Difference-in-Differences Design

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

March 26, 2026

Causal Inference
Control-Based v.s. Design-Based

Control-Based Causal Inference

- So far, we have learned several control-based causal inference methods
  - Matching, regression, and causal machine learning
- These methods are all based on CIA (**selection on observables**)
  - Assumed all confounding factors can be observed
  - Thus, we can eliminate selection bias by comparing the treated and untreated units with the similar observed characteristics

Unobservable Omitted Variable

- If unobservabale confounding factors are time-invariant or common across units
  - We can include **fixed effects** into regression to get causal effects
- Yet, what if important confounding factors are unobserved and time-varying?

Design-Based Causal Inference

- Next four weeks, we will learn several methods to deal with unobservable omitted variables
  - Difference-in-differences design
  - Synthetic control method
  - Instrumental variable design
  - Regression discontinuity design
- The above methods utilize an exogenous factor that drives change in treatment status to estimate the causal effect

Main Idea

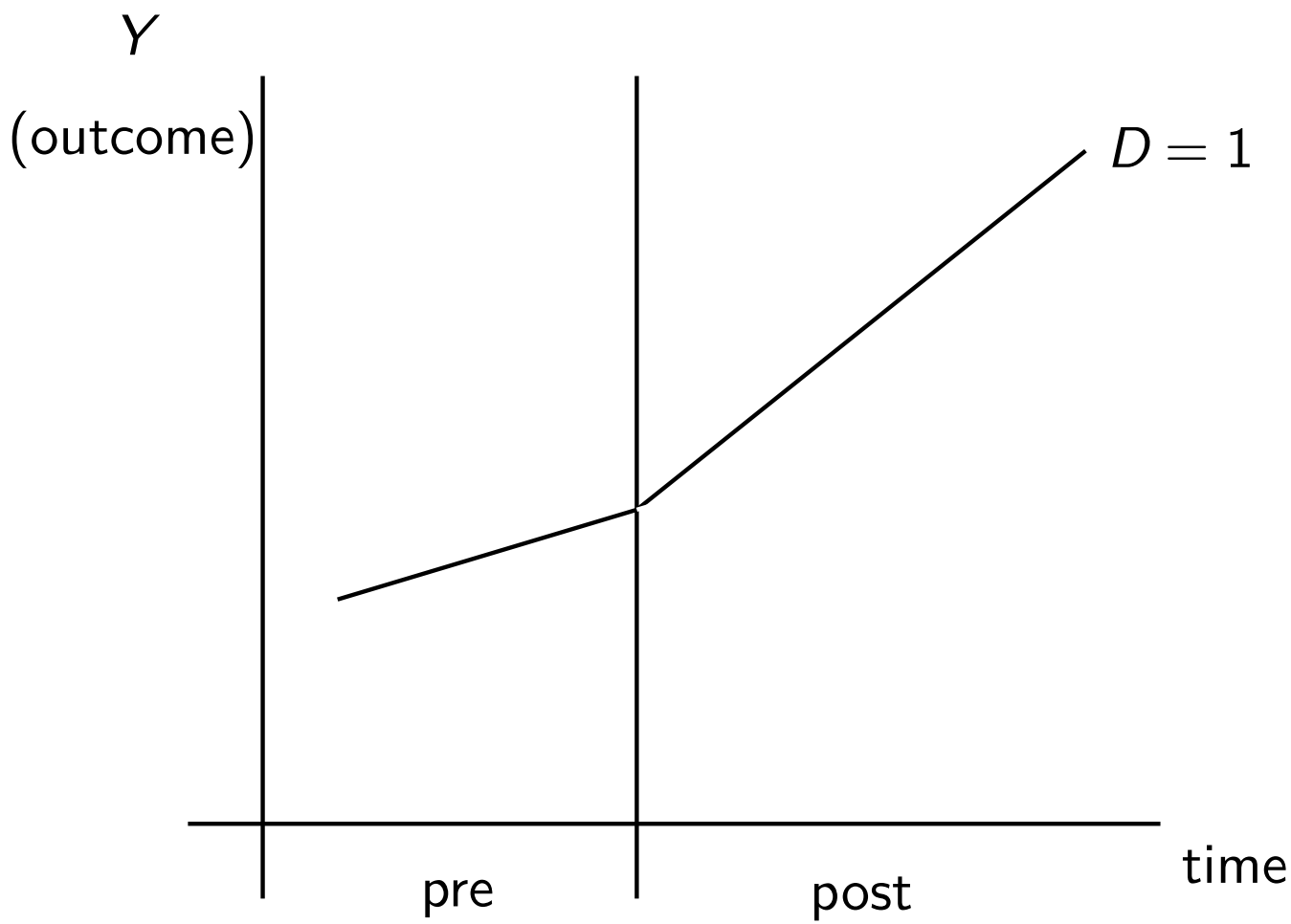
Difference-in-Differences Design

Main Idea

- If we can observe **group-level** outcomes at multiple time points,
  - Especially before and after the treatment,
- And if we assume that, **in the absence of treatment**, the outcomes of the treatment and control groups would have followed **parallel trends**,
- Then we can construct the **counterfactual trend for the treatment group** using
  - The **observed trend in the control group**.
- By comparing the observed trend in the treatment group with its counterfactual trend, we can estimate the **causal effect of the treatment**.

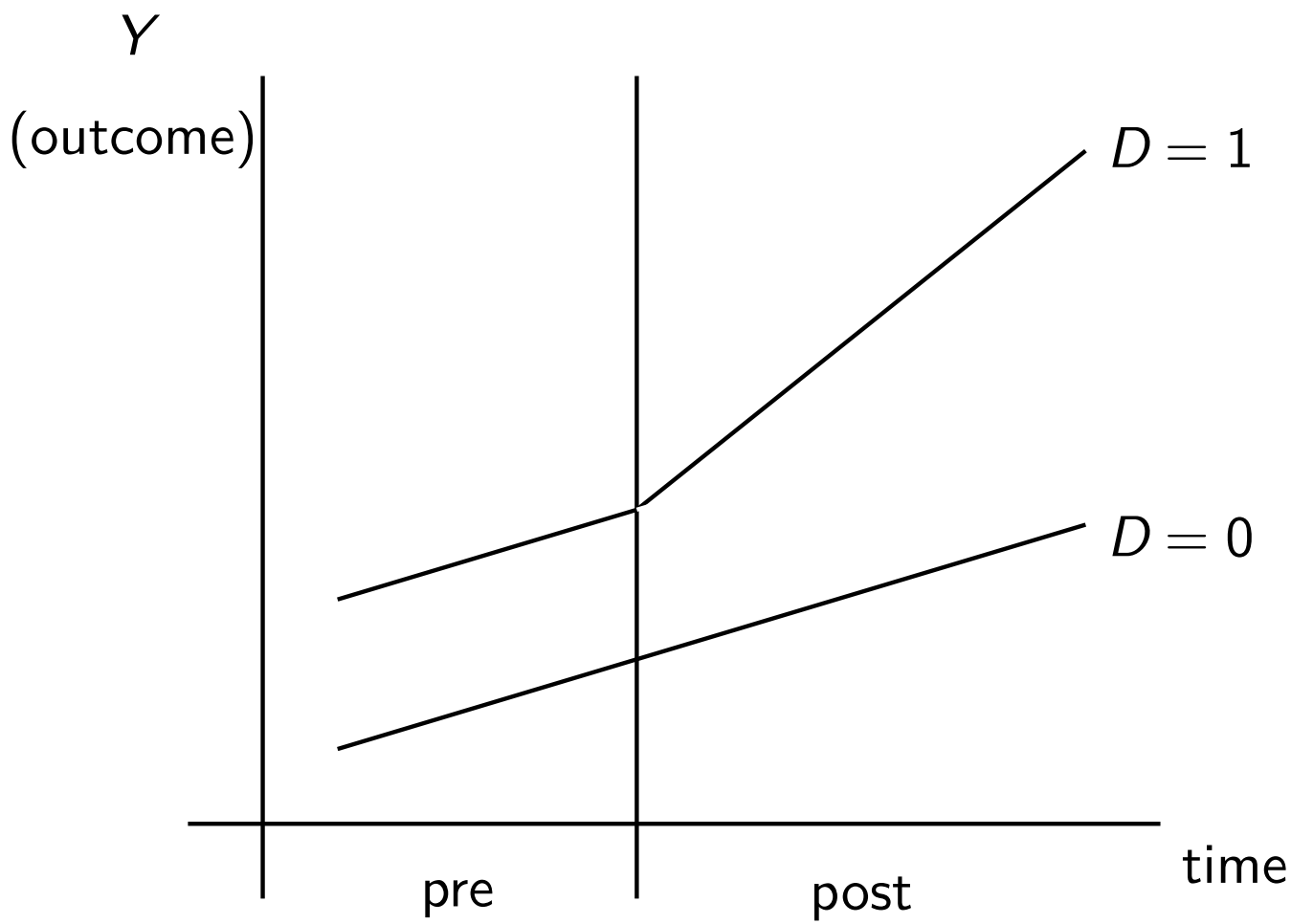
Main Idea

Graph



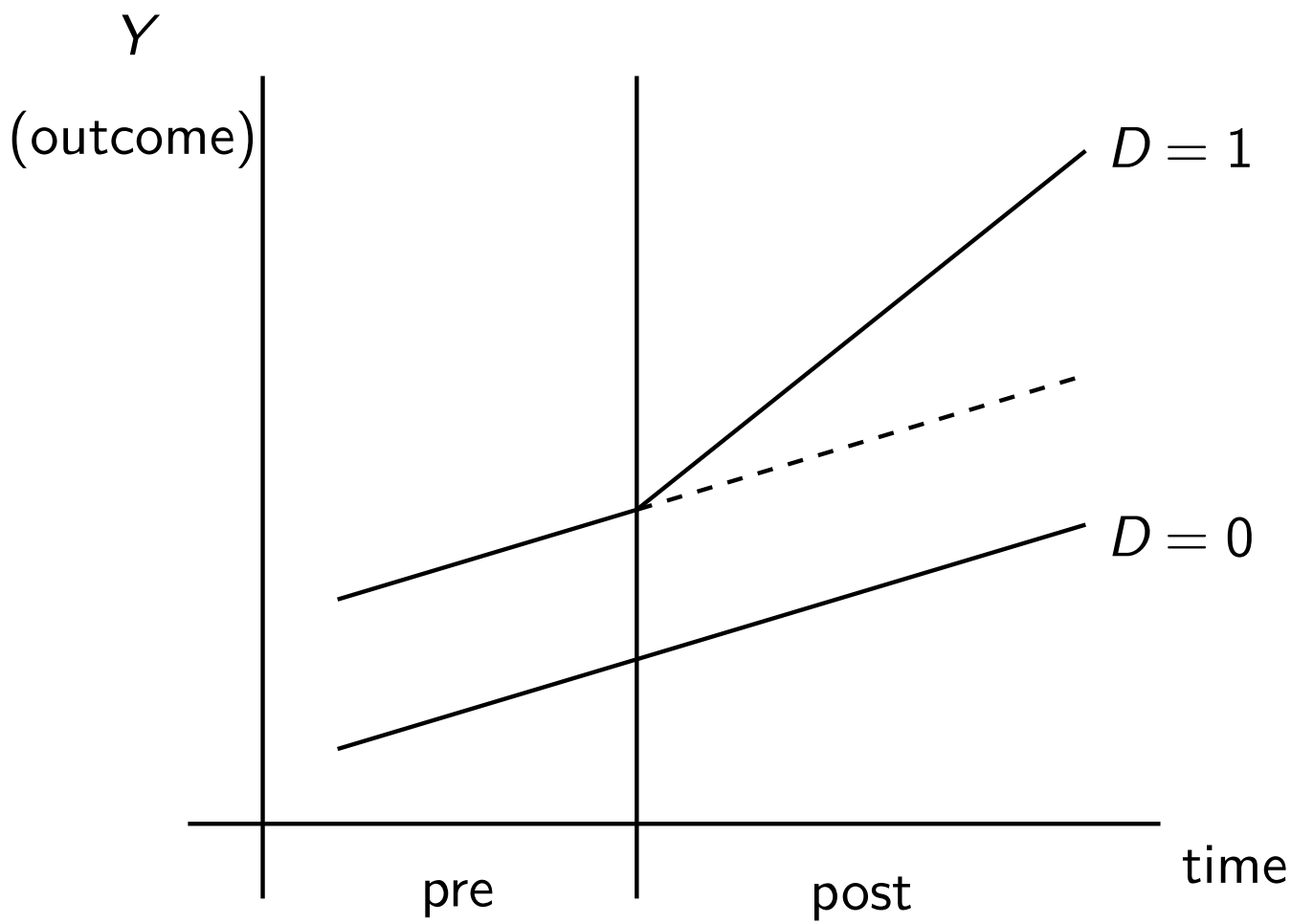
Main Idea

Graph



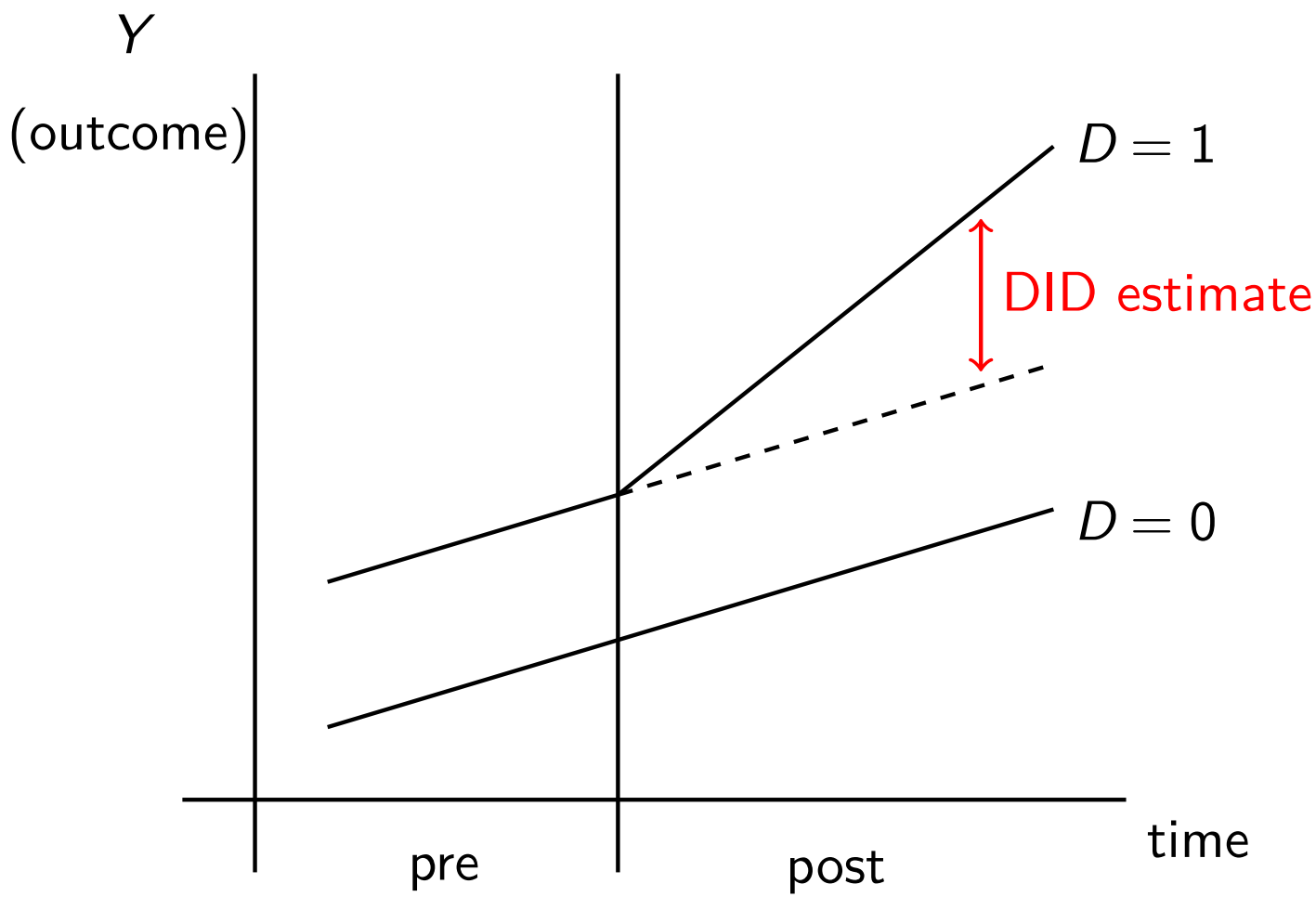
Main Idea

Graph



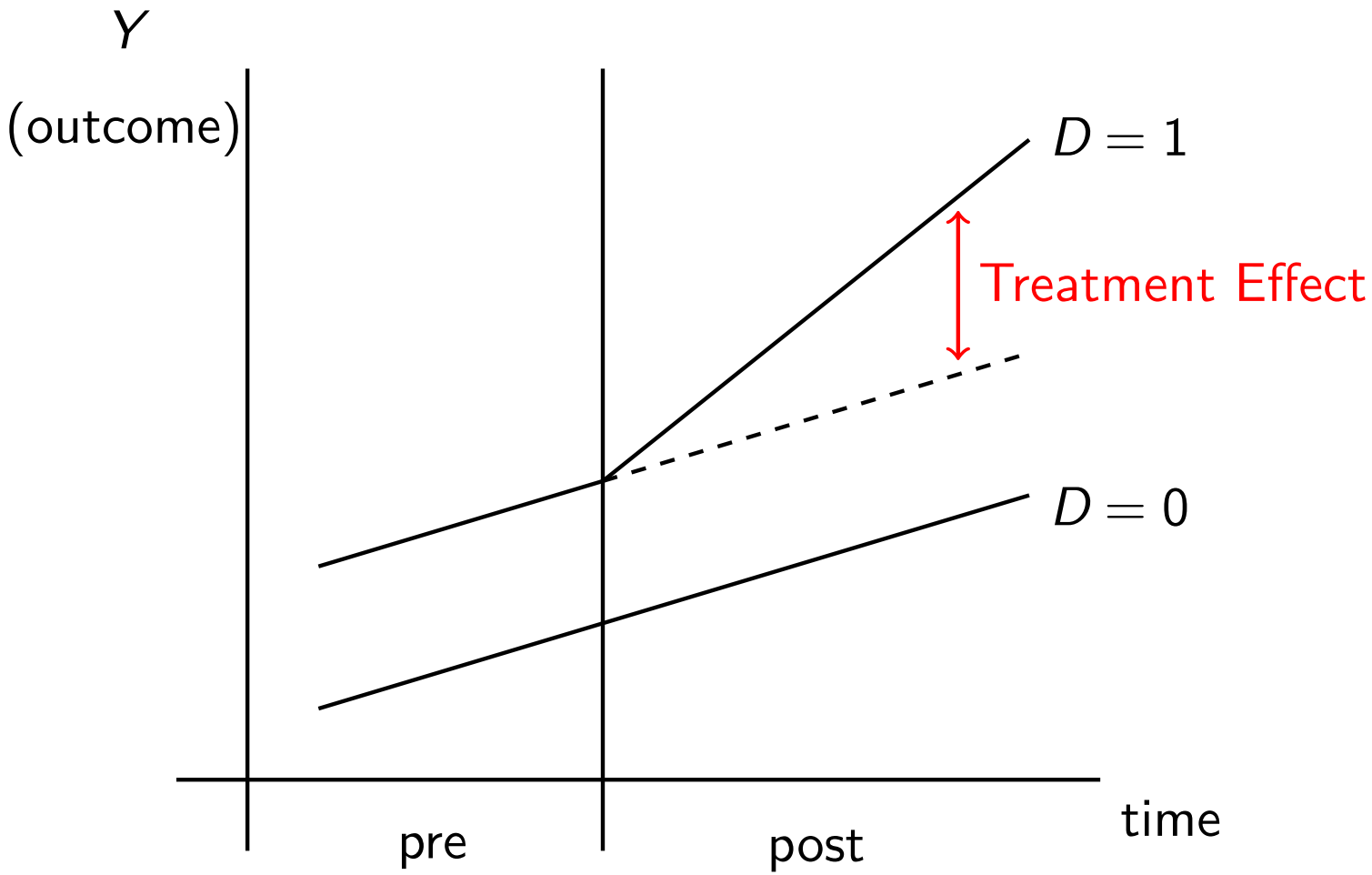
Main Idea

Graph



Main Idea

Graph



A Motivating Example

Card & Krueger (1994)

David Card and Alan B. Krueger (1994) "**Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania**" AER

- They want to estimate the causal effect of raising minimum wage on employment of low-skilled workers

A Motivating Example

Card & Krueger (1994)

- What is the effect of increasing the minimum wage on employment?
- Minimum wage is effective only in certain jobs:
  - Low-skilled jobs
- How much does an increase in the minimum wage reduce demand for low-skilled workers?
  - In a competitive labour market, increases in the minimum wage would move up a downward-sloping labour demand curve.
  - Employment would fall.

A Motivating Example

Card & Krueger (1994)

- Card & Krueger (1994) analyse the effect of a minimum wage increase in New Jersey (NJ) using a DID methodology
- In February 1992 NJ increased the state minimum wage from \$4.25 to \$5.05
- Pennsylvania (PA)'s minimum wage stayed at \$4.25.

- They surveyed about 400 fast food stores both in NJ and in PA both before and after the minimum wage increase in NJ.

A Motivating Example

Card & Krueger (1994)

- Two groups:
  - Treatment group: NJ
  - Control group: PA
- Two periods:
  - Pre-treatment period: February 1992
  - Post-treatment period: November 1992
- Let Y_{st} denote the average employment in state s at time t

A Motivating Example

Card & Krueger (1994)

- To estimate the effect of minimum wage on employment in NJ, we would like to know the following counterfactual:
  - **In absence of raising minimum wage to \$5.05**, what the average employment level in NJ would be ?
- DID design suggests us construct the **counterfactual employment in NJ** by using:
  - Average employment level in NJ before reform +
  - The trend in average employment level in PA (control group)

$$Y_{NJ,Feb} + (Y_{PA,Nov} - Y_{PA,Feb})$$

A Motivating Example

Card & Krueger (1994)

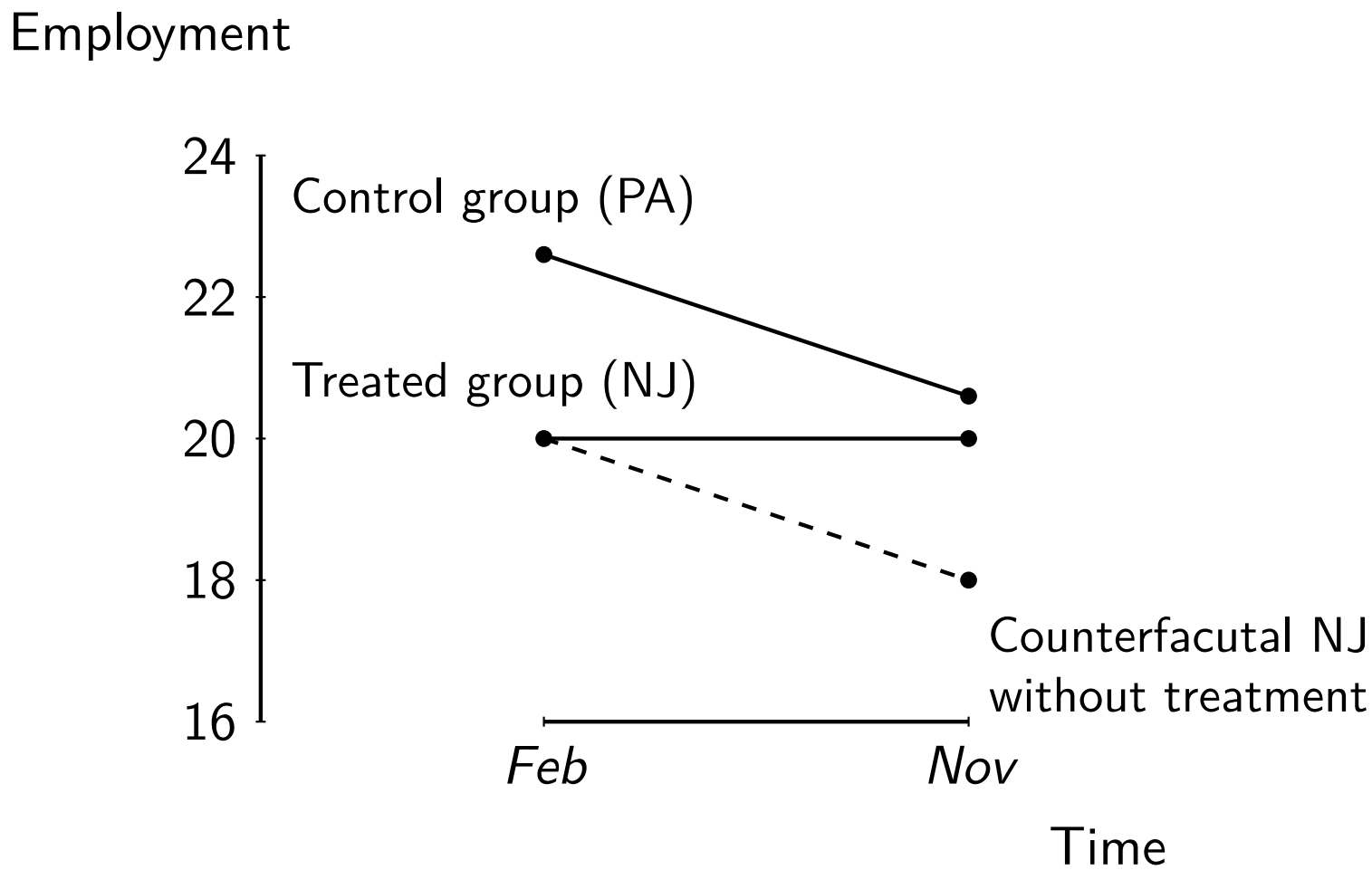
- We can identify the effect of minimum wage on employment in NJ by taking difference in **realized employment** and **counterfactual employment** in NJ:

$$\begin{aligned}\alpha_{DID} &= Y_{NJ,Nov} - [Y_{NJ,Feb} + (Y_{PA,Nov} - Y_{PA,Feb})] \\ &= (Y_{NJ,Nov} - Y_{NJ,Feb}) - (Y_{PA,Nov} - Y_{PA,Feb})\end{aligned}$$

- If PA is a good control group:
  - The trend in employment rate of PA should absorb any other changes in employment that are unrelated to increase minimum wage

A Motivating Example

Card & Krueger (1994)



A Motivating Example

Card & Krueger (1994)

Variable		Stores by state: PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1.	Mean employment at February 1992	23.33	20.44	-2.89
		(1.35)	(0.51)	(1.44)
2.	Mean employment at November 1992	21.17	21.03	-0.14
		(0.94)	(0.52)	(1.07)
3.	Change in mean employment between Feb and Nov	-2.16	0.59	2.76
		(1.25)	(0.54)	(1.44)

- Surprisingly, employment rose in NJ relative to PA after the minimum wage change.

A Motivating Example

Card & Krueger (1994)

$$\begin{aligned}\alpha_{DID} &= (Y_{NJ,Nov} - Y_{NJ,Feb}) - (Y_{PA,Nov} - Y_{PA,Feb}) \\ &= (21.03 - 20.44) - (21.17 - 23.33) \\ &= 0.59 - (-2.16) = 2.76\end{aligned}$$

- Instead of comparing the employment of NJ in February (before reform) and November (after reform)
- DID suggests we need to adjust for change (trend) in labor demand when there was no increase in minimum wage

Identification

Potential Outcomes Framework

- Basic setup: two time periods, two groups
- Two periods
  - In period $t = 1$: one of the groups is treated
  - In period $t = 0$: neither group is treated
- Two groups
  - $D_i = 1$: those that are treated at $t = 1$ (treatment group)
  - $D_i = 0$: those that are always untreated (control group)

Potential Outcomes Framework

- Potential Outcomes
  - Y_{it}^1: the potential outcome for unit i if he would receive treatment at time t
  - Y_{it}^0: the potential outcome for unit i if he would NOT receive treatment at time t

Potential Outcomes Framework

- Observed Outcomes
  - Y_{it} is the observed outcome for unit i at time t
    - Observed outcomes at $t = 0$:

$$Y_{i0} = \mathrm{Y}_{i0}^{0}$$

    - Observed outcomes at $t = 1$:

$$Y_{i1} = \mathrm{Y}_{i1}^{0}(1 - D_i) + \mathrm{Y}_{i1}^{1} D_i$$

Identification Results for DID

- Our main interest is average treatment effect on treated (ATT):

$$\alpha_{\text{ATT}} = E[\text{Y}_{i1}^1 - \text{Y}_{i1}^0 | D_i = 1]$$

- Missing data problem: $E[\text{Y}_{i1}^0 | D_i = 1]$ is unknown
- DID design can help us identify ATT if Parallel Trends Assumption holds

Identification Results for DID

Identification Assumption

Parallel Trends Assumption

$$E[\mathrm{Y}_{i1}^0 - \mathrm{Y}_{i0}^0 | D_i = 1] = E[\mathrm{Y}_{i1}^0 - \mathrm{Y}_{i0}^0 | D_i = 0]$$
$$= E[\mathrm{Y}_{i1} - \mathrm{Y}_{i0} | D_i = 0]$$

- The treatment group and control group would have exhibited the same trend in the absence of the treatment
- We can use Parallel Trends Assumption to construct a counterfactual for treatment group at $t = 1$

$$\mathrm{E}[\mathrm{Y}_{i1}^0 | D_i = 1] = \mathrm{E}[\mathrm{Y}_{i0}^0 | D_i = 1] + \mathrm{E}[\mathrm{Y}_{i1}^0 - \mathrm{Y}_{i0}^0 | D_i = 0]$$
$$= \mathrm{E}[\mathrm{Y}_{i0} | D_i = 1] + \mathrm{E}[\mathrm{Y}_{i1} - \mathrm{Y}_{i0} | D_i = 0]$$

- We can use **observed outcomes** to represent **unobserved** $\mathrm{E}[\mathrm{Y}_{i1}^0 | D_i = 1]$

Identification Results for DID

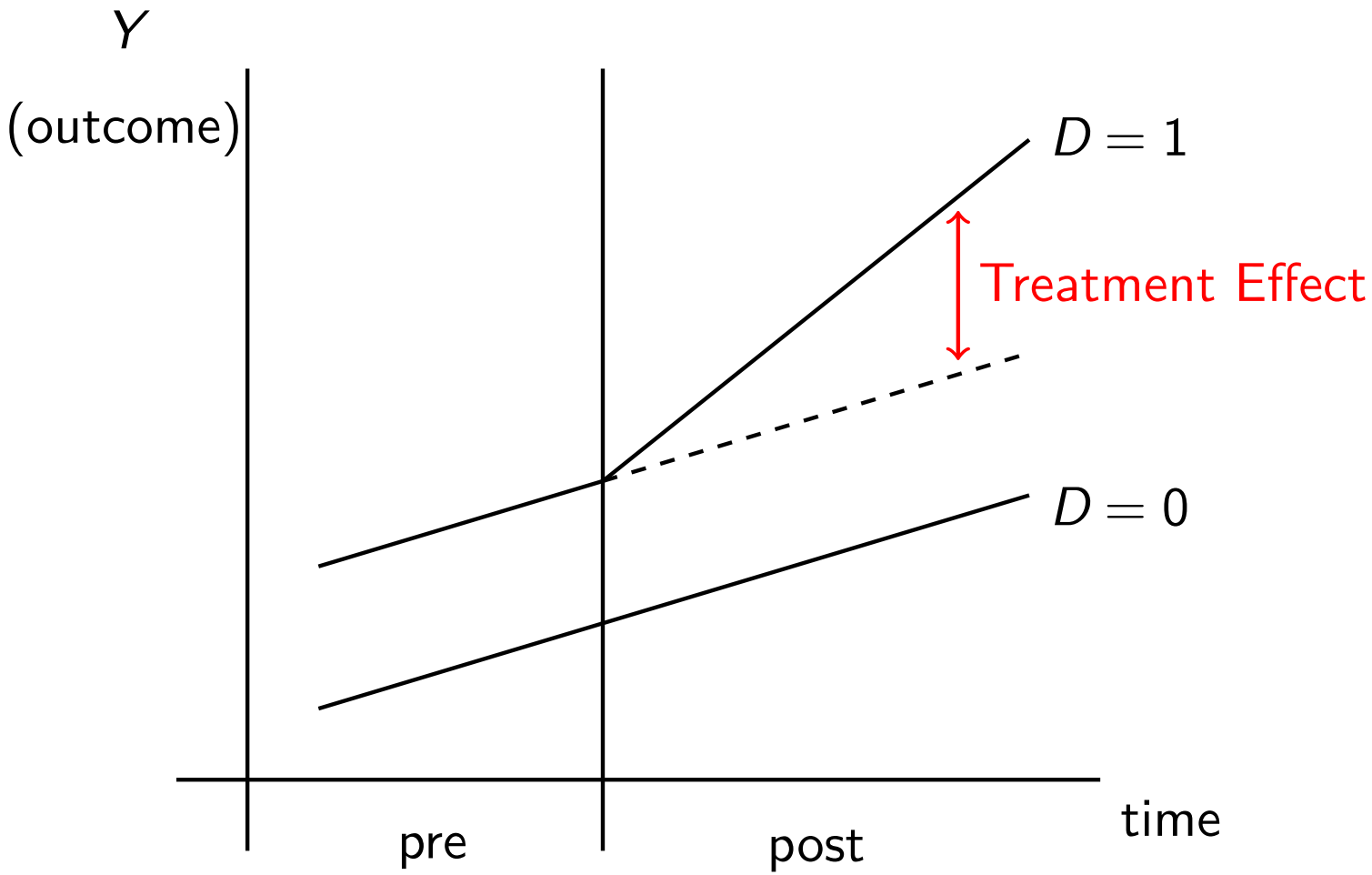
- Apply Parallel Trends Assumption:

$$
\begin{aligned}
\alpha_{\text{ATT}} &= \mathrm{E}[\mathrm{Y}_{i1}^{1} - \mathrm{Y}_{i1}^{0} | D_i = 1] \\
&= \mathrm{E}[\mathrm{Y}_{i1}^{1} | D_i = 1] - \mathrm{E}[\mathrm{Y}_{i1}^{0} | D_i = 1] \\
&= \mathrm{E}[\mathrm{Y}_{i1}^{1} | D_i = 1] - \mathrm{E}[\mathrm{Y}_{i0}^{0} | D_i = 1] - \mathrm{E}[\mathrm{Y}_{i1}^{0} - \mathrm{Y}_{i0}^{0} | D_i = 0] \\
&= \mathrm{E}[\mathrm{Y}_{i1}^{1} - \mathrm{Y}_{i0}^{0} | D_i = 1] - \mathrm{E}[\mathrm{Y}_{i1}^{0} - \mathrm{Y}_{i0}^{0} | D_i = 0] \\
&= \mathrm{E}[\mathrm{Y}_{i1} - \mathrm{Y}_{i0} | D_i = 1] - \mathrm{E}[\mathrm{Y}_{i1} - \mathrm{Y}_{i0} | D_i = 0] = \alpha_{\text{DID}}
\end{aligned}
$$

- The **average treatment effect on treated (ATT)** can be identified by difference in trend of outcome between treatment and control groups

Identification Results for DID

Graphical Interpretation



Estimation

DID Estimation

Basic Two Periods/Groups

- Basic case: two groups and two periods
- We can obtain the DID estimates using a regression framework

$$Y_{it} = \mu + \gamma D_i + \delta Post_t + \alpha(D_i \times Post_t) + \varepsilon_{it},$$

  - D is a dummy indicating treatment group
  - $Post$ is a dummy indicating post-treatment period
  - γ captures differences across groups that are constant over time
  - δ captures differences over time that are common to all groups

DID Estimation

Basic Two Periods/Groups

$$Y_{it} = \mu + \gamma D_i + \delta Post_t + \alpha(D_i \times Post_t) + \varepsilon_{it},$$

- α is the coefficient of interest
  - Capture the different trends in outcome between treatment and control group
- We will show that α can represent the DID estimate:

$$\begin{aligned}\alpha = &\{\mathrm{E}[\mathrm{Y}_{it}|D_i = 1, Post_t = 1] - \mathrm{E}[\mathrm{Y}_{it}|D_i = 1, Post_t = 0]\} \\ &- \{\mathrm{E}[\mathrm{Y}_{it}|D_i = 0, Post_t = 1] - \mathrm{E}[\mathrm{Y}_{it}|D_i = 0, Post_t = 0]\}\end{aligned}$$

DID Estimation

Basic Two Periods/Groups

$$Y_{it} = \mu + \gamma D_i + \delta Post_t + \alpha(D_i \times Post_t) + \varepsilon_{it},$$

- Assume $E[\varepsilon_{it}|D, Post] = 0$
  - **Pre-treatment mean of outcome for control group:** $\mathrm{E}[\mathrm{Y}_{it}|D = 0, Post = 0] = \mu$
  - **Post-treatment mean of outcome for control group:** $\mathrm{E}[\mathrm{Y}_{it}|D = 0, Post = 1] = \mu + \delta$
  - **Pre-treatment mean of outcome for treatment group:** $\mathrm{E}[\mathrm{Y}_{it}|D = 1, Post = 0] = \mu + \gamma$
  - **Post-treatment mean of outcome for treatment group:** $\mathrm{E}[\mathrm{Y}_{it}|D = 1, Post = 1] = \mu + \gamma + \delta + \alpha$

DID Estimation

Basic Two Periods/Groups

- α can represent treatment effect identified by DID design α_{DID}:

$$\begin{aligned}\alpha_{DID} &= \{\mathrm{E}[\mathrm{Y}_{it}|D=1, Post=1] - \mathrm{E}[\mathrm{Y}_{it}|D=1, Post=0]\} \\ &\quad - \{\mathrm{E}[\mathrm{Y}_{it}|D=0, Post=1] - \mathrm{E}[\mathrm{Y}_{it}|D=0, Post=0]\} \\ &= \{(\mu+\gamma+\delta+\alpha) - (\mu+\gamma)\} - \{(\mu+\delta) - \mu\} \\ &= \alpha\end{aligned}$$

DID Estimation

Basic Two Periods/Groups

	Pre	Post	Pre/Post difference
Control Group	μ	$\mu + \delta$	δ
Treatment Group	$\mu + \gamma$	$\mu + \gamma + \delta + \alpha$	$\delta + \alpha$
DID			α

DID Estimation

Use Regression to Get DID Estimator

- Estimate the DID estimator in a regression framework has the following advantages:
  - It is easy to calculate standard errors
  - We can control for other variables which may reduce the selection bias further
  - It is easy to include multiple periods
  - We can study treatments with different **treatment intensity** (continous measure)
    - Example: varying increases in the minimum wage for different states
    - Estimate equation:
      $Y_{it} = \alpha + \delta(\text{TreatmentIntensity}_{it} \times Post_t) + \lambda_i + \gamma_t + \epsilon_{it}$

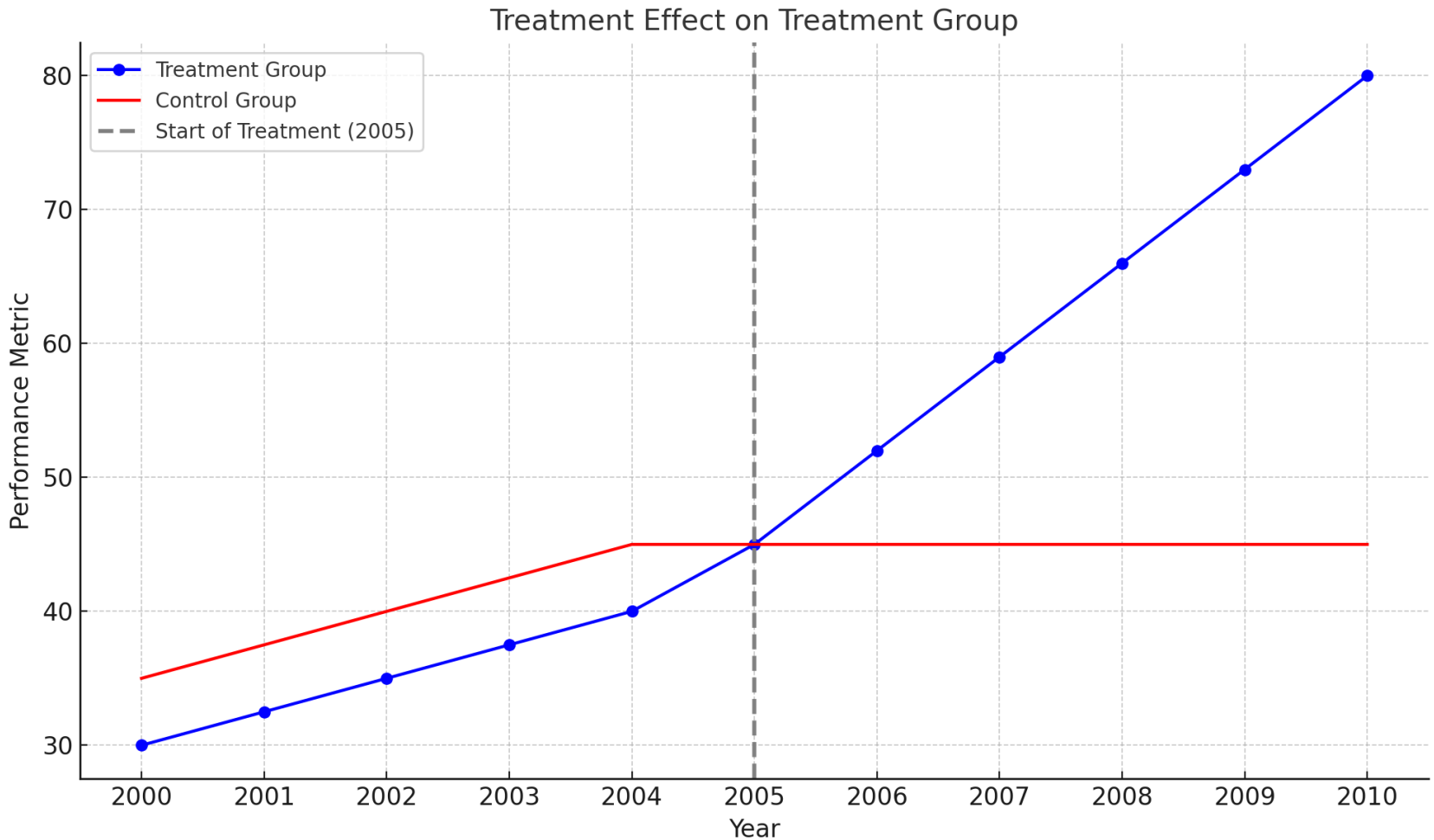
Examine Parallel Trends Assumption

Testing the Parallel Trends Assumption

- The key assumption for any DID design is the **Parallel Trends Assumption**
- The outcome in treatment and control groups would follow **the same time trend in the absence of the treatment**.
  - This does not mean that they have to have the same level (mean) of the outcome!
  - Parallel Trends Assumption is fundamentally untestable in the post-treatment period.
  - However, we can use **pre-treatment data** to provide supporting evidence for this assumption:
    - Graphical evidence showing similar pre-treatment trends
    - Formal tests using event study/dynamic DID specifications
  - Even if pre-trends are similar, we should still be concerned about **other policies or events changing at the same time**

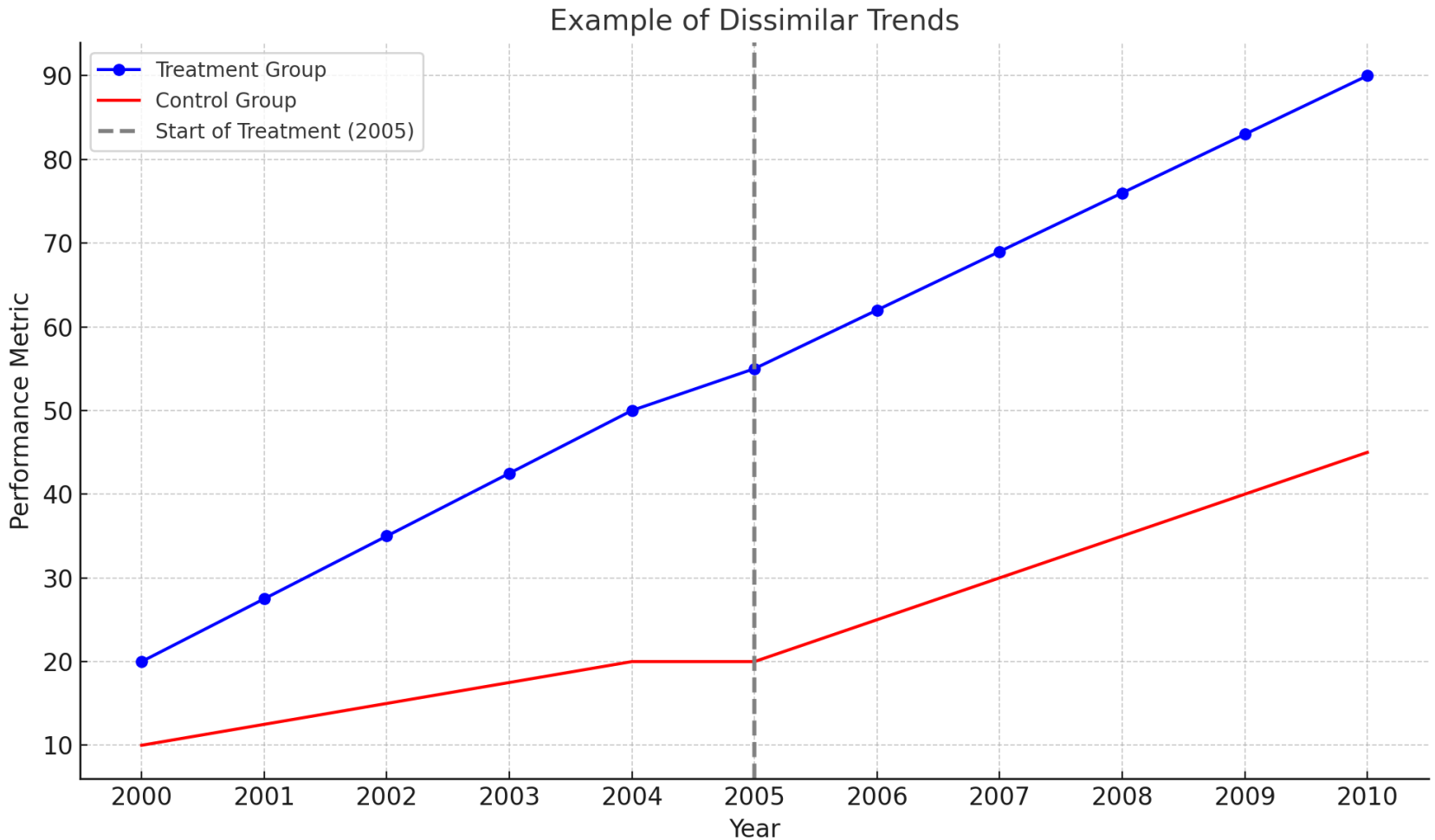
Examining Parallel Pre-Trends

Graphical Evidence



Examining Parallel Pre-Trends

Graphical Evidence



Dynamic DID Design

- We can include leads and lags in the DID specification:
  1. To examine whether treatment and control group share similar pre-treatment trends
  2. To analyze whether the treatment effect changes over time after implementation
- This approach is called the **Dynamic DID Design** or **Event-Study Design**

Dynamic DID Design

- The estimated regression would be:

$$Y_{it} = \alpha + \beta D_i + \sum_{\substack{k=-m \\ k\neq -1}}^{q} \delta_k \mathbf{I}[t - E_i = k]$$

$$+ \sum_{\substack{k=-m \\ k\neq -1}}^{q} \gamma_k D_i \times \mathbf{I}[t - E_i = k] + X_{it}'\theta + \varepsilon_{it},$$

  - E_i represents the timing when treatment happens.
  - $\mathbf{I}[t - E_i = k]$ is an indicator for being k years from the treatment event
  - t is the calendar year
    - Treatment occurs in $k = 0$ $(t = E_i)$
    - For example, $\mathbf{I}[t - E_i = -1]$ is a dummy variable indicating one year before treatment occurs
    - We usually use time $k = -1$ as baseline year

Dynamic DID Design

- The estimated regression would be:

$$Y_{it} = \alpha + \beta D_i + \sum_{\substack{k=-m \\ k\neq -1}}^{q} \delta_k \mathbf{I}[t - E_i = k]$$
$$+ \sum_{\substack{k=-m \\ k\neq -1}}^{q} \gamma_k D_i \times \mathbf{I}[t - E_i = k] + X_{it}'\theta + \varepsilon_{it},$$

  - $\gamma_{-2}, \gamma_{-3}, ..., \gamma_{-m}$ represent pre-trend
    - These coefficients should be zero if pre-tends is parallel
  - $\gamma_0, \gamma_1, ..., \gamma_q$ represent post-treatment effects

Dynamic DID Design

Example

Hsing-Wen Han, Kuang-Ta Lo, Yung-Yu Tsai, and Tzu-Ting Yang (2023), "**The Effect of Financial Resources on Fertility: Evidence from Administrative Data on Lottery Winners**", Working Paper

Empirical Example: Han et al. (2023)

Motivation

- During the past fifty years, fertility rates in developed countries have declined dramatically
- Low fertility rate leads to the growth of an aging population, workforce shortages, and reductions in tax revenue.
- Many countries initiated child-related cash transfer policies to encourage childbearing.
  - On average, the public spending of child-related cash benefits accounts for 1.1% of GDP in OECD countries.
- The rationale behind these policies is that people do not have enough income to afford the expense of raising children, so the government needs to subsidize them.

Empirical Example: Han et al. (2023)

Motivation

- Empirically, there is an endogenous problem between income and fertility.
  - Reverse Causality
  - Income effect confounds with substitution effect
    - Both working and raising children are time-consuming activities
    - A sudden increase in wage income can increase the relative price of having children
    - Higher wage income would make people work more and reduce demand for children

Empirical Example: Han et al. (2023)

Dynamic DID Design

- This paper examines the fertility impact of the large and permanent income shock generated by winning lottery prizes.
- We implement a dynamic DID design to examine the causal effect of large income shock on fertility.
- Compare the trend in fertility before and after receiving a windfall gain between:
  - Households winning 1,000,000 NT$ from lottery prizes.
  - Households winning less than 10,000 NT$.

Empirical Example: Han et al. (2023)

Dynamic DID Design

- We estimate the following regression:

$$Y_{it} = \alpha + \beta D_i + \sum_{k=-3}^{6} \delta_k \mathbf{I}[t - E_i = k] + \sum_{k=-3}^{6} \gamma_k D_i \times \mathbf{I}[t - E_i = k] + X_{it}'\theta + \varepsilon_{it},$$

  - D_i represents treatment group dummy.
  - Treatment Group:
    - Households who earn more than 1,000,000 NT$ by winning lotteries in a given year
  - Control group:
    - Households who earn less than 10,000 NT$ from winning lotteries during sample period

Empirical Example: Han et al. (2023)

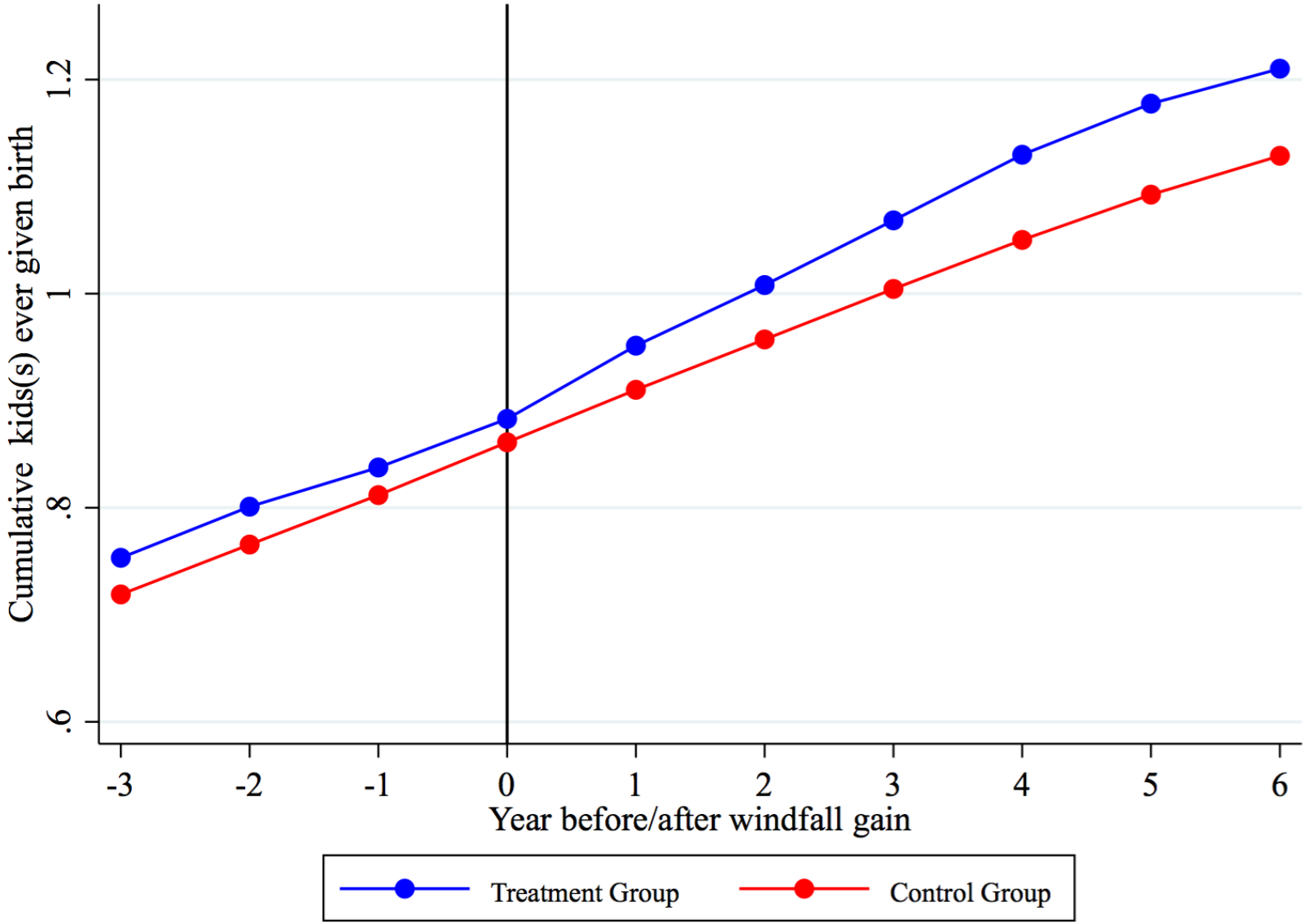
Dynamic DID Design

$$Y_{it} = \alpha + \beta D_i + \sum_{k=-3}^{6} \delta_k \mathbf{I}[t - E_i = k]$$
$$+ \sum_{k=-3}^{6} \gamma_k D_i \times \mathbf{I}[t - E_i = k] + X_{it}'\theta + \varepsilon_{it},$$

- Y_{it}: Cumulative number of children for household i in the year t
- $\mathbf{I}[t - E_i = k]$ denotes dummy variables for the year before and after winning lottery.
  - E_i is the lottery-winning year
  - For example, $\mathbf{I}[t - E_i = 1]$ represents a dummy for the first year after winning lottery.
- Note that we use one year before lottery-winning year as the baseline year (i.e. $k = -1$).

Examining Parallel Pre-Trends

Raw Data: Cumulative Number of Children



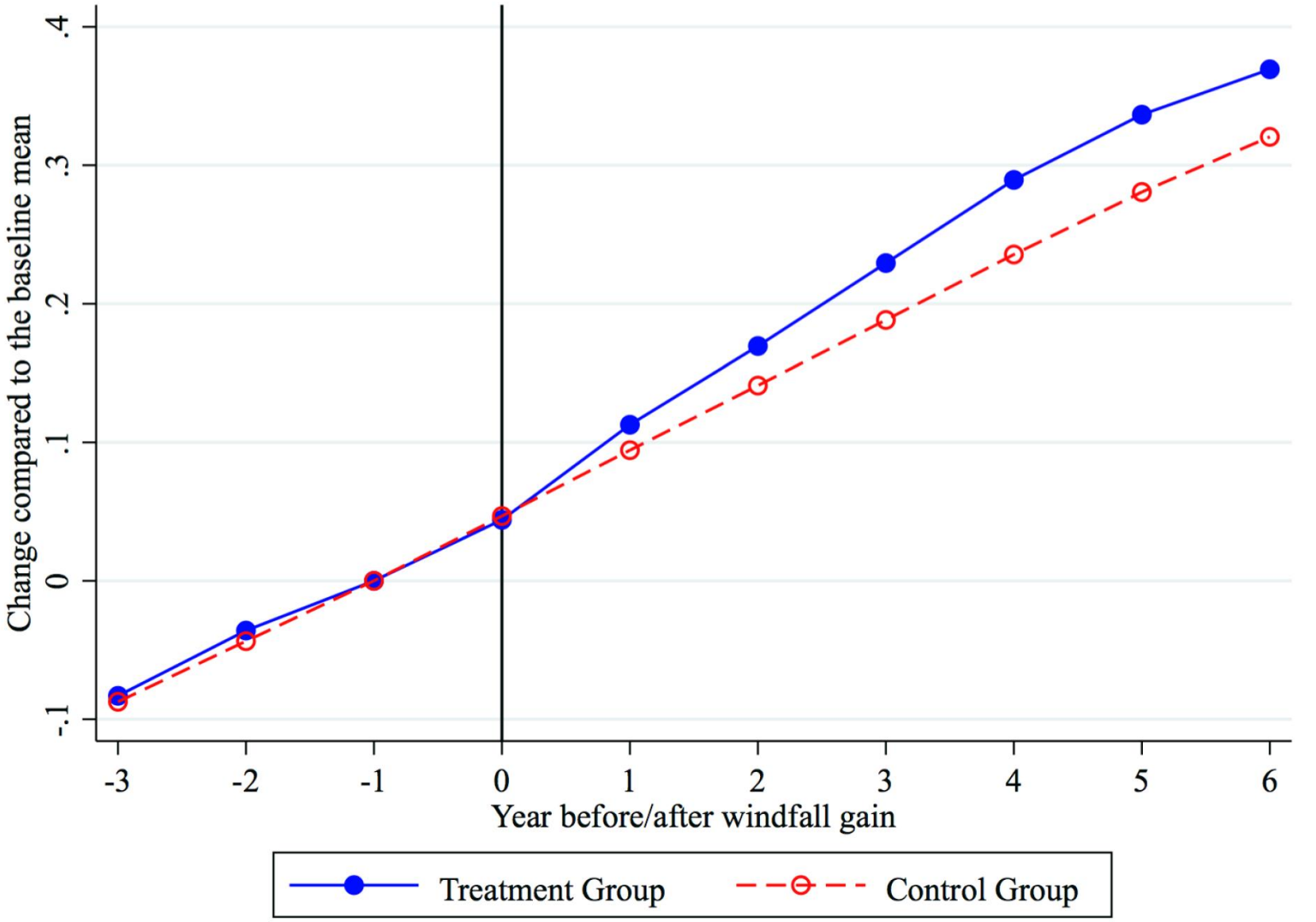
Examining Parallel Pre-Trends

Raw Data: Cumulative Number of Children

- Since we focus on trend rather than level, we sometimes subtract the baseline mean ($k = -1$) from the outcome at each time period

Examining Parallel Pre-Trends

Subtract the Baseline Mean: Cumulative Number of Children



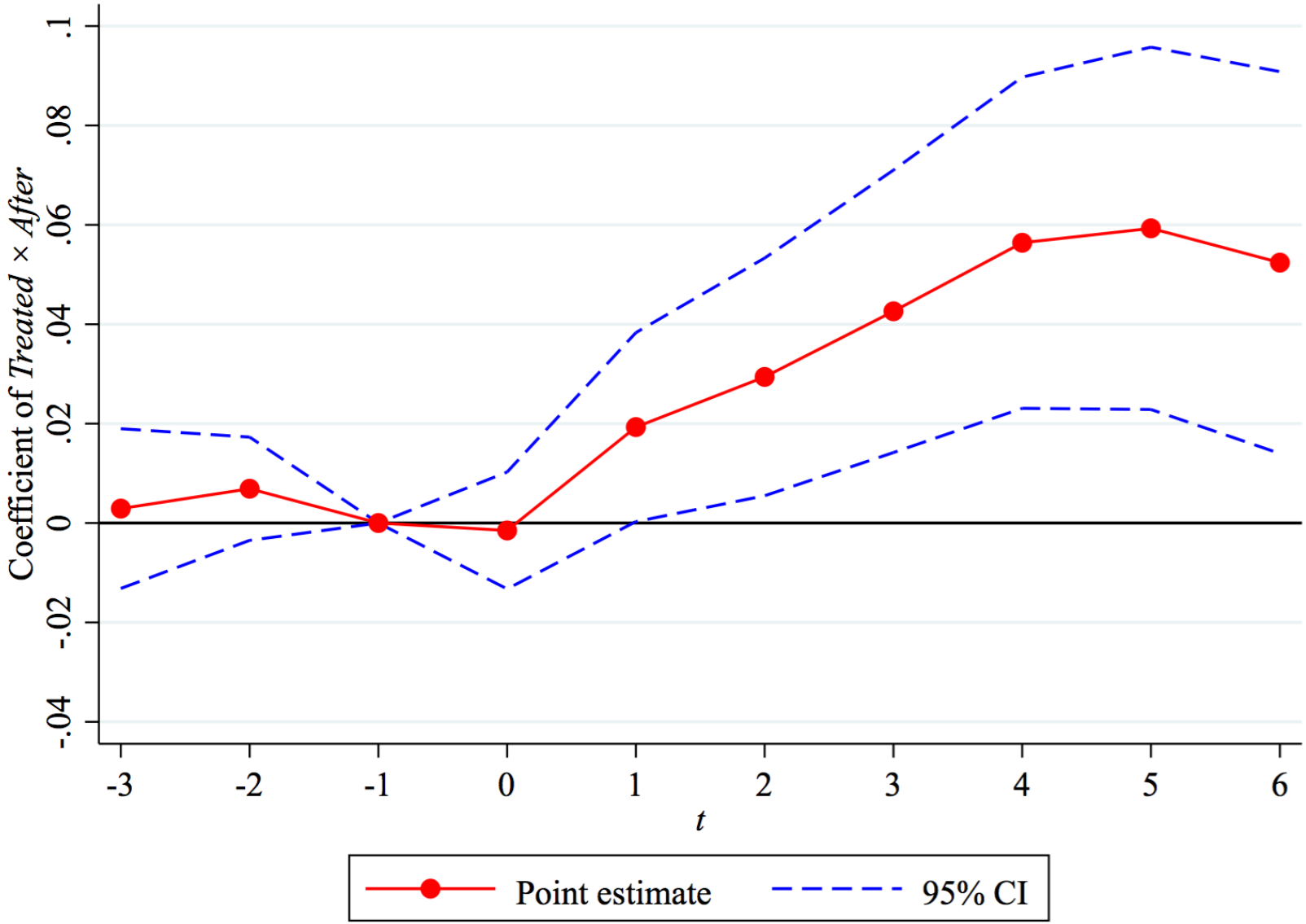
Examining Parallel Pre-Trends

Raw Data: Cumulative Number of Children

- We can formally examine whether two groups share similar pre-treatment trend by showing the estimated coefficients $\gamma_{-2}, \gamma_{-3}, ..., \gamma_6$
- If pre-treatment trend is parallel between two groups, γ_{-2}, γ_{-3} should be close to zero
- $\gamma_0, \gamma_1, ..., \gamma_6$ represent the treatment effects of winning lotteries

Examining Parallel Pre-Trends

Dynamic DID design: Cumulative Number of Children



Another Way to Examine Parallel Pre-Trends

- Conduct a DID estimation using pre-treatment data
- Arbitrarily choose a "treatment timing" in the pre-treatment period

$$Y_{it} = \mu + \gamma D_i + \delta Placebo_t + \alpha(D_i \times Placebo_t) + X_{it}'\beta + \varepsilon_{it},$$

- *Placebo* is a dummy indicating fake "post-treatment" period
- If pre-trend is parallel between two groups, we would expect $\alpha = 0$

STATA Example

Empirical Example 1: Eissa and Jeffrey (1996)

Eissa, Nada, and Jeffrey B. Liebman. (1996) "**Labor Supply Responses to the Earned Income Tax Credit**" QJE

- They want to look at the effect of tax credit on labor supply

Empirical Example 1: Eissa and Jeffrey (1996)

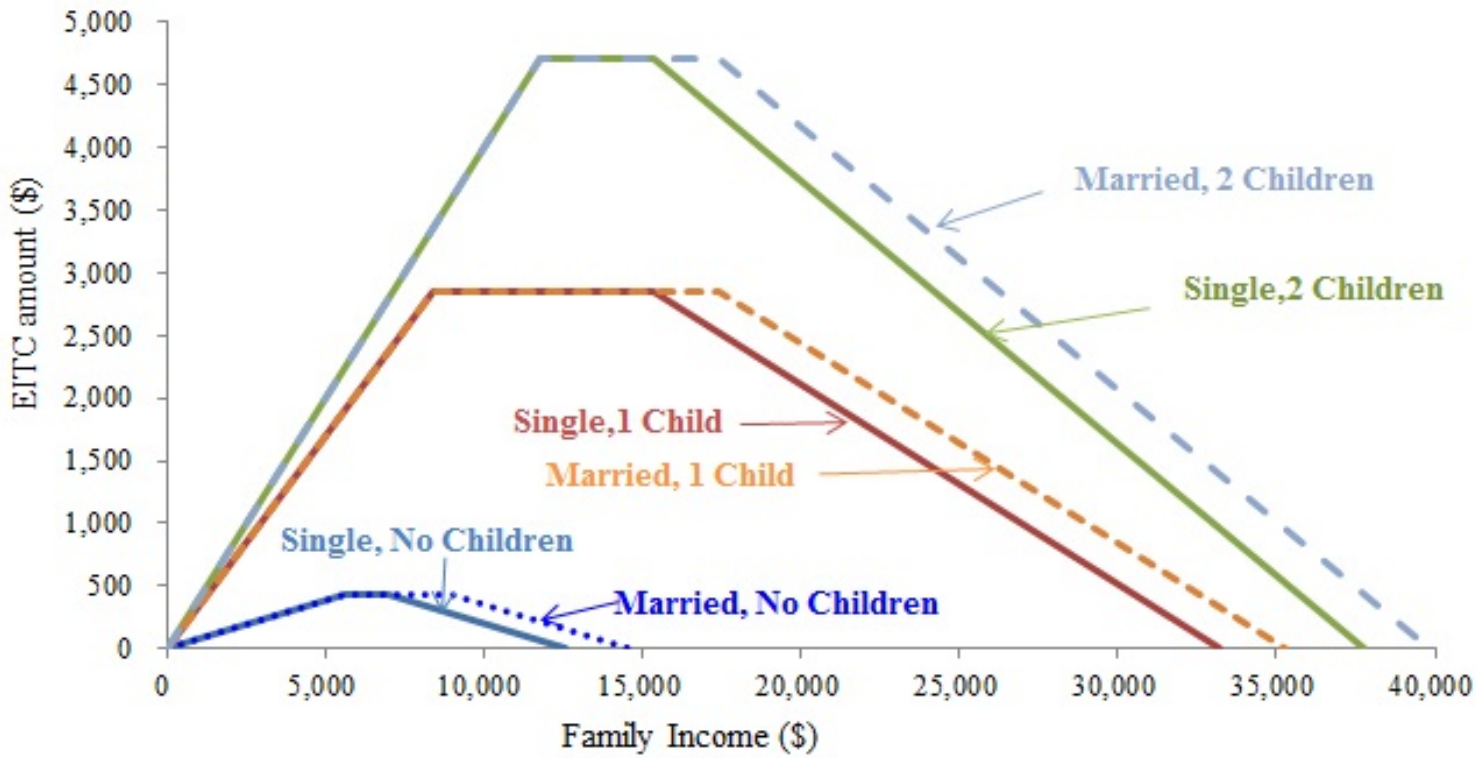
STATA Implementation

- See DID.do
- Use eitc.dta

Empirical Example 1: Eissa and Jeffrey (1996)

- Earned Income Tax Credit (EITC) is a refundable tax credit that subsidizes earnings of working poor in US
  - The amount of cash transfer depends on the number of children and previous year earnings
  - In 1994, the amount of EITC had large increase for those who have children
- The author examined how did labor supply respond to this change in tax credit using DID design

EITC benefit rule



Step 1: Define treatment and control groups

- Treatment group: those who have at least one children
  - They receive much more tax credit after 1994
- Control group: those who do not have children
  - Their tax credit did not increase after 1994

Step 1: Define treatment and control groups

- D: a dummy that indicate whether individual i had children or not

$$D = \begin{cases} 1 & \text{if individual } i \text{ had at least one children} \\ 0 & \text{if individual } i \text{ did not have children} \end{cases}$$

Step 1: Define treatment and control groups

- *Post*: a dummy that indicate whether individual i was observed after 1994 (Post-treatment period)

$$Post = \begin{cases} 1 & \text{if individual } i \text{ was observed after 1994} \\ 0 & \text{if individual } i \text{ was observed before 1994} \end{cases}$$

- $D \times Post$: a treatment dummy that indicate whether individual i was affected by 1994 EITC expansion

Step 1: Define treatment and control groups

STATA Command

```
** a dummy for treatment group
gen treated = (children >= 1)

** a dummy for post-treatment period
gen post = (year >= 1994)

** treatment variable (DID key variable)
gen treated_post = treated*post
```

- Create dummy variables for treatment group, post-treatment period and treatment variable (DID)

Step 2: Graphical Analysis

- Plot the time trend of outcomes for treatment and control groups
  - Check whether there is a **parallel trend** in outcomes of treatment and control groups **before reform**
  - Examine whether the outcomes of treatment group exhibits different trend **after reform**

Step 2: Graphical Analysis

STATA Command

```
1 collapse (mean) work, by(year treated)
```

- **collapse:** This command converts the data into a dataset of summary statistics, such as sums, means, medians, and so on
  - Converts the data into mean of "work" (Labor Force Participation Rates) by group and year - group and year mean

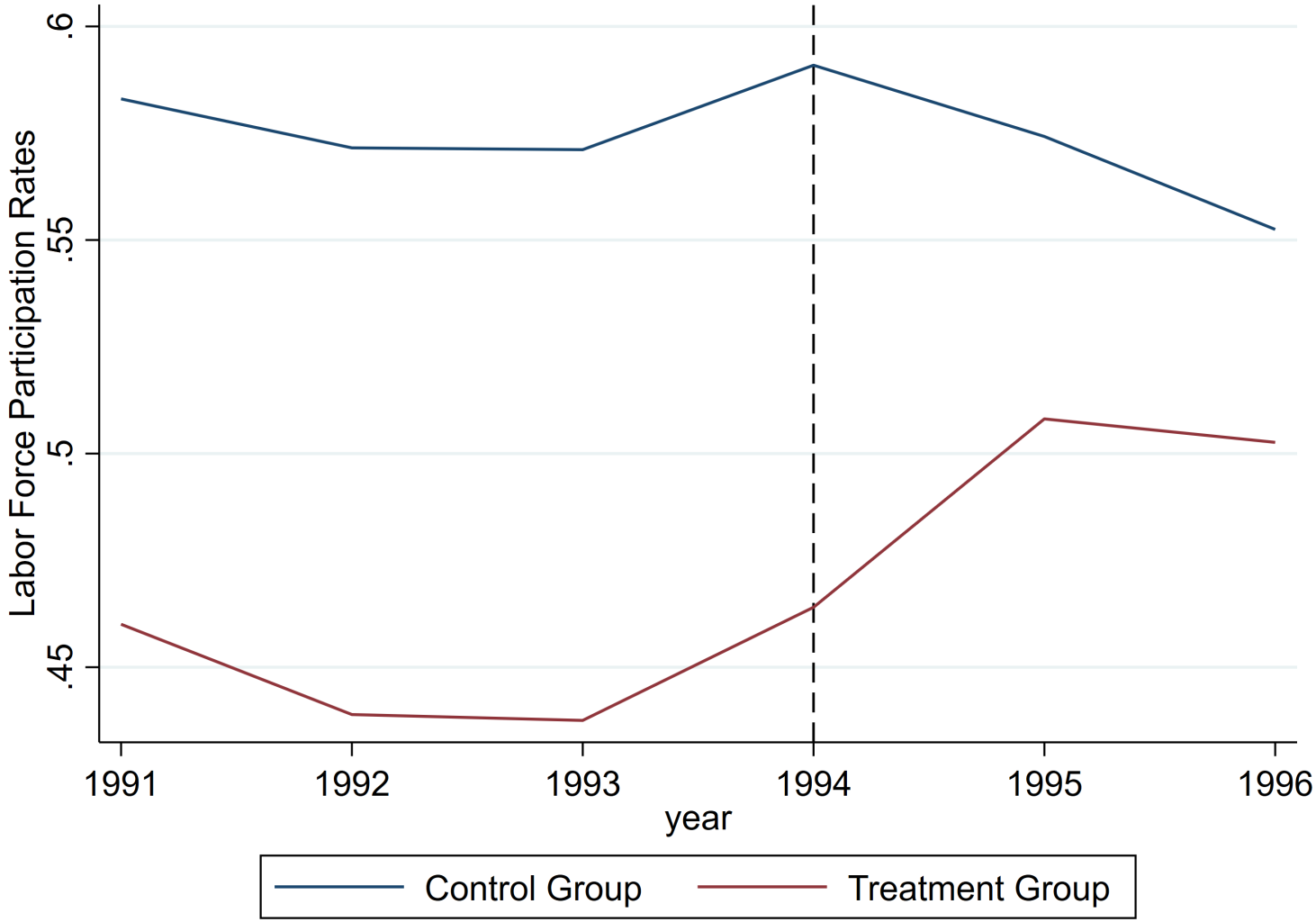
Step 2: Graphical Analysis

STATA Command

```
1 graph twoway (line work year if treated==0) (line
      work year if treated==1), legend(order(1 "Control
       Group" 2 "Treatment Group")) ///
2 graphregion(fcolor(white) lcolor(white) ifcolor(white
      ) ilcolor(white)) ///
3 xline(1994, lp(dash) lc(black) lw(medthin)) ///
4 ytitle(Labor Force Participation Rates)
```

- Create a twoway gragh ("work" by "year") for treatment and control groups

Trend in Outcomes for Treatment and Control Groups



Step 3: Pre/Post DID regression

STATA Command

- We can estimate the following DID regression:

$$Y_{it} = \mu + \gamma D_i + \delta Post_t + \alpha(D_i \times Post_t) + X'_{it}\beta + \varepsilon_{it},$$

Step 3: Pre/Post DID regression

STATA Command: outreg2

```
reg work post treated treated_post,r
outreg2 using "$table\DID_pre_post.xls", replace
    nocon keep(treated_post)  ///
stats(coef se) addstat(Sample Size, e(N)) ///
addtext(Basic, Yes, Age, No, Demo, No, State FE, No,
    Year FE, No)
```

- **outreg2:** Stata command (from a user-written package) to export regression tables to an external file
  - **using "$table\DID_pre_post.xls":** Path and filename for the output Excel file (tex/csv)
  - **replace:** Overwrites any existing file with the same name
  - **nocon:** Suppresses the constant term in the output table
  - **keep(treated_post):** Only output the coefficient for the treated_post variable

Step 3: Pre/Post DID regression

STATA Command: outreg2

```
1 reg work post treated treated_post age age2,r
2 outreg2 using "$table\DID_pre_post.xls", append nocon
      keep(treated_post)  ///
3 stats(coef se) addstat(Sample Size, e(N)) ///
4 addtext(Basic, Yes, Age, Yes, Demo, No, State FE, No,
      Year FE, No)
```

- **outreg2 options:**
  - **append:** Append the output to an existing file
  - **stats(coef se):** Specifies to report coefficients and standard errors
  - **addstat(Sample Size, e(N)):** Adds the sample size to the table using e(N) which returns the number of observations
  - **addtext():** Adds text to the table describing which controls were included in the model
  - **///:** Line continuation marker in Stata, allowing the command to continue on the next line

Pre/Post DID Results

```
. reg work post treated treated_post nonwhite age age2 ed finc nonlaborinc,r

Linear regression                               Number of obs     =     13,746
                                                F(9, 13736)       =     122.54
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1993
                                                Root MSE          =     .44741
```

work	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
post	-.008003	.011667	-0.69	0.493	-.030872	.014866
treated	-.072796	.0118336	-6.15	0.000	-.0959914	-.0496006
treated_post	.0429062	.0156294	2.75	0.006	.0122704	.0735421
nonwhite	-.0636661	.0081387	-7.82	0.000	-.0796191	-.0477131
age	.0333505	.003121	10.69	0.000	.0272328	.0394681
age2	-.0004231	.0000427	-9.92	0.000	-.0005067	-.0003395
ed	.0144307	.001531	9.43	0.000	.0114297	.0174317
finc	9.02e-06	7.44e-07	12.13	0.000	7.56e-06	.0000105
nonlaborinc	-.000027	1.18e-06	-22.83	0.000	-.0000293	-.0000247
_cons	-.1554345	.0584772	-2.66	0.008	-.2700577	-.0408112

Step 3: Pre/Post DID regression

STATA Command

- Show the treatment effect by treatment intensity:

```
** treatment intensity
gen treated_1 = (children==1)
gen treated_2 = (children>=2)
gen treated_post_1 = post*treated_1
gen treated_post_2 = post*treated_2

reg work post treated_1 treated_2 treated_post_1
    treated_post_2 nonwhite age age2 ed finc
    nonlaborinc,r
```

Pre/Post DID Results

Treatment Intensity

```
. reg work post treated_1 treated_2 treated_post_1 treated_post_2 nonwhite age age2 ed finc
> c,r
```

Linear regression

Number of obs	=	13,746	
F(11, 13734)	=	111.21	
Prob > F	=	0.0000	
R-squared	=	0.2018	
Root MSE	=	.44676	

work	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
post	-.0080963	.0116636	-0.69	0.488	-.0309584	.0147659
treated_1	-.0257933	.0140935	-1.83	0.067	-.0534185	.0018319
treated_2	-.1084904	.0134775	-8.05	0.000	-.1349081	-.0820726
treated_post_1	.0194807	.0202555	0.96	0.336	-.0202229	.0591842
treated_post_2	.0576754	.0175247	3.29	0.001	.0233244	.0920263
nonwhite	-.0589521	.0081535	-7.23	0.000	-.0749341	-.0429702
age	.0353075	.0031283	11.29	0.000	.0291756	.0414394
age2	-.0004527	.0000428	-10.57	0.000	-.0005366	-.0003688
ed	.0145814	.0015276	9.55	0.000	.0115872	.0175756
finc	8.95e-06	7.41e-07	12.08	0.000	7.50e-06	.0000104
nonlaborinc	-.0000266	1.18e-06	-22.59	0.000	-.000029	-.0000243
_cons	-.1873079	.0584447	-3.20	0.001	-.3018675	-.0727482

Step 4: Examine Common Trend by a Placebo Test

STATA Command

- Creating a placebo DID model is when you arbitrarily choose a treatment time before your actual treatment time
- Test to see if you get a "significant" treatment effect (Hope not)

```
gen placebo = (year >= 1992)
gen treated_placebo = treated*placebo

reg work treated placebo treated_placebo if year
    <1994,r
```

Placebo Test

```
. reg work treated placebo treated_placebo if year<1994,r

Linear regression                               Number of obs     =      7,401
                                                F(3, 7397)        =      42.00
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0167
                                                Root MSE          =     .49594
```

work	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treated	-.1229792	.0196214	-6.27	0.000	-.1614428	-.0845157
placebo	-.0116737	.0184199	-0.63	0.526	-.047782	.0244346
treated_placebo	-.0101282	.0243824	-0.42	0.678	-.0579245	.0376682
_cons	.5830325	.0148165	39.35	0.000	.553988	.612077

Step 5: Dynamic DID Design

STATA Command

```
1 reg work treated pre_event_3 pre_event_2 post_event_0
    -post_event_2 pre_dd_3 pre_dd_2 post_dd_0 -
    post_dd_2 nonwhite age age2 ed finc nonlaborinc,r
2
3 outreg2 using "$table\DID_dynamic.xls", replace nocon
     keep(pre_dd_* post_dd_*)
4  year<1994,r
```

Dynamic DID Estimates

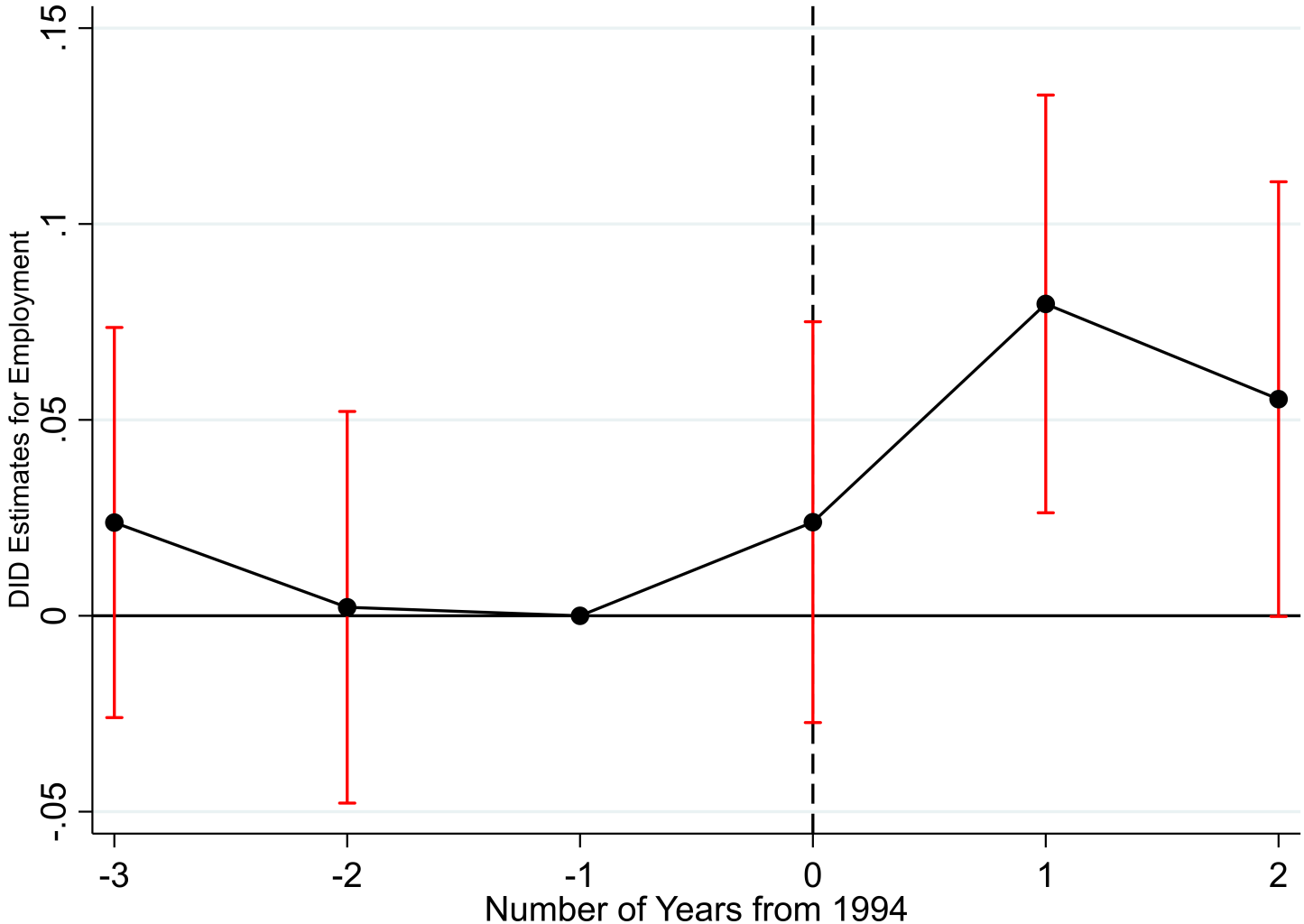
Linear regression

Number of obs	= 13,746
F(17, 13728)	= 66.34
Prob > F	= 0.0000
R-squared	= 0.1998
Root MSE	= .44742

work	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treated	-.0819711	.019021	-4.31	0.000	-.1192548	-.0446874
pre_event_3	-.004421	.0192407	-0.23	0.818	-.0421354	.0332934
pre_event_2	-.0019519	.0190814	-0.10	0.919	-.0393541	.0354503
post_event_0	.0062106	.0194671	0.32	0.750	-.0319476	.0443688
post_event_1	-.0184141	.020402	-0.90	0.367	-.0584048	.0215765
post_event_2	-.0202813	.0211655	-0.96	0.338	-.0617685	.021206
pre_dd_3	.02384	.0253685	0.94	0.347	-.0258858	.0735658
pre_dd_2	.0021661	.0254708	0.09	0.932	-.0477601	.0520923
post_dd_0	.0239192	.0260711	0.92	0.359	-.0271838	.0750221
post_dd_1	.0796404	.0271846	2.93	0.003	.0263549	.132926
post_dd_2	.0552523	.0282647	1.95	0.051	-.0001504	.1106551
nonwhite	-.0637037	.0081406	-7.83	0.000	-.0796604	-.0477469
age	.0334936	.0031221	10.73	0.000	.0273738	.0396134
age2	-.0004249	.0000427	-9.96	0.000	-.0005086	-.0003413
ed	.0144686	.0015302	9.46	0.000	.0114692	.0174681
finc	9.02e-06	7.43e-07	12.14	0.000	7.56e-06	.0000105
nonlaborinc	-.000027	1.18e-06	-22.77	0.000	-.0000293	-.0000246
_cons	-.1561703	.0592764	-2.63	0.008	-.2723601	-.0399805

Dynamic DID Estimates

Graph



R Example

Step 1: Define treatment and control groups

R Command

```
eitc_data <- eitc_data %>%
mutate(treated = as.numeric(children >= 1),
post = as.numeric(year >= 1994),
treated_post = treated * post,
age2 = age^2,
nonlaborinc = finc - earn)
```

- **%>%:** The pipe operator passes the data from the left side to the function on the right side as the first argument,
  - Allowing for more readable code by chaining operations sequentially
- **as.numeric():** Converts logical values (TRUE/FALSE) to numeric values (1/0)

Step 2: Graphical Analysis

R Command

```
data_summary <- eitc_data %>%
group_by(year, treated) %>%
summarise(work_mean = mean(work, na.rm = TRUE))
```

- **data_summary:** Create a summary of the dataset for graphical analysis.
  - **group_by(year, treated):** Groups the data by year and treatment status.
  - **summarise(work_mean = mean(work, na.rm = TRUE)):** Calculates the mean work participation for each group, ignoring missing values.

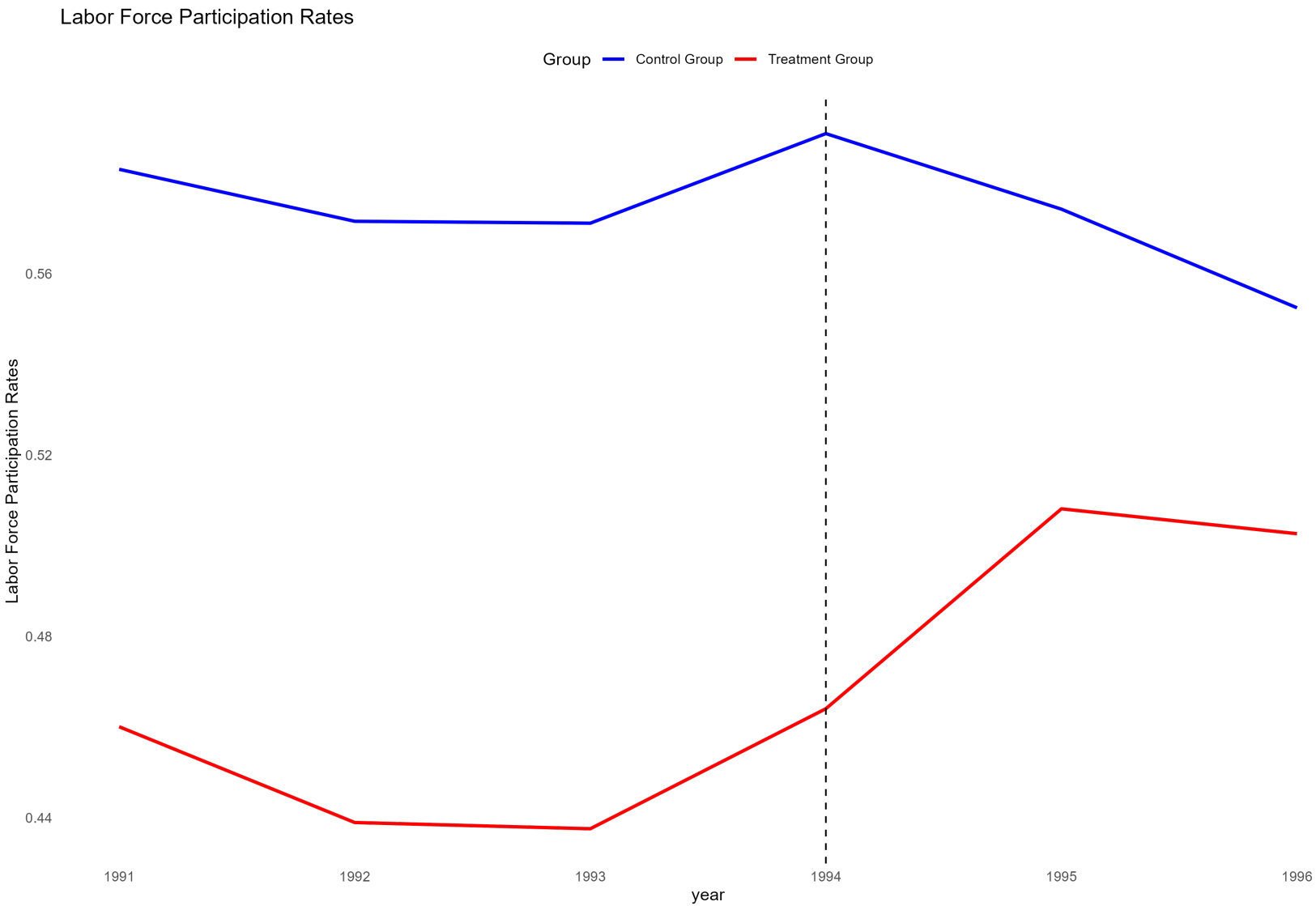
Step 2: Graphical Analysis

R Command

```
ggplot(data_summary, aes(x = year, y = work_mean,
    color = factor(treated))) +
geom_line(size = 1) +
scale_color_manual(values = c("blue", "red"),
labels = c("Control Group", "Treatment Group")) +
geom_vline(xintercept = 1994, linetype = "dashed",
    color = "black") +
labs(title = "Labor Force Participation Rates",
y = "Labor Force Participation Rates",
color = "Group") +
theme_minimal() +
theme(legend.position = "top",
panel.grid.major = element_blank(),
panel.grid.minor = element_blank())

ggsave(paste0(pic, "/eitc_DID.png"), width = 12,
    height = 8, units = "in")
```

DID graph


Labor Force Participation Rates
Group
Control Group
Treatment Group
0.56
0.52
0.48
0.44
Labor Force Participation Rates
1991
1992
1993
1994
1995
1996
year

Step 3: Pre/Post DID regression

R Command

- Show the treatment effect by treatment intensity:

```
1 DID <- lm(work ~ treated_post+post+treated, data
      = eitc_data)
2
3 summary(DID)
```

Step 4: Placebo Test

R Command

```
eitc_data <- eitc_data %>%
mutate(placebo = as.numeric(year >= 1992),
treated_placebo = treated * placebo)

# Regression for placebo test
model_placebo <- lm(work ~ treated + placebo +
    treated_placebo, data = eitc_data, subset = (year
     < 1994))
summary(model_placebo)
```

Step 5: Dynamic DID Design

R Command

```
eitc_data <- eitc_data %>%
mutate(treat_year = 1994,
event_year = year - treat_year,

# Create pre-treatment DID dummies
pre_dd_1 = as.numeric(event_year == -1) * treated,
pre_dd_2 = as.numeric(event_year == -2) * treated,
pre_dd_3 = as.numeric(event_year == -3) * treated,

# Create post-treatment DID dummies
post_dd_0 = as.numeric(event_year == 0) * treated,
post_dd_1 = as.numeric(event_year == 1) * treated,
post_dd_2 = as.numeric(event_year == 2) * treated)
```

Step 5: Dynamic DID Design

R Command

```
1 model_dynamic_did <- lm(work ~ treated + pre_event_3
      + pre_event_2 + post_event_0 + post_event_1 +
      post_event_2 + pre_dd_3 + pre_dd_2 + post_dd_0 +
      post_dd_1 + post_dd_2 + nonwhite + age + age2 +
      ed + finc + nonlaborinc, data = eitc_data)
2 summary(model_dynamic_did)
```

Step 5: Dynamic DID Design

R Command

```
# Export regression results for Dynamic DID
dynamic_did_results <- tidy(model_dynamic_did) %>%
filter(grepl("pre_dd_|post_dd_", term))

write.xlsx(dynamic_did_results, file = paste0(table,
    "/DID_dynamic.xlsx"), overwrite = TRUE)
```

- **tidy():** From the broom package, used to organize regression output into a data frame.
  - **filter(grepl("pre_dd_|post_dd_", term)):** Extracts coefficients related to the dynamic DID variables.
    - This helps focus the analysis specifically on the treatment effects across different time periods.

Step 5: Dynamic DID Design

R Command

```
# Export regression results for Dynamic DID
dynamic_did_results <- tidy(model_dynamic_did) %>%
filter(grepl("pre_dd_|post_dd_", term))

write.xlsx(dynamic_did_results, file = paste0(table,
    "/DID_dynamic.xlsx"), overwrite = TRUE)
```

- **write.xlsx():** Writes the filtered regression results to an Excel file.
  - **file = paste0(table, "/DID_dynamic.xlsx"):** Specifies the file path and name for exporting the results.
  - **overwrite = TRUE:** Overwrites the existing file if it already exists, ensuring the latest results are saved.

Statistical Inference

Statistical Inference in DID Estimation

- Many papers using a DID design use data from many years
  - Not only 1 pre and 1 post period
- The variables of interest in DID setups typically vary at the group level, and outcome variables are often serially correlated
  - The observation is at individual level
  - But the group (policy variation) is defined at city level

Statistical Inference in DID Estimation

- The conventional (heteroskedasticity-robust) standard errors depend on assumption of independence
  - IID assumption: independent and identically distributed
  - The IID assumption implies that observations are random draws from some population and are uncorrelated with each other
- In DID design or panel data, independence assumption is often unrealistic

Statistical Inference in DID Estimation

Example

$$Y_{ist} = \mu + \gamma Treat_s + \delta Post_t + \alpha^{DD} d_{st} + X'_{ist}\beta + \epsilon_{ist}.$$

- $d_{st} = Treat_s \times Post_t = 1$ if a state s raises minimum wage at time t.
  - $d_{11} = 0, d_{12} = 0, d_{13} = 0, d_{14} = 0, \ldots$
  - $d_{21} = 0, d_{22} = 0, d_{23} = 1, d_{24} = 1, \ldots$
  - $d_{31} = 0, d_{32} = 1, d_{33} = 1, d_{34} = 1, \ldots$

  $\Rightarrow$ This implies strong within-state serial correlation in the treatment variable over time, which can lead to biased inference if not properly accounted for.

Statistical Inference in DID Estimation

- As Bertrand, Duflo, Mullainathan (2004) point out:
  - Conventional standard errors often severely **understate** the standard error of the DID estimators
- Intuition:
  - Conventional standard errors do not account for the fact that observations within the same group may be correlated
  - This leads to an underestimation of the true standard errors, potentially causing over-rejection of the null hypothesis.

Statistical Inference in DID Estimation

Heteroskedasticity-Robust Standard Errors

- The variance-covariance matrix of the OLS estimator $\hat{\beta}$, adjusting for heteroskedasticity but not autocorrelation:

$$\hat{\text{SE}}(\hat{\beta})_{\text{Robust}} = (X'X)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}$$

  - X is the matrix of covariates and treatment variables
  - $\hat{\epsilon}_i$ are the residuals from the OLS regression
  - n is the number of observations
  - Assumes observations are uncorrelated

- This is a heteroskedasticity-robust estimator that assumes no autocorrelation and relies on the independence of residuals across units.

Statistical Inference in DID Estimation

Clustered Standard Errors

- Clustered standard errors take into account not just heteroskedasticity but also autocorrelation within clusters
- For data clustered into G groups, the clustered variance-covariance matrix of $\hat{\beta}$ is:

$$\hat{\text{SE}}(\hat{\beta})_{\text{cluster}} = (X'X)^{-1} \left(\sum_{g=1}^{G} \sum_{i \in g} \sum_{j \in g} \hat{\epsilon}_i \hat{\epsilon}_j X_i X_j' \right) (X'X)^{-1}$$

  - g indexes the clusters (e.g., states, schools, firms)
  - i and j index observations within clusters
  - This allows for intra-cluster correlation, leading to more accurate (typically larger) standard error estimates
  - When clustering, we allow for arbitrary correlation of residuals within a cluster but assume independence across clusters

Statistical Inference in DID Estimation

Solutions

- Simple solution:
  - Clustering standard errors at the group level
  - In STATA one would simply add **vce(cluster state)** to the regression equation if one analyzes state level variation
- Asymptotic consistency of estimated clustered standard errors depends on number of clusters
  - Only guaranteed to get precise estimate of correct standard errors if we have a lot of clusters
  - If too few clusters, standard errors will be too low
  - Hansen (2007) suggests that having at least 10 clusters may be sufficient for consistent inference under certain conditions, though more is generally better.
- Compare the robust SE to the cluster SE and take maximum of the two

Statistical Inference in DID Estimation

Summary

- DID estimators often rely on variation at the group level and involve serially correlated outcomes.
  - Conventional (robust) standard errors assume independence across observations, which is unrealistic in DID setups.
  - Ignoring within-group correlation leads to understated standard errors and false precision.
- Clustered standard errors account for intra-group correlation and provide more accurate inference.
- Reliable inference requires a sufficient number of clusters; 10 or more is a commonly used rule of thumb.

Suggested Readings

- Chapter 5, Mastering Metrics: The Path from Cause to Effect
- Chapter 5, Mostly Harmless Econometrics
- Chapter 9, Causal Inference: The Mixtape