2.89
	(1.35)	(0.51)	(1.44)
2. FTE employment after, all available observations	21.17	21.03	-0.14
	(0.94)	(0.52)	(1.07)
3. Change in mean FTE employment	-2.16	0.59	2.76
	(1.25)	(0.54)	(1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The

Regression DD - Card and Krueger

- DID model:

$$Y_{ts} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + u_{its}$$

  - NJ is a dummy equal to 1 if the observation is from NJ,
  - otherwise equal to 0(from Penny)
  - d is a dummy equal to 1 if the observation is from November (the post period),
  - otherwise equal to 0(Feb. the pre period)
- Which estimate coefficient does present DID estimator?

Regression DD - Card and Krueger

- A 2×2 matrix table

		treat or control	
		NJ=0(control)	NJ=1(treat)
pre or post	d=0(pre)	α	$\alpha+\gamma$
	d=1(post)	$\alpha+\lambda$	$\alpha+\gamma+\lambda+\delta$

- Then DID estimator

$$\begin{aligned}\hat{\beta}_{DID} &= (\bar{Y}_{treat,post} - \bar{Y}_{treat,pre}) - \\ &(\bar{Y}_{control,post} - \bar{Y}_{control,pre}) \\ &= (NJ_{post} - NJ_{pre}) - (PA_{post} - PA_{pre}) \\ &= [(\alpha+\gamma+\lambda+\delta) - (\alpha+\gamma)] - [(\alpha+\lambda) - \alpha] \\ &= \delta\end{aligned}$$

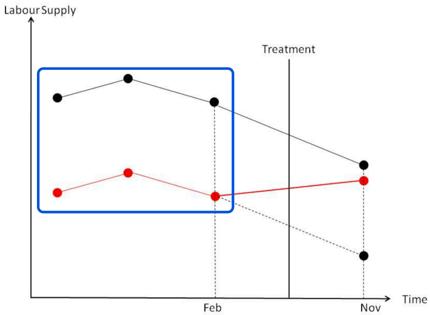
Key Assumption For DID

Paralled Trend

- A key identifying assumption for DID is: **Common trends** or **Parallel trends**
  - Treatment would be the same "trend" in both groups in the absence of treatment.
- This doesn't mean that they have to have the same mean of the outcome.
- There may be some unobservable factors affected on outcomes of both group. But as long as the effects have the same trends on both groups, then DID will eliminate the factors.
- It is difficult to verify because technically one of the parallel trends can be an unobserved counterfactual.

Assessing Graphically

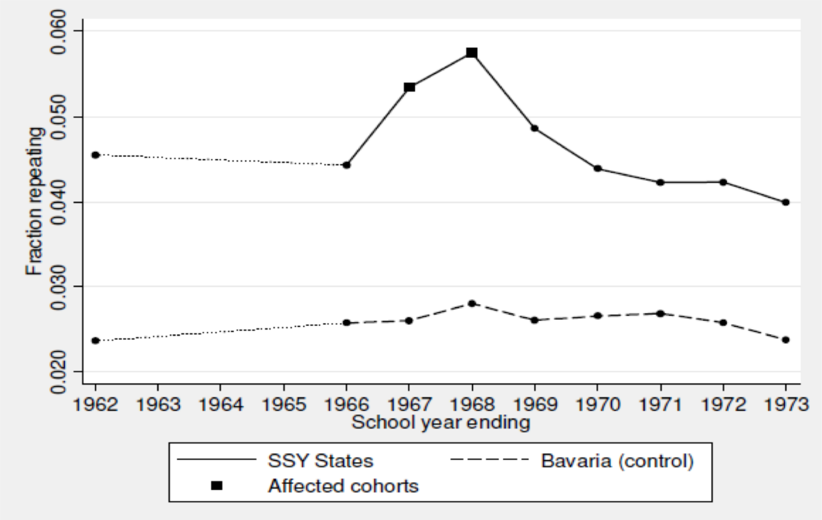
- **Common Trend**: It is difficult to verify but one often uses pre-treatment data to show that the trends are the same.
  - If you only have two-period data, you can do **nothing**.
  - If you luckly have multiple-period data, then you can show something graphically.



An Encouraging Example: Pischeke(2007)

- Topic: the length of school year on student performance
- Background:
  - Until the 1960s, children in all German states except Bavaria started school in the Spring. In 1966-1967 school year, the Spring moved to Fall.
  - It make two shorter school years for affected cohort, 24 weeks long instead of 37.
- Research Design:
  - Dependent Variable: Retreating rate
  - Independent Variable: spending time on school
  - Treatment group: Students in the German **States except Bavaria**.
  - Control group: Students in **Bavaria**.

An Encouraging Example: Pischeke(2007)



An Encouraging Example: Pischeke(2007)

- This graph provides strong visual evidence of treatment and control states with a common underlying trend.
- A treatment effect that induces a sharp but transitory deviation from this trend.
- It seems to be clear that a short school years have increased repetition rates for affected cohorts.

Extensions of DID

A Simple DID Regression

- The simple DID regression

$$Y_{ist} = \alpha + \beta(Treat \times Post)_{st} + \gamma Treat_s + \delta Post_t + u_{ist}$$

  - $Treat_s$ is a dummy variable indicate whether or not is **treated**.
  - $Post_t$ is a dummy variable indicate whether or not is **post-treatment** period.
  - γ captures the outcome gap between treatment and control group that *are constant over time*.
  - δ captures the outcome gap across post and pre period that *are common to both two groups*.
  - β is the coefficient of interest which is the *difference-in-differences* estimator
- Note: *Outcomes are often measured at the individual level i,while treatment takes place at the group level s*.

A Simple DID Regression with Covariates

- Add more covariates as **control variables** which may reduce the residual variance (lead to smaller standard errors)

$$Y_{ist} = \alpha + \beta(Treat \times Post)_{st} + \gamma Treat_s + \delta Post_t + \Gamma X'_{ist} + u_{ist}$$

  - X_{ist} is a vector of control variables. Γ is the corresponding estimate coefficient vector.
- X_{ist} can include individual level characteristics and time-varying measured at the group level.
- Those time-invariant Xs may not helpful because they are part of fixed effect which will be differential.
- Time-varying Xs may be problematic if they are the outcomes of the treatment which are **bad controls**.
- So *Pre-treatment covariates* which could include Xs on both group and individual level are more favorable.

A Simple DID Regression with More Periods

- We can slightly change the notations and generalize it into

$$Y_{ist} = \alpha + \beta D_{st} + \gamma Treat_s + \delta Post_t + \Gamma X'_{ist} + u_{ist}$$

  - Where D_{st} means $(Treat \times Post)_{st}$
- Using Fixed Effect Models further to transform into

$$Y_{ist} = \beta D_{st} + \alpha_s + \delta_t + \Gamma X'_{ist} + u_{ist}$$

  - α_s is a set of groups fixed effects, which captures $Treat_s$.
  - δ_t is a set of time fixed effects, which captures $Post_t$.
- Note:
  - Samples enter the treatment and control groups **at the same time**.
  - The frame work can also apply to **Repeated(Pooled) Cross-Section Data**.

DID for different treatment intensity

- Study treatments with different treatment intensity. (e.g., varying increases in the minimum wage for different states)'
- Card(1992) exploits regional variation in the impact of the federal minimum wage. The regression is

$$Y_{ist} = \beta(Intense_s \times D_t) + \gamma_s + \delta_t + u_{ist}$$

- Where the variable $Intense_s$ is a measure of the fraction of teenagers likely to be affected by a minimum wage increase in each state and D_t is a dummy for observations after 1990, when the federal minimum increased from $\$3.35$ to $\$3.80$.
- β means that how much does wage increase when increasing the one fraction of affected teenagers by an increase of the federal minimum wage.

Loose or Test Common Trend Assumption

Add group-speicific time trends

- This setting can eliminate the effect of group-specific time trend in outcome on our DID estimates

$$Y_{ist} = \beta D_{st} + \alpha_s + \delta_t + \tau_{st} + \Gamma X'_{ist} + u_{ist}$$

- τ_{st} is group-specific dummies multiplying the time trend variable t, which can be quadratic to capture some nonlinear trend.
- The **group specific time trend** in outcome means that treatment and control groups can follow different trends.
- It make DID estimate more robust and convincing when the pretreatment data establish a clear trend that can be extrapolated into the posttreatment period.

Add group-specific time trends

- Besley and Burgess (2004),“Can Labor Regulation Hinder Economic Performance? Evidence from India”, *The Quarterly Journal of Economics*.
  - Topic: labor regulation on businesses in Indian states
  - Method: Difference-in-Differences
  - Data: States in India
  - Dependent Variable: log manufacturing output per capita on states levels
  - Independent Variable: Labor regulation(lagged) coded $1 = pro - worker$;$0 = neutral$;$-1 = pro - employer$ and then accumulated over the period to generate the labor regulation measure.

TABLE 5.2.3
Estimated effects of labor regulation on the performance of firms in Indian states

	(1)	(2)	(3)	(4)
Labor regulation (lagged)	−.186 (.064)	−.185 (.051)	−.104 (.039)	.0002 (.020)
Log development expenditure per capita		.240 (.128)	.184 (.119)	.241 (.106)
Log installed electricity capacity per capita		.089 (.061)	.082 (.054)	.023 (.033)
Log state population		.720 (.96)	0.310 (1.192)	−1.419 (2.326)
Congress majority			−.0009 (.01)	.020 (.010)
Hard left majority			−.050 (.017)	−.007 (.009)
Janata majority			.008 (.026)	−.020 (.033)
Regional majority			.006 (.009)	.026 (.023)
State-specific trends	No	No	No	Yes
Adjusted R^2	.93	.93	.94	.95

- Controlling the group specific time trend- thus the long-term propensity of pro-labor of the states- makes the estimate to zero.

Within control group – DDD(Triple D)

- More convincing analysis sometime comes from higher-order contrasts: **DDD** or **Triple D** design.
  - Build the third dimension of contrast to eliminate the potential bias.
- e.g: Minimum Wage
  - Treatment group: Low-wage-workers in NJ.
  - Control group 1: High-wage-workers in NJ.
  - Assumption 1： the low wage group would have the same trends as high wage group if there were not the new law.
  - Control group 2: Low-wage workers in PA.
  - Assumption 2： the low wage group in NJ would have the same trends as those in PA if there were not the new law.
- It can loose the simple *common trend* assumption in simple DID.

Within control group – DDD(Triple D)

- Jonathan Gruber (1994), "The Incidence of Mandated Maternity Benefits", *American Economic Review*
  - Topic: how the *mandated maternity* benefits affects female's wage and employment.
  - Several state government passed the law that mandated childbirth be covered comprehensively in health insurance plans.
  - Dependent Variable: log hourly wage
  - Independent Variable: mandated maternity benefits law
- Econometric Method: **Triple D**
  1. DID estimates for treatment group (women of childbearing age) in treatment state v.s. control state before and after law change.
  2. DID estimates for control group (women not in childbearing age) in treatment state v.s. control state before and after law change.
  3. DDD DDD estimate of the effect of mandated maternity benefits on wage is $(1) - (2)$

Within control group – DDD(Triple D)

- DDD in Regression

$$Y_{isct} = \beta D_{sct} + \alpha_s + \gamma_c + \delta_t + \lambda_{1st} + \lambda_{2sc} + \lambda_{3ct} + \Gamma X'_{icst} + u_{isct}$$

- α_s:a set of dummies indicating whether or not treatment state
- δ_t: a set of dummies indicating whether or not law change
- γ_c: a set of dummies indicating whether or not women of childbearing age

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** **(0.026)**		

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** **(0.026)**		

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** **(0.026)**		

The Event Study Design: Including Leads and Lags

- If you have a multiple years panel data, then including leads into the DD model is an easy way to analyze pre-treatment trends.
- Lags can be also included to analyze whether the treatment effect changes over time after assignment.
- The estimated regression would be

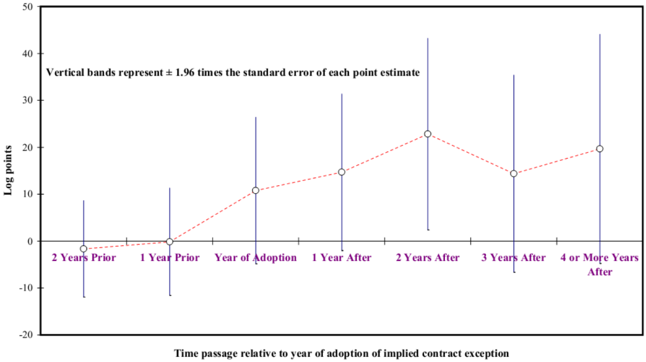
$$Y_{its} = \alpha_s + \delta_t + \sum_{\tau=-q}^{-1} \theta_\tau D_{st} + \sum_{\tau=0}^{p} \delta_\tau D_{st} + X_{ist} + u_{its}$$

- Treatment occurs in year 0
- Includes q leads or anticipatory effects
- Includes p leads or post treatment effects

Study including leads and lags – Autor (2003)

- Autor (2003) includes both leads and lags in a DD model analyzing the effect of increased employment protection on the firm's use of temporary help workers.
- In the US employers can usually hire and fire workers at will.
- U.S labor law allows 'employment at will' but in some state courts have allowed a number of exceptions to the doctrine, leading to lawsuits for 'unjust dismissal'.
- The employment of temporary workers in a state to dummy variables indicating state court rulings that allow exceptions to the employment-at-will doctrine.
- The standard thing to do is normalize the adoption year to 0
- Autor(2003) then analyzes the effect of these exemptions on the use of temporary help workers.

Study including leads and lags – Autor (2003)



- The leads are very close to 0: Common trends assumption may hold.
- The lags show that the effect increases during the first years of the treatment and then remains relatively constant.

Standard errors and Other Threats

Standard errors in DD strategies

- Many paper using DD strategies use data from many years: not just 1 pre and 1 post period.
- The variables of interest in many of these setups only vary at a group level (say a state level) and outcome variables are often serially correlated.
- In the Card and Krueger study, it is very likely that employment in each state is not only correlated within the state but also serially correlated.
- As Bertrand, Duflo and Mullainathan (2004) point out, conventional standard errors often severely *understate* the standard deviation of the estimators – standard errors are biased downward.

Standard errors in Practice

- Simple solution:
  - Clustering standard errors at the group level,but the number of groups does matter.
  - It may also cluster at both the group level and time level.
- Other solutions: Bootstrapping

Other Threats to Validity

- Non-parallel trends
- Other simultaneous shock
- Functional form dependence
- Multiple treatment times(Stagger DID)

Non-parallel trends

- Often policymakers will select the treatment and controls based on pre-existing differences in outcomes – practically guaranteeing the parallel trends assumption will be violated.
- "Ashenfelter dip"
  - Participants in job trainings program often experience a "dip" in earnings just prior to entering the program.
  - Since wages have a natural tendency to mean reversion,comparing wages of participants and non-participants using DD leads to an upward biased estimate of the program effect.

DD with multiple treatment times

- What happens if we have treated units who get treated at different times?
- The simple DID model

$$Y_{ist} = \alpha + \beta D_{st} + \gamma Treat_s + \delta Post_t + \Gamma X'_{ist} + u_{ist}$$

- But now DT_{it} can turn from 0 to 1 at different times for different units.
- **Caution**: this specification gets you a weighted average of several comparisons.This may not be exactly what you want!

Function Forms

- So far our specifications of DID regression equation is linear, but what if it is wrong?
- Several nonparametric or semi-parametric methods can be used
  - Matching DID: Propensity Score Matching and Kernel Density Matching DID
  - Semiparametric DID

Checks for DD Design

- Very common for readers and others to request a variety of “robustness checks” from a DID design.
- Think of these as along the same lines as the leads and lags
  - Falsification test using data for prior periods
  - Falsification test using data for alternative control group(kind of triple DDD)
  - Falsification test using alternative “placebo” outcome that should not be affected by the treatment

Summary

Wrap up

- Difference-in-differences is a special case of fixed effect model with much more powers in our toolbox to make causal inference.
- The key assumption is common trend which is not easy to testify using data.
- Noting that using the right way to inference the standard error.

Extensions of DID: Synthetic Controls

Introduction

Basic Idea

- The **synthetic control method**(SCM) were originally proposed in Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the aim to estimate the effects of aggregate interventions,
- Interventions that are implemented at an aggregate level affecting a small number of large units (such as a cities, regions, or countries), on some aggregate outcome of interest.
- The basic idea behind synthetic controls is that a combination of units often provides a better comparison for the unit exposed to the intervention than any single unit alone.
- It is a data-driven procedure to use a small number of non-treated units to build the suitable counterfactuals.

Introduction

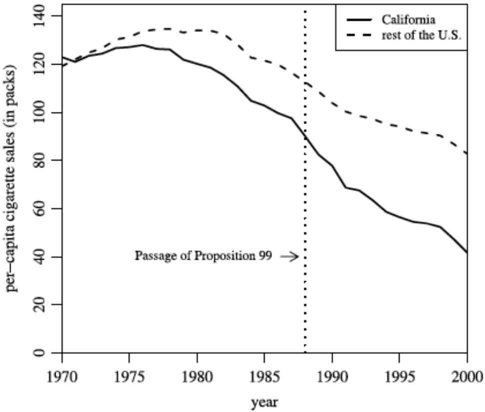
- Synthetic control has been called the most important innovation in causal inference of the last 15 years(Athey and Imbens 2017).
- It is useful for case studies, which is nice because that is often all we have.
- Continues to also be methodologically a frontier for applied econometrics and is widely used in many field, even outside academia.

Extensions of DID: Synthetic Controls Method

- The basic idea is use (long) longitudinal data to build the **weighted average of non-treated** units that best reproduces characteristics of the treated unit over time in pre-treatment period.
- The weighted average of non-treated units is the **synthetic cohort**.
- Causal effect of treatment can be quantified by a simple difference after treatment: *treated vs synthetic cohort*.

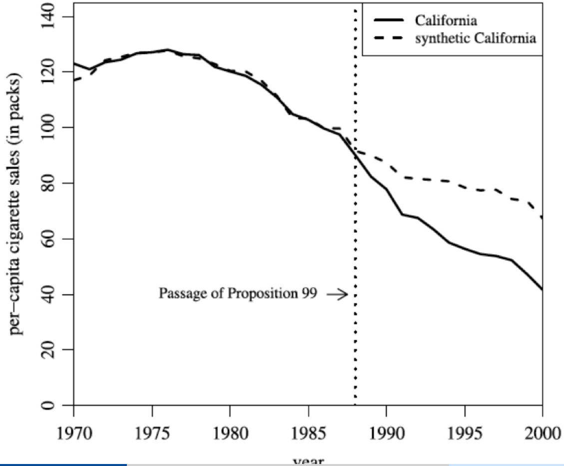
Abadie et.al(2010): Tax on Cig-Consumption

- In 1988, California passed comprehensive tobacco control legislation: Increased cigarette taxes by $0.25 per pack ordinances.
- It estimates the effect of the policy on cigarette consumption.



Abadie et.al(2010): Tax on Cig-Consumption

- Using 38 states that had never passed such programs as controls: **Synthetic CA**



Predictor Means: Actual vs Synthetic California

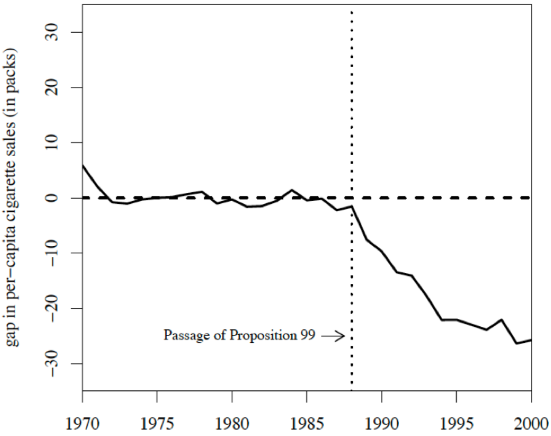
- Most observables are similar between Actual and Synthetic

Variables	California Real	California Synthetic	Average of 38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

The Application: Actual vs Synthetic California

- The treatment effect is measured by the gap in ciga-sales between Actual and Synthetic



Formalization

Formalization: The Setting

- Suppose that we obtain data for $J+1$ units: $j = 1, 2, ..., J+1$
  - Assume that the first unit (j = 1) is the **treated unit**, that is, the unit affected by the policy intervention of interest.
  - Then the set of potential comparisons,$j = 2, ..., J+1$ is a collection of **untreated units**, not affected by the intervention.
- Assume also that our data span T periods and that the first T_0 periods are before the intervention.
- Let Y_{jt} and Y_{jt}^N be the real and potential outcomes of interest for unit j of $J+1$ aggregate units at time t with and without intervention.
- the effect of the intervention of interest for the affected unit in period $t(t > T_0)$

$$\tau_{1t} = Y_{1t} - Y_{1t}^N$$

Formalization: The Setting

- How to reproduce Y_{1t}^N which is totally unobservable? Use unaffected units in control groups to predict it.
- More specifically, a weighted average of the units in the comparison group use to construct the potential outcome of treated units, which define as **synthetic control**.Thus,

$$\hat{Y}_{1t}^N = \Sigma_{j=2}^{J+1} w_j Y_{jt}$$

- Then the question is how to determine these values of the weights, w_j

Formalization: Weights

- Let more specifically, $W = (w_2, ..., w_{J+1})'$ have to satisfy two restriction conditions
  - $w_j \geq 0$ for $j = 2, ..., J+1$
  - $\Sigma_{j=2}^{J+1} w_j = 1$
- how to determine these values of the weights, w_j
  - The simplest way: assigning equal weights, thus

$$w_j = \frac{1}{J}$$

  - Or a population weighted version is a fraction of the total population in the comparison group(at the time of the intervention),thus

$$w_j = \frac{N_j}{\Sigma_{j=2}^{J+1} N_j}$$

Formalization: Weights

- For each unit, j, we also observe a set of characteristics which can be use to predict the outcome Y_{jt}, denoted as $X_{1j}, ...X_{kj}$
- Let X_1 is a $k \times 1$ vector of these characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- Abadie et. al (2010) proposes that we can determine the value of w_j^* by using Matching method,which is a re-weighted method in nature.
- Let X_1 is a $k \times 1$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.

Matching Estimator

- Suppose we have treated and untreated groups but the here assignment is not random. Then we can't obtain the causal effect δ directly by
$$E(Y_1|D=1)-E(Y_0|D=0)$$
for the presence of selection bias.
- The idea of matching method is quite simple. **What if we can construct a reasonable "control" group by selecting some(or all) samples in untreated group** then we can estimate the treatment effect
$$\hat{\delta}=\frac{1}{N_T}\Sigma_{i=1}^{N_T}(Y_i-Y_i^c)$$
    - N_T is the sample size in treatment group
    - Y_i^c is the corresponding counterfactual outcomes by matching(selecting) the sample in untreated group.

Matching Estimator: an example

	Potential Outcome			
unit	under Treatment	under Control		
i	Y_i^1	Y_i^0	D_I	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

- the only covariates is X, which is used to select the “proper” counterfactuals

Matching Estimator: an example

	Potential Outcome			
unit	under Treatment	under Control		
i	Y_i^1	Y_i^0	D_I	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

- Then

$$\hat{\delta} = \frac{1}{3}[(6-9)+(1-0)+(0-9)] = -3.7$$

Matching Estimator

- But what if we have multiple covariates using to match, thus $X = (X_1, X_2, ...X_k)'$
- If $X = (x_1, x_2, ...x_k)$ is a k-class vector, then the **distance** to measure "closeness" or "similarity" between two vectors such as X_i and X_j is the **Euclidean distance**

$$\| (X_i - X_j) \| = \sqrt{(X_i - X_j)'(X_i - X_j)} = \sqrt{\Sigma_{n=1}^{k}(X_{ni} - X_{nj})^2}$$

Normalized Euclidean distance

- The *Euclidean distance* is not invariant to changes in the scale of the X's. For this reason, alternative distance metrics that are invariant to changes in scale are used.
- A commonly used distance is the **normalized Euclidean distance**:

$$\| (X_i - X_j) \| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$$

- where V is some $(k \times k)$ symmetric and positive semidefinite matrix. More specifically,

$$\hat{V}^{-1} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}_k^2 \end{pmatrix}$$

Formalization: Weight by Matching

- The rule to choose the optimal weight vector $W^* = (w_2, ..., w_{J+1})'$ will be

$$argmin_W \parallel (X_1 - X_0 W) \parallel$$

- Thus,the optimal vector should **minimize the "distance"** between treated unit and unaffected group,subject to two weight constraints.
- More specifically, *Abadie, et al(2010)* consider

$$\parallel (X_1 - X_0 W) \parallel_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

where V is some $(k \times k)$ symmetric and positive semidefinite matrix.

Formalization: More on the V matrix

- Typically, V is diagonal with main diagonal $v_1, ..., v_k$. Then the synthetic control weights minimize

$$\sum_{m=1}^{k} v_m \big(X_{1m} - \sum_{j=2}^{J+1} w_j^* X_{jm}\big)^2$$

- Where v_m is a **weight** that reflects the *relative importance* that we assign to the m^{th} variable when we measure the discrepancy between the treated unit and the synthetic controls.
- And v_m is critical because it weights directly shape w_j,which help reproducing the counterfactual outcome for the treated unit in the absence of the treatment.

Formalization: Estimating the V matrix

- Various ways to choose V
  - In practice,most people choose V that minimizes *the mean squared prediction error(MSPE)*.Thus,

$$\sum_{t=1}^{T_0}\left(Y_{1t}-\sum_{j=2}^{J+1}w_j^*(V)Y_{jt}\right)^2$$

- If the number of pre-intervention periods in the data is “large”, then matching on pre-intervention outcomes can allow us to control for the heterogeneous responses to multiple unobserved factors.
- The intuition here is that only units that are alike on unobservables and unobservables would follow a similar trajectory pre-treatment.

A Machine learning procedure

1. Divide the pre-intervention periods(T_0) into a initial **training** period($t = 1, ...t_0$) and a subsequent **validation** period($t = t_0 + 1, ...T_0$).
2. Select a value V^* make the MSPE is small

$$\sum_{t=t_0+1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j(V)Y_{jt}\right)^2$$

3. Use the resulting V^* and data on the predictors for the last t_0 before in the intervention,$t = t_0 + 1, t_0 + 2, ..., T_0$ to calculate $w^* = w(V^*)$

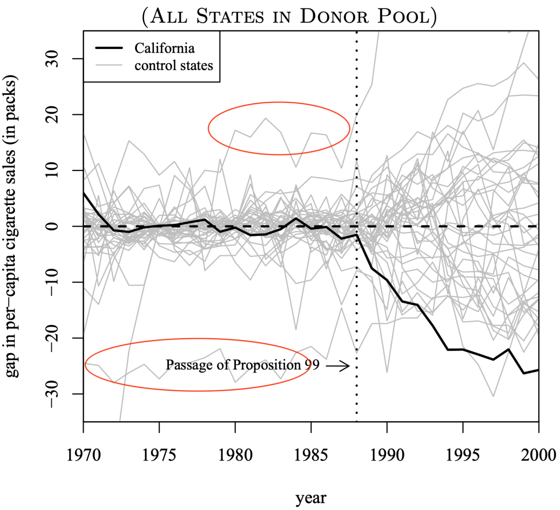
Inference

- Permutation Strategy: whether the effect estimated by the synthetic control for the unit affected by the intervention is **large** relative to the effect estimated for a unit chosen at random.
- Implementation: "randomization" of the treatment to each unit, re-estimating the model, and calculating a set of root mean squared prediction error (RMSPE) values for the pre- and post-treatment period.
- For $0 \leq t_1 \leq t_2 \leq T$ and $j = 1, 2, ..., J+1$,let

$$R_j(t_1, t_2) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (Y_{jt} - \hat{Y}_{jt}^N)^2\right)^{\frac{1}{2}}$$

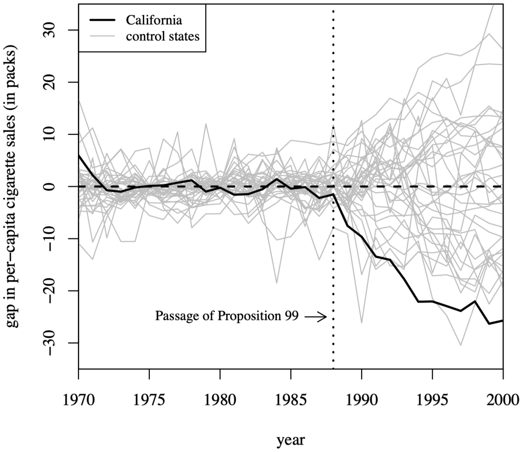
- Some states whose pre-treatment RMSPE is considerably different than California's can be dropped.

Inference: Dropping Sample



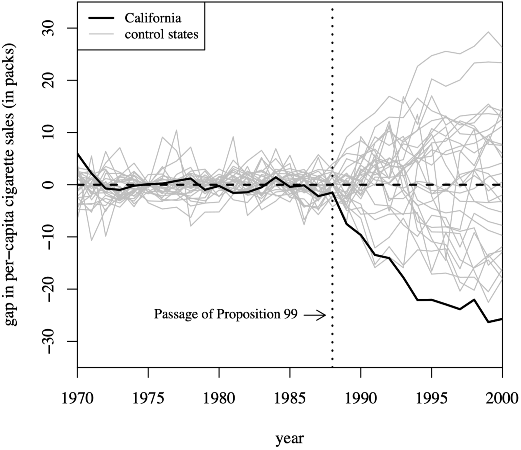
Inference: Dropping Sample

(PRE-PROP. 99 MSPE $\leq$ 20 TIMES PRE-PROP. 99 MSPE FOR CA)



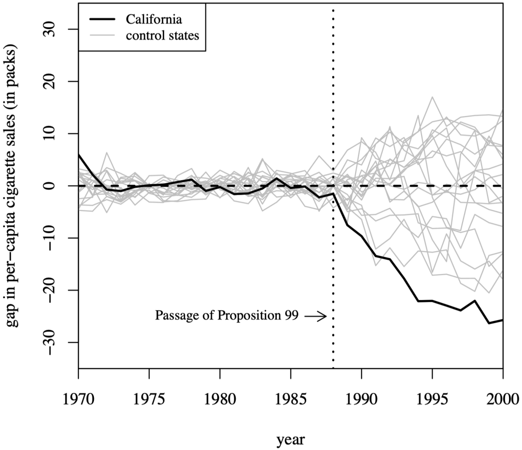
Inference: Dropping Sample

(PRE-PROP. 99 MSPE $\leq$ 5 TIMES PRE-PROP. 99 MSPE FOR CA)



Inference: Dropping Sample

(Pre-Prop. 99 MSPE ≤ 2 Times Pre-Prop. 99 MSPE for CA)



Inference: Procedure

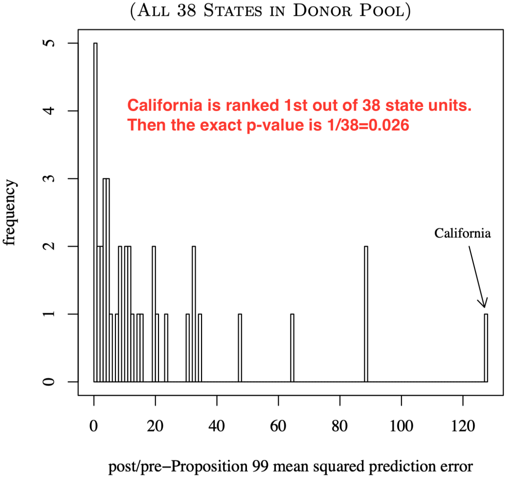
1. Iteratively apply the synthetic method to each state in the unaffected group and obtain a distribution of placebo effects.
2. Calculate the RMSPE(root mean squared prediction error) for *each placebo* for the pre-treatment and post-treatment.
   - Post-treatment $R_{j,post} = RMSPE_j(T_0 + 1, T)$
   - Pre-treatment $R_{j,pre} = RMSPE_j(1, T_0)$
3. Compute the ratio of the post-to-pre-treatment and sort it in descending order from greatest to highest. Thus

$$r_j = \frac{R_{j,post}}{R_{j,pre}}$$

4. The exact p-value is defined as

$$p - value = \frac{rank_{th}}{J + 1}$$

Inference: P-Value

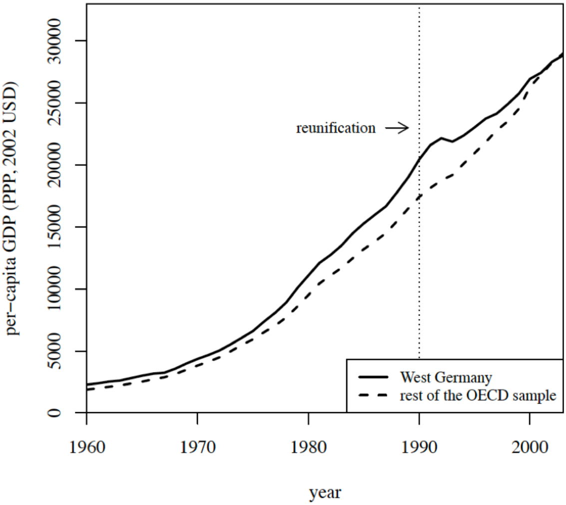


An Application: The 1990 German Reunification

Topic: The Economic Effect of the German Reunification on West Germany

- Cross-country regressions are often criticized because they put side-by-side countries of very different characteristics.
    - "What do Thailand, the Dominican Republic, Zimbabwe,Greece and Bolivia have in common that merits their being put in the same regression analysis? Answer: For most purposes, nothing at all." (Harberger 1987)
- Application: The economic effect of "Berlin Wall" Falling,thus the 1990 German reunification,on West Germany.
- Control group is compositional restricted to 16 OECD countries

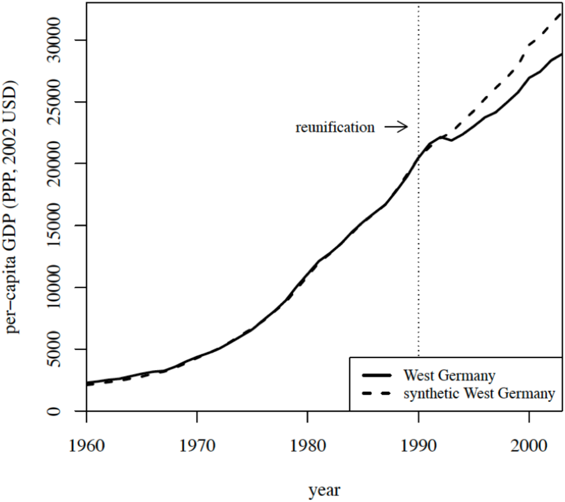
West Germany v.s. OECD



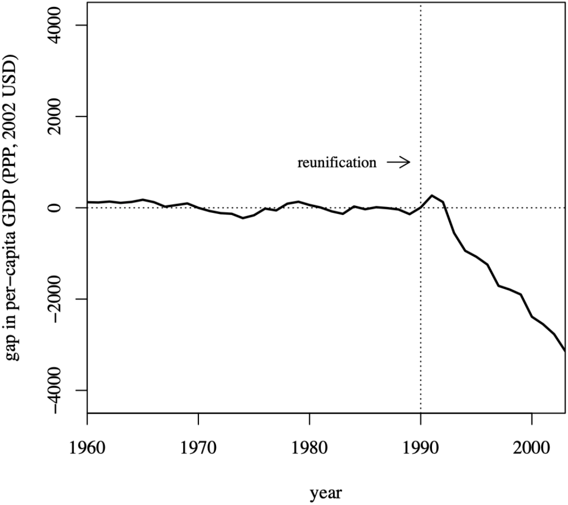
Economic Growth Predictors Means across groups

	West Germany	Synthetic West Germany	OECD Sample
GDP per-capita	15808.9	15800.9	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

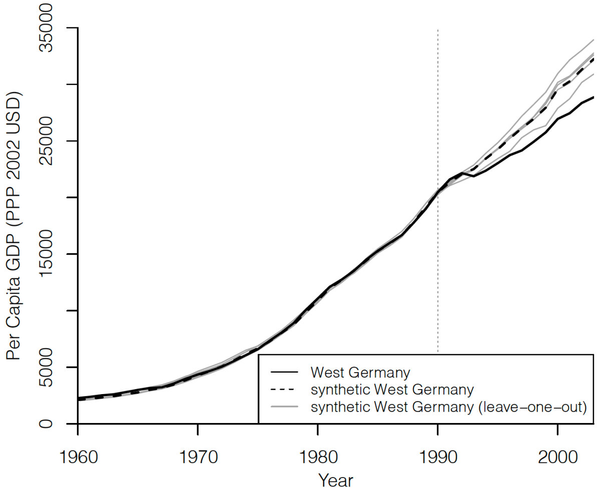
West Germany v.s Sythetic West Germany



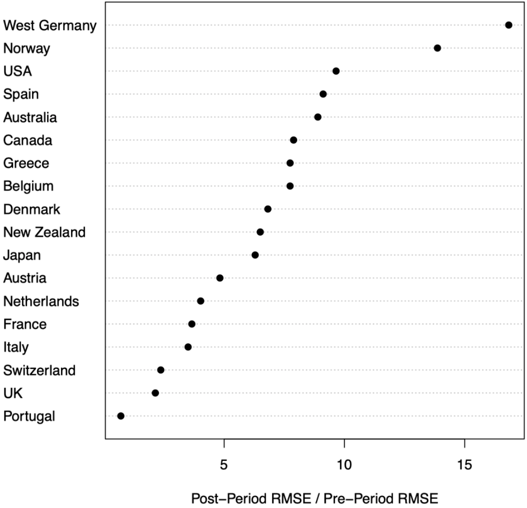
GDP Gap: West Germany and synthetic West Germany



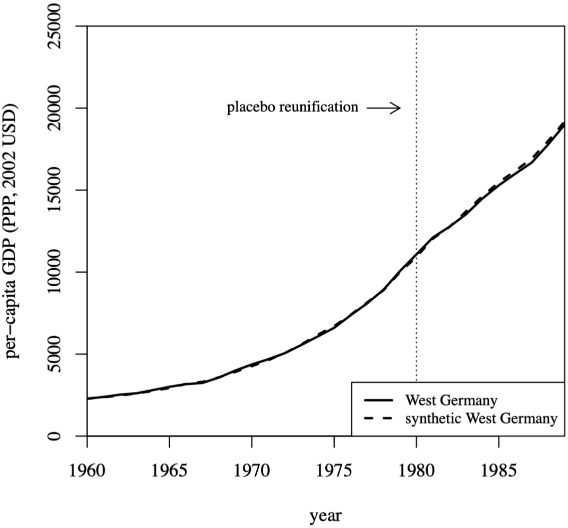
The 1990 German Reunification: Leave-one-out estimates



RMSE Test


West Germany
Norway
USA
Spain
Australia
Canada
Greece
Belgium
Denmark
New Zealand
Japan
Austria
Netherlands
France
Italy
Switzerland
UK
Portugal
5
10
15
Post−Period RMSE / Pre−Period RMSE

Placebo Test: What if '1980' German Reunification



Wrap Up

- **Synthetic control method** provide many practical advantages for causal inference.
- The credibility of the results depends on
  - the level of diligence exerted in the application
  - whether contextual and data requirements are met

Summary for Causal Inference

Final Thoughts(Angrist and Pischeke,2008)

- A good research design is one you are excited to tell people about
  - that's basically what characterizes all research designs, whether **instrumental variable**,**regression discontinuity designs** or **difference-in-differences**,**synthetic control method** among others(**Seven Magic Weapons**).
- Causality is *easy and hard*. Don't get confused which is the hard part and which is the easy part.
- Always understand *what assumptions you must make*, be clear which parameters you are and are not identifying.
- Last but not least, Remember: **Good question is always the first priority**. Along with good research design is in the second place.

Though still a long way to go but now we could take a break and enjoy the landscape.