Differences in Differences

Compass-Lexecon

Raquel Carrasco

Universidad Carlos III de Madrid

May-June 2024

Outline Session 1

- Introduction to Differences in Differences (DD) Methodology
  - Canonical DD
  - Identification Assumptions
  - TWFE model
  - Event studies
- Advancing to Triple Differences Model (DDD)
  - Conceptual Framework
  - Applications
- Further issues
  - The Scale Dependence of the DD Identifying Assumptions
  - Standard Errors in DD Strategies

DD: Introduction

- Purpose: Estimation of **causal** effects.
- Let's consider the case that data do not come from a randomized experiment.
- Since we are dealing with observational data, we have several options:
  - A treatment-control comparison is not necessarily a causal comparison because of the potential systematic differences between two groups.
    - We need to have data on everything that affects treatment timing and the outcome of interest (unconfoundedness assumption).
  - A before-after comparison is not necessarily a causal comparison because of the potential change in time.
    - Does not account for potential trends in outcomes.
    - This is more reasonable if we study very short-run effects, but that is not usually the case.

DD: Introduction

- DD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.
- DD combines previous approaches to avoid their pitfalls.
  - Advantage: Allow for selection on unobservables and for time-trends.
  - Assumption: Absent the treatment, the outcome of interest would grow similarly across groups/cohorts - Parallel Trends assumption.
    - Is Parallel Trends a plausible assumption in our application?

Data requirements

- Basic setup: two or more groups, with units observed in two or more periods.
- In some periods and some groups are exposed to the treatment (repeated cross-sections, or panel data)
- The individual fixed-effects strategy requires longitudinal data.
- Recall how the fixed effects model assumes:

$$E(Y_{i0t} \mid i, t) = \alpha_i + \lambda_t.$$

- The DD model makes a very similar assumption but conditioning on a group level instead of an individual level effect:

$$E(Y_{i0st} \mid i, t) = \gamma_s + \lambda_t,$$

where s could be, for example, a state.

- While the basic strategy is the same, the data requirements are much less. We do not need repeated observations on unit i (i.e. a panel).
  - The source of omitted variable bias consists of unobserved variables at the state/cohort level, s.

Example 1:

Effect of compulsory laws on years of education

- Effect of compulsory schooling laws on schooling obtained in the US.
- These laws are set at the state level. For example:
  - Florida raised its compulsory schooling requirement from 5 to 7 grades in 1935.
  - Georgia, neighbor state, required 6 grades both before and after 1935.
- We can think of FL as the treatment state and GA as the control state.
- 1934 is a control period and 1935 is the treatment period.

Example 1:

Effect of compulsory laws on years of education

- Let D_{st} denote a dummy for the treatment and period, i.e. a compulsory schooling requirement in FL:

$$D_{st} = \begin{cases} 1 \text{ if } s = FL \text{ and } t = 1935 \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

  Thus, $D_{st} = 1$ if s was treated in t.

- Consider this model:

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + e_{ist}$$

Example 1:

Effect of compulsory laws on years of education

- It is easy to see that

$$\begin{aligned} E\left(Y_{ist}|\, s=GA, t=1935\right)-E\left(Y_{ist}|\, s=GA, t=1934\right) &= \lambda_{1935}-\lambda_{1934}, \\ E\left(Y_{ist}|\, s=FL, t=1935\right)-E\left(Y_{ist}|\, s=FL, t=1934\right) &= (\lambda_{1935}-\lambda_{1934}) \\ & \quad +\beta. \end{aligned}$$

- The population difference-in-difference mean is

$$\begin{aligned} &[E(Y_{ist}|\, s=FL, t=1935)-E(Y_{ist}|\, s=FL, t=1934)] \\ -&[E(Y_{ist}|\, s=GA, t=1935)-E(Y_{ist}|\, s=GA, t=1934)]=\beta. \end{aligned}$$

Example 2:

Effect of minimum wage on employment

- Effect of minimum wage on employment (Card and Krueger, 1994).
- In March 1992, the state of New Jersey (NJ) increased the legal minimum wage by 19% (from 4.25$ to 5.05$ per hour), whereas the bordering state of Pennsylvania (PA) kept it constant.
- Card and Krueger (1994) evaluated the effect of this change on the employment of low wage workers.
- They collected data on employment at 400 fast food restaurants (big minimum-wage employers) in NJ in February 1992 and November 1992, and the same type of data for eastern PA.

Example 2:

Effect of minimum wage on employment

- Outcome: employment (average per store), observed in both areas, and both right-before and after the change.
- Classic study in labor economics, updated in Card and Krueger (2000) with additional number of years.
- Card and Krueger found that rising the minimum wage didn't cause an employment reduction.
- This article originated much economic and political debate.
- DD estimation has become a very popular method of obtaining causal effects, especially in the US, where the federal structure provides cross state variation in legislation.


I
federalism
(for the natural experiment)

Example 3:

Impact of immigration on labor market

- The impact of the Mariel Boatlift on the Miami Labor Market.
- There has been a debate about the impact of immigration on the labor market opportunities for the less-skilled natives in the US.
- Card (1990) used the Mariel Boatlift of 1980 as a natural experiment to measure the effect of immigration on local labor markets.
  - It was a mass emigration of Cubans who departed from Cuba's Mariel Harbor, motivated by a sharp recession in the Cuban economy, that led the government to allow anyone who wanted to leave to do so.
  - The Mariel Boatlift increased the Miami labor force by 7%.
- Card used data on unemployment for Miami and four comparison cities (Atlanta, Los Angeles, Houston, and Tampa) which showed trend in unemployment similar to Miami.
- The results do not show evidence of deterioration of labor markets due to immigration.

Baseline framework

Canonical DD Setup without Covariates

- Two periods, $t = 1$ (before treatment), $t = 2$ (after treatment).
- Let
  - $D_i = 1$ if the individual is in the treated group, and $D_i = 0$ otherwise.
  - $Y_{ji}(t)$: potential outcomes for treatment status $j = 0, 1$ in period t:
    - $Y_{1i}(t)$: Potential outcome at period t if units were exposed to treatment in period t.
    - $Y_{0i}(t)$: Potential outcome at period t if units were not exposed to treatment in period t.
  - $Y_i(t)$: observed outcome in period t:

$$Y_i(t) = D_i Y_{1i}(t) + (1 - D_i) Y_{0i}(t).$$

Baseline framework

Canonical DD Setup without Covariates

- In the simplest scenario, treatment is only provided after period $t = 1$ (between $t = 1$ and $t = 2$), and there is no anticipation.
- Thus, in $t = 1$, $D_{it} = 0$ always, so that observed outcome in $t = 1$ is

$$Y_i(1) = Y_{0i}(1) = Y_{1i}(1) \text{ (always observed).}$$

- In $t = 2$, observed outcome depends on treatment status:

$$Y_i(2) = D_i Y_{1i}(2) + (1 - D_i) Y_{0i}(2).$$

- The causal estimand of primary interest is the average treatment effect on the treated (ATT) in period $t = 2$,

$$\begin{aligned} \tau &= E[Y_{1i}(2) - Y_{0i}(2) | D_i = 1]. \\ &= \underbrace{E[Y_{1i}(2) | D_i = 1]}_{\text{estimable from data}} - \underbrace{E[Y_{0i}(2) D_i = 1]}_{\text{counterfactual}} \end{aligned}$$

- Identification of the ATT is achieved via three main assumptions:

(1) Stable Unit Treatment Value Assumption (SUTVA): unit i outcomes do not depend on the treatment status of unit $j \neq i$.

$$Y_i(t) = D_i Y_{1i}(t) + (1 - D_i) Y_{0i}(t).$$

This rules out spillover and general equilibrium effects, which would come about when the policy affects outcomes of participants **and** non-participants (i.e. displacement effects).

(2) No-Anticipation: the treatment has no causal effect before its implementation.

- Otherwise the changes in the outcome for the treated group between period 1 and 2 could reflect not just the causal effect in period $t = 2$ but also the anticipatory effect in period $t = 1$:

$$Y_i(1) = Y_{0i}(1) = Y_{1i}(1)$$

(3) Parallel Trends Assumption: The average outcomes for treated and controls would have followed parallel paths over time in the absence of treatment:

- The comparison group has a trend in $E[Y_0]$ that is the same as the counterfactual:

$$E\left[Y_{0i}(2) - Y_{0i}(1) \middle| D_i = 1\right] = E\left[Y_{0i}(2) - Y_{0i}(1) \middle| D_i = 0\right].$$

- Under these assumptions, we can use the average trend followed by control units as counterfactual.
- If we observe several periods before treatment, this assumption is typically checked by showing that trends before treatment coincided for treated and control units.
  - BUT, this is neither necessary nor sufficient for post-treatment parallel trends.
  - Parallel trend is scale-dependent: hold for Y but may not for a nonlinear monotone transformation of Y.

Baseline framework

Identification assumptions

- How can these assumption help us?
- From the definition of the ATT and SUTVA, we have:

$$\begin{aligned} \tau &= E\left[Y_{1i}(2) - Y_{0i}(2) \middle| D_i = 1\right] \\ &= \underbrace{E\left[Y_i(2) \middle| D_i = 1\right]}_{\text{by SUTVA}} - \underbrace{E\left[Y_{0i}(2) \middle| D_i = 1\right]}_{\text{counterfactual}} \end{aligned}$$

- $E\left[Y_{0i}(2) \middle| D_i = 1\right]$ depends on potential outcomes, and our goal is to find ways to "impute" it.
- This is where PT and no-anticipation become crucial.

- First, recall the PT assumption:

$$E\left[Y_{0i}(2) - Y_{0i}(1) \middle| D_i = 1\right] = E\left[Y_{0i}(2) - Y_{0i}(1) \middle| D_i = 0\right]$$

- By simple manipulation, we can write it as:

$$E\left[Y_{0i}(2) \middle| D_i = 1\right] = E\left[Y_{0i}(1) \middle| D_i = 1\right] + E\left[Y_{0i}(2) - Y_{0i}(1) \middle| D_i = 0\right]$$

- Now, exploiting No-Anticipation and SUTVA:

$$E\left[Y_{0i}(2) \middle| D_i = 1\right] = \underbrace{E\left[Y_i(1) \middle| D_i = 1\right]}_{\text{by No-Anticipation and SUTVA}} + \underbrace{E\left[Y_i(2) - Y_i(1) \middle| D_i = 0\right]}_{\text{by No-Anticipation and SUTVA}}$$

- Combining these results together, we have that, under SUTVA + No-Anticipation + PT assumptions, it follows that

$$\begin{aligned}\tau &= E\left[Y_{1i}(2) - Y_{0i}(2) \middle| D_i = 1\right] \\ &= E\left[Y_i(2) \middle| D_i = 1\right] - E\left[Y_i(1) \middle| D_i = 1\right] - E\left[Y_i(2) - Y_i(1) \middle| D_i = 0\right] \\ &= E\left[Y_i(2) - Y_i(1) \middle| D_i = 1\right] - E\left[Y_i(2) - Y_i(1) \middle| D_i = 0\right]\end{aligned}$$

- Lots of restrictions placed on difference-in-differences
  - No-Anticipation: you chose a baseline that is not treated
  - SUTVA: your comparison group outcomes are not affected by the treated outcomes
  - PT: your comparison group has the same trend as the counterfactual
  - Only when you have NA, SUTVA, and PT does DD equal ATT.

Baseline framework

Identification assumptions

- DD estimand is non-parametrically identified.
- Equivalently, the PT assumption can be viewed as a **selection bias stability** assumption:

$$E\left[Y_{0i}(2)|D_i=1\right]-E\left[Y_{0i}(2)|D_i=0\right]=E\left[Y_{0i}(1)|D_i=1\right]-E\left[Y_{0i}(1)|D_i=0\right]$$

- PT allows for selection bias! However, the selection bias has to be the same in both periods.
- Based on the "constant difference" interpretation of the PT assumption, the DD estimator can be written as:

$$\begin{aligned}\tau &= E\left[Y_{1i}(2)-Y_{0i}(2)|D_i=1\right] \\ &= \left(E\left[Y_i(2)|D_i=1\right]-E\left[Y_i(2)|D_i=0\right]\right) \\ &\quad -\left(E\left[Y_i(1)|D_i=1\right]-E\left[Y_i(1)|D_i=0\right]\right)\end{aligned}$$

-
  - What if the composition of the cross-sectional units changes over time?

Estimation

The two-way fixed effects (TWFE) regression specification

- In practice, we typically rely on the following two-way fixed effects (TWFE) regression specification:

$$\begin{aligned} Y_{it} &= \delta_0 + \delta_D D_i + \delta_T T_t + \beta (T_t \times D_i) + u_{it} \\ \beta &\equiv \tau = ATT, \end{aligned}$$

  where we assume that $E(u_{it} \mid D_i, T_t) = 0$.
- Pros of regression-based DD:
  - easy inference (point estimate and standard errors)
  - easy to incorporate covariates
- Cons of regression-based DD:
  - Parametric assumptions, subject to misspecification
    - Abadie (2005) proposed a semiparametric approach
    - Sant'Anna and Zhao (2020) proposed a Double-robust DD
  - However, if you have Staggered DD (i.e., treatment is applied at different times to different groups) and heterogenous effects, TWFE is really bad.

Estimation

The two-way fixed effects (TWFE) regression specification

- More than 2 groups (multiple treatments and multiple controls), and more than 2 periods (pre and post):

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + e_{ist},$$

  D_{st} is a dummy for treated in treatment periods.

- Notice that we could run this on the micro data or aggregate to the state level

$$Y_{st} = \gamma_s + \lambda_t + \beta D_{st} + e_{st}$$

- Both regressions give the same estimates (the second weighted by the number of observations in the cell) since regressors just vary at the group level.
  - Perfect fit¡

Estimation

The two-way fixed effects (TWFE) regression specification

- PT assumption more credible if we condition on covariates.
- We use an augmented specification in which we add further regressors.
  - Notice that only covariates at the state/year level matter for identification.
  - We might want to include individual level covariates, but the within state variation does not matter for identification although may reduce standard errors.
  - But, careful, we need to impose additional linearity and homogeneity assumptions.

Including Leads and Lags in the TWFE model (Event study regressions)

- Most DD applications have several periods before anyone was treated and several periods after
- We can incorporate leads and lags before/after treatment kicks in (event).
  - We can analyze **pre-trends** and treatment effect changes over time after treatment.
- In particular, we can test whether the groups were moving in parallel prior to the treatment
  - If so, then PT assumption more credible
  - Even if pre-trends are the same, one still has to worry about other policies changing at the same time

Including Leads and Lags in the TWFE model (Event study regressions)

- Consider the model with a binary treatment.
  - Let k be the time at which the treatment is being switched on in group s.
  - If we have m pre-treatment periods and q post-treatment periods, regress

$$Y_{ist} = \gamma_s + \lambda_t + \sum_{j=-m}^{q} \beta_j D_{st}^j + e_{ist},$$

The event dummies around the event window $[-m; q]$ take the value 1 if the treated is j periods away in period t from initial treatment and 0 otherwise: $D_{st}^j = 1(t = k + j) \times D_s$.

- Estimated $\beta_j \ \forall j \geq 0$ coefficients are estimated ATT parameters assuming parallel trends and $\beta_j = 0 \ \forall j < 0$ is part of your evidence for that.

Including Leads and Lags in the TWFE model (Event study regressions)

- Example: all states enter into treatment in period $k = 0$. And we have one period before and one period after:

$$\begin{aligned} Y_{ist} &= \gamma_s + \lambda_t + \beta_{-1} D_{st}^{-1} + \beta_0 D_{st}^{0} + \beta_1 D_{st}^{1} + e_{ist}, \\ D_{st}^{-1} &= 1(t - k = -1) = 1(t = k - 1) \\ D_{st}^{0} &= 1(t - k = 0) \\ D_{st}^{1} &= 1(t - k = 1) \end{aligned}$$

Including Leads and Lags in the TWFE model (Event study regressions)

- The DD literature refers to $\beta_{-3}, \beta_{-2}, \beta_{-1}$ as "leads" or anticipatory effects, and $\beta_3, \beta_2, \beta_1$ as "lags" or post-treatment effects, even though they are merely interactions of the treatment indicator with time dummies and are not actually leads and lags of the treatment indicator in a time-series jargon sense.
- Under Parallel pre-trends, and No Anticipation, and SUTVA, then mechanically $\beta_j \forall j < 0$ will be zero.
- Moreover, β_j, $j > 0$ may not be identical for different periods. For example, the effect of the treatment could accumulate over time, so that β_j increases in j.

Including Leads and Lags in the TWFE model (Event study regressions)

- The DD always identifies the sum of two terms:

$$\beta_0 = \overbrace{E\left[Y_{1i}(2) - Y_{0i}(2) \middle| D_i = 1\right]}^{\text{ATT}} + \underbrace{\left\{E\left[Y_{1i}(2) - Y_{0i}(1) \middle| D_i = 1\right] - E\left[Y_{0i}(2) - Y_{0i}(1) \middle| D_i = 0\right]\right\}}_{\text{non-PT bias}}$$

- But this was post-treatment. Still, DD always identify the sum of those terms, even in the pre-period.
- In pre-periods the ATT is implicitly zero, and the only thing that you can be measuring with pre-trend DD coefficients is differential trends:

$$\beta_{-1} = \left\{E\left[Y_{1i}(1) - Y_{0i}(0) \middle| D_i = 1\right] - E\left[Y_{0i}(1) - Y_{0i}(0) \middle| D_i = 0\right]\right\}$$

Including Leads and Lags: Effect of job protection on employment outsourcing

- Autor (2003, JOLE) includes both leads and lags in a DD model analyzing the effect of increased employment protection on the firm's outsourcing of temporary workers.
- Common law in the US enables employers to hire and fire workers at will.
  - In some states, courts have ruled some exceptions limiting employers' discretion, therefore increasing job protection.
  - Different states have passed these court exceptions into laws at different points in time.
- The standard thing to do is to normalize the adoption year to 0.
- Autor then analyzes the effect of these exceptions that increased job protection on the outsourcing of temporary workers, finding a significant and positive substantial effect.

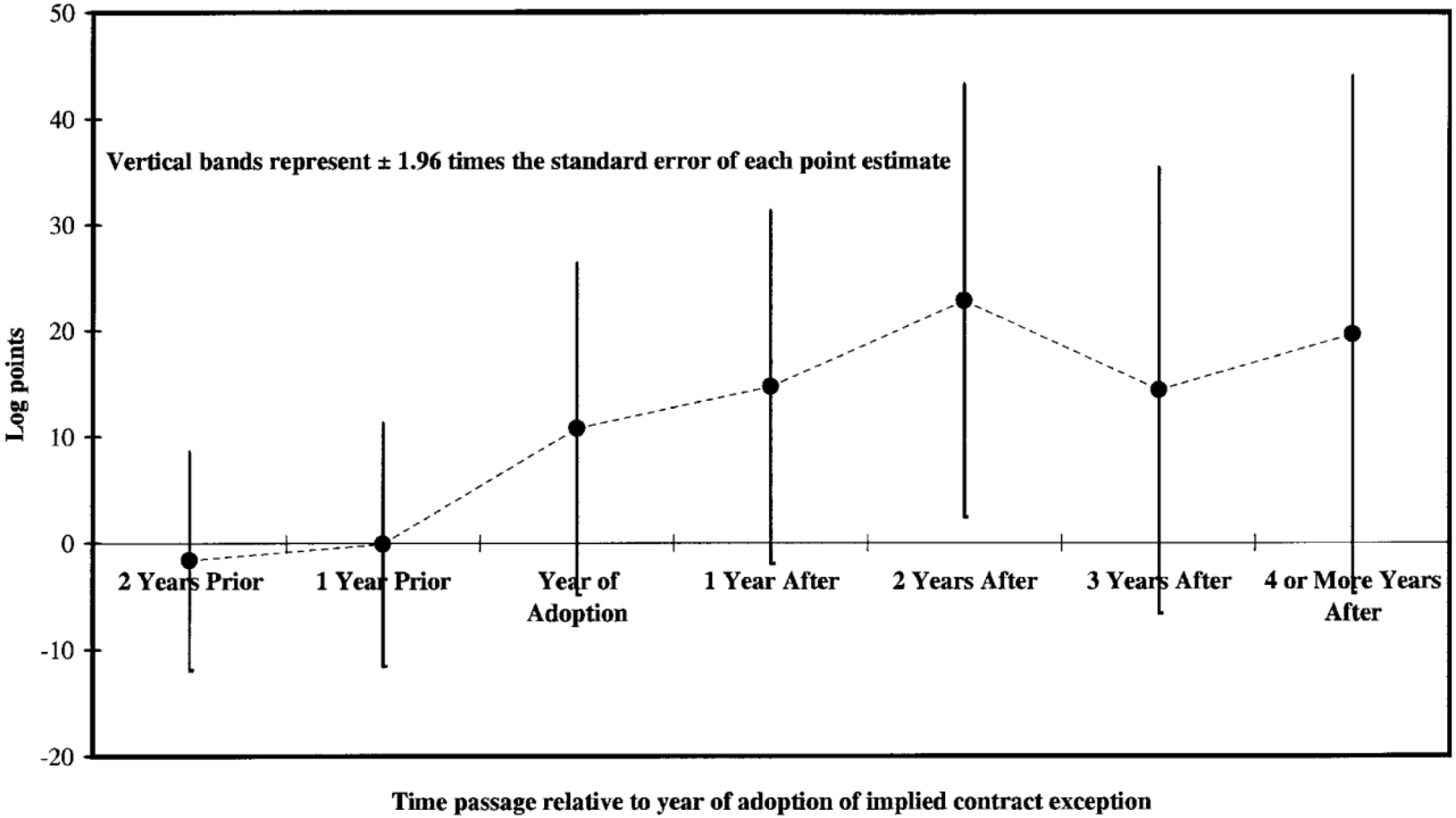

FIG. 3.—Estimated impact of implied contract exception on log state temporary help supply industry employment for years before, during, and after adoption, 1979–95.

Including Leads and Lags: Limitations

- Testing for pre-existing trends is a very natural way to assess the plausibility of the PT assumption
- But it also has several limitations:
- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
    - If other policies change at the same time as the one of interest we can have parallel pre-trends but non-parallel post-trends

Including Leads and Lags: Limitations

- Low power (the probability that we correctly reject H_0 when it is false)
  - even if pre-trends are non-zero, we may fail to detect it statistically
  - Roth (2022, AER: "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends"):
    - Diagnostics of power: If we are relying on a pre-trends test to verify the parallel trends assumption, we'd like that test to have power to detect relevant violations of parallel trends.
    - To assess the power of a pre-trends test, we can calculate its ex ante power: how likely we would be to detect a particular hypothesized violation of parallel trends.
    - The `pretrends` package provides methods for doing these calculations.

Including Leads and Lags: Limitations

- We might also consider the sensitivity analysis provided by Rambachan and Roth (2023, Restud, "A More Credible Approach to Parallel Trends"):
  - Instead of requiring that parallel trends holds exactly, they impose restrictions on how different the post-treatment violations of parallel trends can be.
    - They develop confidence intervals for the treatment effect that take into account the uncertainty over how big the pre-treatment violations of parallel trends are.
    - `HonestDiD` package: Rather than relying on the significance of a pre-test, the HonestDiD approach imposes that the post-treatment violations of parallel trends are not "too large" relative to the pre-treatment violations.

Drawbacks of TWFE model

- This model is extremely popular in applied research
- But it is not clear how β (or β_j) can be interpreted particularly on settings where units are treated at different point in times and the effects are not homogenous.
    - Autor (2003) clues us into the problems of TWFE:

Drawbacks of TWFE model

- The coefficients from standard TWFE models may not represent a straightforward weighted average of unit-level treatment effects:
  - TWFE regressions make both "clean" comparisons between treated and not-yet-treated units as well as "forbidden" comparisons between units who are both already-treated.
  - When treatment effects are heterogeneous, these "forbidden" comparisons potentially lead to severe drawbacks such as TWFE coefficients having the opposite sign of all individual-level treatment effects due to "negative weighting" problems.

Picking a good control group

- It is often the main challenge for the researcher to identify a particularly appropriate comparison group, which satisfies the necessary identifying assumption, that the treated groups in the absence of treatment would behave similarly as the untreated groups.
- Sometimes the choice of treatment group is obvious but sometimes it is not.
- For example, Abadie and Gardeazabal (2003) try to identify the cost in terms of lost output of terrorism in the Basque country region of Spain.
  - They compare growth in the Basque region to other regions in Spain. However, no single other region in Spain is a good comparison group, since the Basque country has relatively more manufacturing than the rest of the country.
  - They propose a method (Synthetic Control) of constructing a counterfactual for the Basque region, using a weighted average of all other Spanish regions. The weights are chosen so as to mimic the pre-terrorism growth trends of the Basque country as closely as possible.

Picking a good control group

- More supporting evidence of the identification assumptions can be obtained by running placebo tests:
  - There should be no effect on units that are plausibly unaffected by treatment.
  - If there is an effect, this may indicate a violation of parallel trends and your estimator is probably picking up some underlying trends
- Plausibly unaffected units can also form an additional control group in a triple difference design (DDD)
  - Some people use DDD as a falsification exercise using a placebo DD to give evidence that the DD design is valid.
  - It can also be used when DD is biased (i.e. PT violation) to "improve" the estimates of the causal effects.

Differences in Differences in Differences (DDD)

- Suppose a state implements a health care policy change aimed at the elderly (65+), and the response variable, Y, is a health outcome.
- Alternative 1: Use data only on people in the state with the policy change, both before and after the change.
  - Control group: people under 65 in the state with the policy change.
    Treatment group: people above 65 in the state with the policy change.
  - Potential problem: other factors unrelated to the state's policy change might affect the health of the elderly relative to younger population (e.g., changes in health care emphasis at the federal level).
- Alternative 2: Use data only on elderly people in the state with the policy change and use another state as the control group.
  - Control group: elderly people from the non-policy state.
  - Potential problem: changes in the health of the elderly might systematically differ across states, e.g., due to income, environmental and wealth differences, rather than the policy change.

Differences in Differences in Differences (DDD)

- More robust Alternative: Use both a different state and a control group within the treatment state.
- In the two-period case, let S the binary variable for the treatment state, and G the binary variable for elderly people (so that treatment group is that with $S \times G = 1$). Then:

$$\rho = [\underbrace{E\left(Y(2)-Y(1)|\, S=1, G=1\right)}_{\text{Elderly in treatment state}} - \underbrace{E\left(Y(2)-Y(1)|\, S=1, G=0\right)}_{\text{Non-elderly in treatment state}}]$$
$$-[\underbrace{E\left(Y(2)-Y(1)|\, S=0, G=1\right)}_{\text{Elderly in control state}} - \underbrace{E\left(Y(2)-Y(1)|\, S=0, G=0\right)}_{\text{Non-elderly in control state}}]$$

which exploits the difference between states and between control and treated groups in the same state.

- Assumption: the bias of the true DD is the same as the bias of the placebo DD

Differences in Differences in Differences (DDD)

- In the two-period case:

$$\begin{aligned} Y &= \beta_0 + \beta_1 G + \beta_2 S + \beta_3 (G \times S) \\ &\quad + \beta_4 T + \beta_5 (T \times G) + \beta_6 (T \times S) \\ &\quad + \theta (T \times G \times S) + u \end{aligned}$$

  The coefficient of interest is now θ, corresponding to the triple interaction term $(T \times G \times S)$. This parameter is call **Difference-in-difference-in differences (DDD)**.

- The DDD estimate is the difference between the DD of interest and the placebo DD (that is supposed to be zero).
  - But....if the placebo DD is non-zero, it might be difficult believe that the DD removes all the bias.
- Recent reference: Olden and Moen (2022, The Econometrics Journal).

DDD Example: Effect of Mandated Benefit on Wage

- Gruber (1994, AER) estimated the effect of mandated maternity benefits on log (wage) using the CPS.
  - Main interest: Whether maternity benefits lower the wage of childbearing-age women.
- Three dimensions:
  - Treatment states that passed such laws ($S = 1$): IL, NJ, and NY. Also, some control states were picked.
  - Treatment group ($G = 1$) individuals: workers "at risk" of having a child (married childbearing-age women). Control ($G = 0$) individuals: those directly unaffected by the law (persons over 40 or single males of age 20-40).
  - Treatment period: individual data two years before (1978,1979) and after (1981,1982) the legislation were taken.

DDD: Mandated Benefit Effect on Wage

- The sample size varies depending on the year and state, but it is around $1,500$ for the treatment states and $5,000$ for the control states.
- Some covariates such as education, experience, marital status, race and industry are also included.

	$E\left(\Delta \log(wage) \middle\vert \cdot\right)$
$G = 1, S = 1$	-0.034
$G = 1, S = 0$	0.028
DD1	-0.062
$G = 0, S = 1$	-0.011
$G = 0, S = 0$	-0.003
DD2	-0.008
DDD	-0.054

DDD: Mandated Benefit Effect on Wage

- There was a 3.4% fall in the wages for women in the treated states over this period, compared to a 2.8% increase in the control states.
- Thus, there was a 6.2 pp. relative fall, and this would be the DD estimate of the law's impact.
- However, if there was a distinct labor-market shock to the treated states over this period, this estimate does not identify the impact of the law.
- In the bottom panel, we find that there is a fall in wages for men in the treated states relative to the other states of 0.8 pp.
- This suggests that we should control for state specific shocks in estimating the impact of the law.

Violations of Parallel Trends: functional form dependence

- A reason to be skeptical of PT is that its validity is functional form dependent:
  - i.e., if it holds for the level of Y , it may not hold for monotone transformations of Y.
- Consider an example:
  - In period 1: $E\left[Y_{0i}(1)|D_i=0\right]=10; E\left[Y_{0i}(1)|D_i=1\right]=5$.
  - In period 2: $E\left[Y_{0i}(2)|D_i=0\right]=15$
  - If treatment hadn't occurred, would treated units' outcome have increased by 5 also (PT in levels), that is, 100%

$$\begin{aligned} 5 &= \underbrace{E\left[Y_{0i}(2)|D_i=1\right]-E\left[Y_{0i}(1)|D_i=1\right]}_{} \\ &= \quad \mathbf{10} \quad - \quad 5 \end{aligned}$$

- Or would they have increased by 50% (PT in logs)?

$$\begin{aligned} \ln\frac{15}{10} &= \underbrace{E\left[\ln(Y_{0i}(2))|D_i=1\right]-E\left[\log(Y_{0i}(1))|D_i=1\right]}_{} \\ &= \quad \ln(\mathbf{7.5}) \quad - \quad \ln(5) \end{aligned}$$

Violations of Parallel Trends: functional form dependence

- Important issue because there are often many possible ways to parameterize the outcome:
  - Earnings could be measured in: levels, logs, percentiles relative to the national distribution
- We may often be unsure which is the "right" parameterization for identification.
- Therefore, desirable if identification doesn't depend on functional form.

Violations of Parallel Trends: functional form dependence

- This functional form dependence is the starting point for the so-called "changes-in-changes model" (CC) that has been proposed by Athey and Imbens (2006, Econometrica).
- The idea is to compare the cumulative distribution functions (cdfs) of the outcomes in the four groups.
- The key difference to the standard DD approach is that the assumptions made as well as the information exploited for identification and estimation comes from the whole outcome distribution and not just from the first moments.
- This model is considerably more general than the standard DD model: Its assumptions are invariant to monotone transformations of the outcome.

Violations of Parallel Trends: functional form dependence

- Roth and Sant'Anna (2023, Econometrica) provide characterizations of when parallel trends is insensitive to functional form, in the sense that it holds for all strictly monotonic transformations of the outcome.
- They show that parallel trends is insensitive to functional form if and only if a "parallel trends"-type condition holds for the entire cumulative distribution function (CDF) of Y_0.

Violations of Parallel Trends: functional form dependence

- They further show that this condition can be satisfied if and essentially only if we are in one of three cases:
  - (i) when treatment is as-if randomly assigned,
  - (ii) when the distribution of Y_0 is stable over time, and
  - (iii) a hybrid of the first two cases in which the population is a mixture of a subpopulation that is (as-if) randomized between treatment and control and another subpopulation that has stable untreated potential outcome distributions over time.
- They conclude that to obtain any consistent estimator of the ATT, one must impose assumptions that either are sensitive to functional form or that identify the full distribution of Y_0.
  - There are fundamental trade offs between dependence on functional form and modeling the full distribution

Standard errors in DD strategies

- Problem with DD: underestimated standard errors.
- Many papers using a DD strategy use data from many years (not only 1 pre and 1 post period).
- Bertrand, Duflo, and Mullainathan (QJE, 2004) point out that conventional standard errors often understate the standard deviation of the estimators:

$$Y_{ist} = \mu_s + \delta_t + \tau D_{st} + \varepsilon_{ist}$$

Standard errors in DD strategies

- Three well-known facts about how serial correlation biases OLS estimates of st. errors:
  - Positive serial correlation of the error term will under-estimate the st. error, while negative serial correlation will cause overestimation.
    - Intuition: Positive serial correlation means that there is less information in each new year of data than OLS assumes.
  - Correlation in the explanatory variable
    - Typically in DD D_{st} is very serially correlated (for the treated it is equal to 0 until one year where it turns to 1 and then stays at 1)
  - As T increases, the bias also increases
- A large fraction of published DD papers report $t-$statistics around 2.
  - Some of these findings may not be as conclusive as previously thought

Standard errors in DD strategies

- How can one address these issues?
  - One solution: collapse data into one observation per unit "before", and one "after" treatment
    - Takes care of serial correlation, but power declines fast
  - No collapsing: conventional variance estimator often led to over-rejection, thus:
    - Block-bootstrap standard errors: a variant of bootstrap which maintains the autocorrelation structure by keeping all the observations that belong to the same group (e.g., state) together.
    - Clustering standard errors at the group level, e.g., state (option `cl(state)` in Stata), provided that the number of groups is large!
    - Parametric modelling of the serial correlation: specify an auto-correlation structure for the error term, estimate its parameters, and use these parameters to compute standard errors.