Chapter 8: Causal Inference with Panel Data I

Frontier Topics in Empirical Economics

Lecture notes based on slides by Zibin Huang
College of Business, Shanghai University of Finance and Economics

November 2025

Contents

1 Introduction

In the previous chapters we mostly worked with cross-sectional data (except for chapter 7): one observation per unit, all measured at a single point in time. This chapter adds a new dimension. When we observe the same units repeatedly over time, we have *panel data*, and a whole new source of identifying variation opens up: variation across time *within* the same individual, firm, city, or country. Panel data let us compare a unit to itself, which eliminates every confounder that stays constant over time.

This chapter develops the panel data toolkit for causal inference in four steps. First, we introduce the fixed effect model and its three equivalent or nearly equivalent estimators: the fixed effect estimator, the dummy variable estimator, and the first difference estimator. Second, we combine unit and time fixed effects into the two way fixed effect (TWFE) model and study its most famous special case, the difference in differences (DID) design, using the minimum wage study of Card and Krueger (1994) as the running example. The key identifying assumption is parallel trends, and we discuss the traditional ways of assessing it: plotting raw trends and running event study regressions. Along the way we emphasize a question that every applied researcher must be able to answer: *which variation in the data identifies the parameter?*

Third, we present three modern extensions. The recent literature has taught us to be much more careful about pre-trend testing: (1) Roth (2022) shows that conventional event study pre-tests often have low power, and that pre-testing itself distorts inference. (2) When only a few units are treated with no clean control group, the synthetic control method from Abadie et al. (2010) constructs an artificial control unit as a weighted combination of untreated units. (3) When a second within-state comparison group is available, triple differences (DDD) can difference out non-parallel trends.

Fourth, a preview: the next chapter continues with panel data and treats staggered adoption designs, where treatment turns on at different times for different units and the TWFE regression can behave badly. The tools of this chapter are the foundation for that discussion.

2 Fixed Effects

2.1 Setup

Consider a classic question: what is the impact of military service on wages? Let i index individuals and t index time. We observe wage Y_{it}, military service status D_{it}, and time varying covariates X_{it}. The worry is an unobserved individual attribute, ability A_i, that affects both the decision to serve and later wages. Assume a constant treatment effect and a linear CEF:

$$\begin{aligned} Y_{it} &= \alpha + \rho D_{it} + A_i'\gamma + X_{it}'\beta + \epsilon_{it}, \\ E[Y_{it} \mid A_i, X_{it}, D_{it}] &= \alpha + \rho D_{it} + A_i'\gamma + X_{it}'\beta. \end{aligned} \tag{1}$$

The unobserved confounder A_i has two crucial properties: it is *unobserved*, so we cannot control for it directly, and it is *time invariant*, which is what panel data will let us exploit. The identifying assumption is that selection into treatment is fully captured by A_i and X_{it}:

$$\epsilon_{it} \perp\!\!\!\perp D_{it} \mid A_i, X_{it}.$$

If we could observe A_i, we would simply control for it. We cannot, but with panel data we do not need to. There are three simple ways to estimate ρ.

2.2 Three Estimators

Method 1: The fixed effect (within) estimator. The FE estimator is a deviation from means estimator. Take individual level time averages of both sides of (1):

$$\bar{Y}_i = \alpha + \rho \bar{D}_i + A_i'\gamma + \bar{X}_i'\beta + \bar{\epsilon}_i,$$

where $\bar{Y}_i = \frac{1}{T}\sum_t Y_{it}$ and so on. Now subtract this from the original equation:

$$\begin{aligned} Y_{it} - \bar{Y}_i &= \rho\,(D_{it} - \bar{D}_i) + (A_i'\gamma - A_i'\gamma) + (X_{it} - \bar{X}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i) \\ &= \rho\,(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i). \end{aligned} \tag{2}$$

The unobserved time invariant term $A_i'\gamma$ is identical in every period for individual i, so it equals its own time average and cancels. Running OLS on the demeaned equation (2) delivers ρ without ever observing A_i.

Method 2: The dummy variable estimator. Alternatively, saturate the individual dimension with a full set of individual dummies. Collect the constant and the individual specific term into one intercept per person:

$$Y_{it} = \alpha_i + \rho D_{it} + X'_{it}\beta + \epsilon_{it}, \qquad \alpha_i \equiv \alpha + A'_i\gamma. \tag{3}$$

The unobserved A_i is absorbed into the estimated dummy coefficient α_i. Running OLS on (3) with N individual dummies again delivers ρ. Because OLS on group indicators just fits the group mean, then By the Frisch–Waugh–Lovell logic, regressing on a full set of individual dummies is numerically identical to demeaning within individuals. Thus, the dummy estimator and the FE estimator produce exactly the same point estimates, standard errors, and other main statistics. For this reason the two are used interchangeably and jointly called the "FE model."

Method 3: The first difference (FD) estimator. A third route removes A_i by differencing across adjacent periods. Let $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$. Subtracting the period $t-1$ equation from the period t equation:

$$\Delta Y_{it} = \rho\, \Delta D_{it} + \Delta X'_{it}\beta + \Delta\epsilon_{it}. \tag{4}$$

Again the time invariant A_i drops out, this time through differencing rather than demeaning.

2.3 Comparing FE, Dummy, and FD

All three estimators exploit the same source of variation: changes over time within the same person. But they are not always numerically identical, and the differences matter in practice.

FE, Dummy, and FD: How They Relate

- **FE and Dummy are always identical.** Point estimates, standard errors, and the main regression statistics coincide. We refer to both as the FE model.
- **FE and FD are identical when $T = 2$.** With two periods, the within deviation and the first difference are the same transformation up to scale.
- **FE and FD differ when $T > 2$,** and the choice between them is an inference question about the error process:

- If ϵ_{it} are serially *uncorrelated* shocks, FE is more efficient. First differencing manufactures serial correlation in $\Delta\epsilon_{it}$ (adjacent differences share a common ϵ term), which FD then has to live with. Serial correlation invalidates the usual statistical inference based on i.i.d. assumption and damages efficiency.
- If ϵ_{it} follows a *random walk*, that is, $\epsilon_{it} - \epsilon_{it-1} = \Delta\epsilon_t$, $\Delta\epsilon_{it}$ is i.i.d., then FD is better. Because the original errors are serial correlated but the differenced errors $\Delta\epsilon_{it}$ are not.

3 Difference in Differences and TWFE

3.1 The Two Way Fixed Effect Model

With panel data we can usually control for both individual and time fixed effects:

$$Y_{it} = \rho D_{it} + X_{it}'\beta + \lambda_t + \alpha_i + \epsilon_{it}. \tag{5}$$

This is the **two way fixed effect (TWFE) model**. The individual effects α_i absorb everything constant within a unit, and the time effects λ_t absorb everything common to all units in a period (macro shocks, seasonality, national policy).

The **difference in differences (DID)** design is a special case of the TWFE model. In a typical DID application, some policy is implemented at a higher level of aggregation than the individual, such as a province or a city, and D_{it} is a binary indicator for whether individual i at time t is covered by the policy. We control for unit (or province, or city) fixed effects and time fixed effects, and read the policy effect off ρ.

3.2 The Card and Krueger Setting

The canonical DID example is Card and Krueger (1994) on the minimum wage. On April 1, 1992, New Jersey raised its state minimum wage. In neighboring Pennsylvania, nothing happened. Card and Krueger collected employment data from fast food restaurants in New Jersey and Pennsylvania in February 1992 (before the change) and November 1992 (after the change). The question is whether the minimum wage increase reduced employment, as a simple competitive model of the labor market predicts.

For restaurant i in state s at time t, denote employment by Y_{ist} and the policy dummy by D_{st}. In this two state, two period setting, $D_{st} = NJ_s \cdot d_t$, where NJ_s indicates New Jersey and d_t indicates the post policy period. Our target is the average treatment effect on the treated,

$$E[Y_{1ist} - Y_{0ist} \mid D_{st} = 1].$$

The problem is the usual missing counterfactual: after the policy, we observe only the treated outcome Y_{1ist} for New Jersey restaurants. How would employment in New Jersey have evolved *without* the policy?

3.3 The Parallel Trend Assumption

The DID answer: use the untreated state as the control group, and assume the treated state would have followed the same time path as the control state in the absence of the policy.

Assumption 3.1 (Parallel Trends)**.** There are no differential trends across treated and untreated states absent the policy change:

$$E[Y_{0ist} \mid s, t] = \gamma_s + \lambda_t. \tag{6}$$

The untreated potential outcome decomposes into a state component and a time component, with no interaction: no term like η_{st} appears in $E[Y_{0ist} \mid s, t]$.

Assumption 3.1 is additive separability of the untreated outcome in state and time. States can differ in levels (γ_s), and time can shift all states together (λ_t), but there is no state specific trend. This is what makes Pennsylvania an acceptable stand-in for the New Jersey counterfactual.

3.4 Identification

Under parallel trends, the policy effect δ is identified by the regression

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \epsilon_{ist}. \tag{7}$$

To see why, take two differences. The *first difference* compares the same state across time:

$$E[Y_{ist} \mid s = PA, t = Nov] - E[Y_{ist} \mid s = PA, t = Feb] = \lambda_{Nov} - \lambda_{Feb}, \tag{8}$$

$$E[Y_{ist} \mid s = NJ, t = Nov] - E[Y_{ist} \mid s = NJ, t = Feb] = \lambda_{Nov} - \lambda_{Feb} + \delta. \tag{9}$$

Each within state change nets out the state level γ_s but still contains the common time shift. The *second difference* compares the two state level changes:

$$(9) - (8) = \delta. \tag{10}$$

The common time trend cancels, and what remains is the policy effect. We difference once over time and once over states, hence the name difference in differences. The logic is drawn in Figure 1: the control group's observed trend is used to construct the treated group's counterfactual, and the treatment effect is the gap between the observed and counterfactual treated outcomes after the policy.

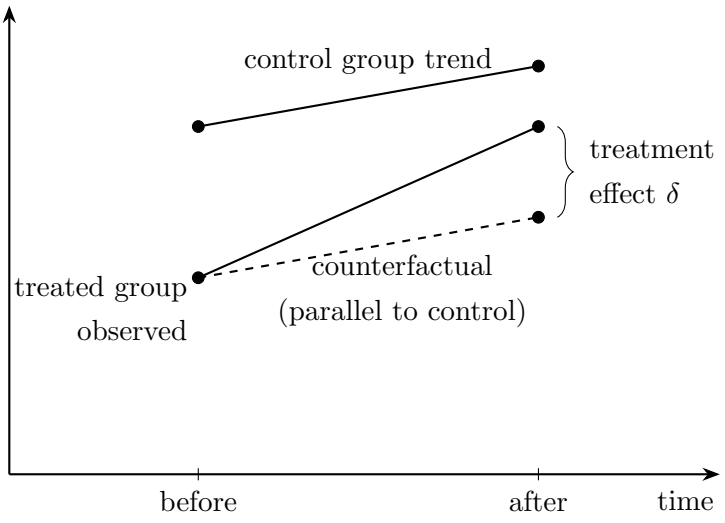

Figure 1: The DID identification logic. The treated group's counterfactual path (dashed) is constructed by shifting the control group's trend down to the treated group's pre-period level. The treatment effect is the post-period gap between the observed and counterfactual treated outcomes.

3.5 Testing the Parallel Trend Assumption

After the policy is implemented at time t_0, we can no longer observe Y_0 for the treated group, so the parallel trend assumption is intrinsically untestable in the post period. What we *can* examine is the pre period: did the two groups trend in parallel before t_0? This is the **pre-trend test**, and there are two simple ways to do it.

Method 1: Draw the picture. Plot the average outcome for treated and control groups over time and look at the pre-policy period. In the Card and Krueger follow-up data, employment in New Jersey and Pennsylvania fast food restaurants tracks each other reasonably well before the policy dates, which supports the design. In other applications the picture can be damning: when the treated group is already on a visibly different trajectory before the policy (rising while the control is flat, say), the parallel trend assumption loses credibility no matter what the post period shows. Always draw this picture first. It is the cheapest and most honest diagnostic available.

Method 2: The event study regression. Suppose we have data from $-T$ to T' in event time, with the policy implemented at $t = 0$ for the treated group. Let D_s be the treated group dummy. Run

$$Y_{ist} = \gamma_s + \lambda_t + \sum_{\tau \neq -1} \mathbf{1}(t = \tau)\, \delta_\tau D_s + \epsilon_{ist}. \tag{11}$$

Each event time τ gets its own coefficient δ_τ, which measures the difference in trends between treated and untreated groups at τ, relative to the omitted baseline. The period just before the policy, $\tau = -1$, is conventionally omitted as the baseline. Note the contrast with the static DID: there we split time into just two bins, before and after, with $D_{st} = D_s d_t$ switching on only for the treated group after implementation; in the event study we give every time point its own parameter.

The estimated δ_τ are then plotted against event time with confidence intervals, as in Figure 2. The reading is:

- Points *before* $t = 0$ insignificant and close to zero $\Rightarrow$ the pre-trend is parallel.
- Points *after* $t = 0$ show the dynamic policy effect, period by period.

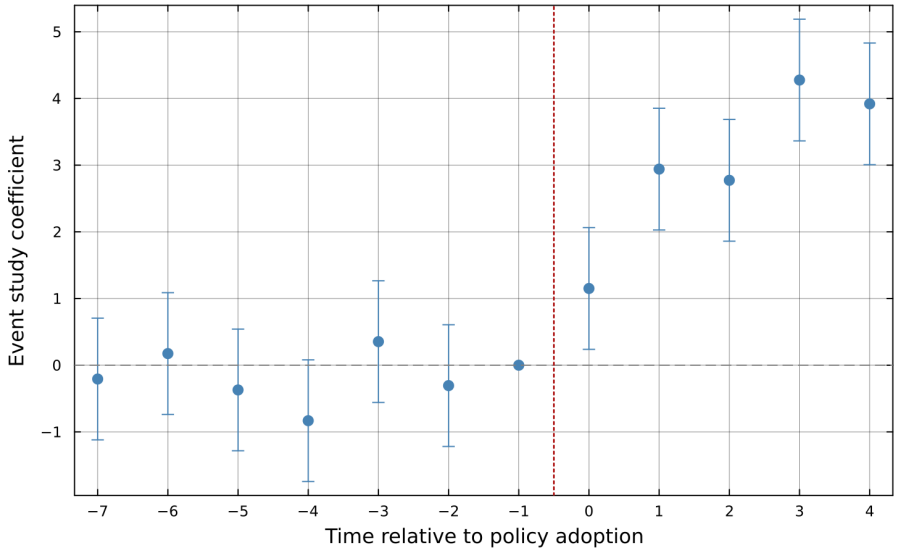

Figure 2: An event study coefficient plot from a simulated DID panel. Each point is $\hat{\delta}_\tau$ with a 95% confidence interval; the period $\tau = -1$ is the omitted baseline (normalized to zero, shown without an interval). Pre-period coefficients are small and insignificant, consistent with parallel trends; post-period coefficients trace out a dynamic treatment effect.

A Traditional DID Procedure

1. Draw the changes of Y over time for treated and control groups as descriptive evidence.
2. Run the main DID regression.
3. Run the event study regression to check the pre-trend and to trace the dynamic effect.
4. Remember to cluster your standard errors. (More details on clustering in Chapter 11.)

4 Identification Source: Which Variation Are You Using?

When you are doing causal research, a central question is: *what kind of variation is used to identify the causal effect?* This question is very, very important. It determines how you can interpret your results, it determines which assumption you are actually invoking, and it therefore determines how you should defend your research design.

With cross-sectional data the answer is usually easy. With panel data it becomes complicated, because there are more dimensions along which variation lives. Applied papers control for many fixed effects at different levels, sometimes combined with IV, RD, or other designs. Whatever the combination, you should always be able to state precisely where the identifying variation comes from.

Consider a simple example: the effect of working experience on wages, for individual i from family j at time t:

$$wage_{ijt} = \beta_0 + \beta_1 exp_{ijt} + \epsilon_{ijt}. \tag{12}$$

The fixed effects you include determine the comparison the regression makes.

Fixed Effects and the Variation They Leave

For regression (12):

- **Time FE only:** you are using variation across individuals and families (i, j level) within the same year.
- **Individual FE only:** you are using variation across time (t level) for the same person.
- **Family FE and time FE:** you are using variation across individuals within the same family in the same year ($i \mid j$ level).
- **Individual FE and time FE:** you are using variation in time trends across different people ($i \times t$ level).

Each row of the box corresponds to a different comparison, a different exogeneity requirement, and a different economic interpretation. Two researchers running "the same" regression with different fixed effects are, in identification terms, running different research designs.

5 Pre-Trend Testing Reconsidered

We now turn to the first of three modern extensions of DID: a more careful look at pre-trend testing. Is the event study a perfect tool for testing parallel trends? It is good, but far from perfect. The key reference is Roth (2022), which examines what event study pre-tests actually deliver in practice.

5.1 Problem 1: Low Statistical Power

The first problem is that pre-trend tests frequently have low power: they are likely to commit type II errors, failing to detect violations that are large enough to matter. Pre-existing trends that produce meaningful bias in the DID estimate may simply not be detected by the test.

To make this concrete, suppose parallel trends is violated linearly: the differential trend between treated and control groups is $\delta_{1t} - \delta_{0t} = \gamma t$ for some slope γ. Roth (2022) runs Monte Carlo simulations calibrated to data from a large set of published papers and asks: how big must the violation be before the standard pre-test detects it reliably? The answer is sobering. For the test to detect the violation 80% of the time, the implied bias in the treatment effect estimate has to be roughly as large as the estimated treatment effect itself. In other words, a pre-existing trend big enough to explain the entire headline result can still slip past the test most of the time. The bias has to be very large for you to detect it.

Why does this happen? It goes back to the nature of statistical testing. Recall the two error types:

- **Type I error:** H_0 is true but we reject it. Its probability is the significance level α.
- **Type II error:** H_0 is false but we fail to reject it. Its probability is β.
- **Power:** the probability of rejecting H_0 when it is false, $1 - \beta$.

The four possible outcomes of a hypothesis test are summarized in Figure 3.

Null hypothesis is . . .	**True**	**False**
Rejected	**Type I error** False positive Probability $= \alpha$	**Correct decision** True positive Probability $= 1 - \beta$
Not rejected	**Correct decision** True negative Probability $= 1 - \alpha$	**Type II error** False negative Probability $= \beta$

Figure 3: Type I and type II errors. the four possible outcomes of a hypothesis test, with the probability of each outcome.

There is an unavoidable trade-off. When you choose a critical value, moving it in one direction raises α, and moving it in the other raises β. You can decrease the type I error rate only by increasing the type II error rate, and vice versa. If you want H_0 to be rejected less easily, you must tolerate a larger probability of a false negative.

Figure 4 visualizes the trade-off. The critical value splits the horizontal axis between the two sampling distributions: the right tail of the null distribution beyond the critical value is the type I error rate α, and the left tail of the alternative distribution below it is the type II error rate β. Pushing the critical value to the right means you are setting a higher bar to reject the null hypothesis. It shrinks α but inflates β. Pulling it to the left does the reverse. There is no choice that shrinks both.

Figure 4: The probability of making type I and type II errors. The null sampling distribution (blue) and the alternative (green) overlap; at the critical value, the shaded right tail of the null is α and the shaded left tail of the alternative is β. Moving the critical value trades one error rate against the other.

Here is the catch. Traditional hypothesis testing is deliberately conservative about *rejecting* H_0: we cap α at 10%, 5%, or 1%, and accept whatever β results. That convention is designed for situations where a false rejection is the costly mistake. But in pre-trend testing the roles are reversed. The null hypothesis is "trends are parallel," and the researcher is hoping *not* to reject it. What we should be conservative about is *failing to reject* a false null, that is, we care about power. A testing convention that makes rejection very hard mechanically makes it very likely that your design passes

the pre-test, whether or not parallel trends actually holds.

5.2 Problem 2: Pre-Testing Distorts Inference

The second problem is subtler. In practice, researchers run the pre-test and proceed only if the design passes. This turns the pre-test into a sample selection device operating across research designs, and conditioning on passing has consequences:

2a. **Variance understatement.** Conditional on having passed a pre-test, the usual standard errors understate the true sampling variability of the estimator.

2b. **Bias exacerbation.** If parallel trends is in fact violated, conditioning on passing the event study test can make the bias of the point estimate *worse*, not better. The draws in which the violation happens to look small in the pre-period, which are exactly the draws that pass, are systematically unrepresentative. Roth (2022) shows the bias is certainly exacerbated in common configurations, such as monotone trend violations with homoskedastic errors.

The overall effect of pre-trend testing is therefore ambiguous: it screens out some non-parallel designs, which is good, but among designs with a real violation that survive the screen, it makes the surviving estimates more biased, which is bad.

5.3 Practical Suggestions

Practical Suggestions Following Roth (2022)

- **Most important: always use your economic knowledge to assess the parallel trend assumption.** The pre-test is a diagnostic, not a substitute for thinking about why the treated and control groups should trend together.
- Do not conclude you are safe just because the event study does not detect any unparalleled pre-trend.
- Do power calculations against economically relevant violations of parallel trends.
- If you know the functional form of the differential trend, control for it directly.
- Conduct sensitivity analysis using Rambachan and Roth (2023): compute bounds on your estimates under a specified amount of parallel trend violation, and report how large a violation would be needed to overturn the conclusion.

6 Synthetic Control

6.1 Main Idea

In previous analysis, we consider micro-level units like individuals and restaurants. In a large micro-level dataset, it is not hard to find comparable treated and control groups. However, what if the treated unit is only one country or one province? It may be hard to find another proper untreated country or province that is similar enough. What should we do? Just create one synthetic control unit. This is the synthetic control method, developed in Abadie et al. (2010, 2015) and reviewed in Abadie (2021). At heart it is a matching method applied to aggregate units.

When the units of observation are a small number of aggregate entities, such as states or countries, a combination of unaffected units often provides a more appropriate comparison than any single unaffected unit alone. Rather than picking one control state, we construct a weighted average of many control states that mimics the treated state as closely as possible before the treatment.

The canonical application is Abadie et al. (2010) on California's Proposition 99, a large scale tobacco control program implemented in 1988 that combined a 25 cent per pack excise tax with restrictions such as a ban on cigarette vending machines in public areas accessible to juveniles. The empirical problem is that California's cigarette sales were not trending in parallel with other states before 1988, and even the average of all other states tracks California poorly. No single state, and not the national average, is a credible control. The solution: combine the other states into a man-made "synthetic California" whose pre-1988 sales path matches the real California almost exactly. After 1988, the gap that opens between real and synthetic California estimates the effect of Proposition 99.

6.2 Formal Setting

Suppose we have units $j = 1, 2, \ldots, J+1$ (provinces, states, countries) observed over T periods. Treatment starts at period T_0, and $j = 1$ is the treated unit. The remaining units $j = 2, \ldots, J+1$ are called the **donor pool**: the synthetic control will be built from them. For each unit we observe a vector of characteristics X_{kj}, which may include pre-treatment values of the outcome itself.

The Predictors Must Be Unaffected by Treatment

The characteristics X_{kj} used to construct the synthetic control must be unaffected by the treatment. Pre-treatment outcomes can be included; post-treatment outcomes must never be.

Define potential outcomes Y_{jt}^I (with intervention) and Y_{jt}^N (without intervention). The treatment effect of interest for the treated unit is

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \quad \text{for } t > T_0,$$

which can vary across time.

Definition 6.1 (Synthetic Control Estimator)**.** A synthetic control is a weighted average of the units in the donor pool. Given weights $w_j \geq 0$ for $j = 2, \ldots, J+1$, the synthetic control estimate of the treated unit's untreated outcome is

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

and the estimated treatment effect is

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N.$$

6.3 Choosing the Weights

How are the weights determined? Stack the characteristics of the treated unit into X_1 and those of the donors into X_0, with $X = (X_1, X_2, \ldots, X_h, \ldots, X_k)$ indexing the predictor variables. The unit weights $W = (w_2, \ldots, w_{J+1})'$ are chosen to minimize the weighted Euclidean distance

$$\|X_1 - X_0 W\| = \left(\sum_{h=1}^{k} v_h \left(X_{h1} - w_2 X_{h2} - \cdots - w_{J+1} X_{hJ+1} \right)^2 \right)^{1/2}. \tag{13}$$

We are searching for the combination of donors that mimics our treated unit the best, predictor by predictor. Watch out for the two different sets of weights in (13):

- w_j is the weight assigned to each *unit* (state) in forming the synthetic control.

- v_h is the weight assigned to each *characteristic* (predictor) when we compute the distance that determines w. It encodes how important each predictor is for the match.

A concrete illustration comes from the German reunification application of Abadie et al. (2015): how would West German GDP have evolved without the 1990 reunification? The treated unit is West Germany; the donor pool is a set of other OECD countries. The predictor weights v are attached to economic growth predictors such as GDP per capita, trade openness, inflation, industry share, population education level, and the investment rate. The resulting synthetic West Germany matches the real West Germany's pre-1990 predictor means far better than the OECD average or any single country. The country weights w turn out to concentrate on a handful of donors (in the published application: Austria, the United States, Japan, Switzerland, and the Netherlands, with all other weights at zero), and the synthetic Germany tracks real West German GDP per capita almost perfectly for three decades before reunification.

6.4 Choosing v by Cross-Validation

The predictor weights v are a tuning parameter, and they are chosen by a training and validation split of the pre-treatment data:

1. **Divide** the pre-treatment sample into two parts: a training period $t = 1, \ldots, t_0$ and a validation period $t = t_0 + 1, \ldots, T_0$.
2. **For every candidate** V, compute the unit weights $\tilde{w}(V)$ by minimizing (13) on the training data.
3. **Select** V^* as the candidate that minimizes the mean squared prediction error on the validation data:

$$\sum_{t=t_0+1}^{T_0} \left(Y_{1t} - \tilde{w}_2(V) Y_{2t} - \cdots - \tilde{w}_{J+1}(V) Y_{J+1,t} \right)^2.$$

4. **Recompute** the final unit weights W using V^* by $\tilde{w}(v)$ on the validation data.

This is exactly the logic of model selection by cross-validation from Chapter 3, transplanted into the comparative case study world: fit on one part of the data, evaluate predictive performance on another, and pick the tuning parameter that predicts best

out of sample.

Figure 5 illustrates the method on a simulated example: the synthetic control tracks the treated unit closely through the pre-period, then a gap opens after the treatment date, and that gap is the estimated effect.

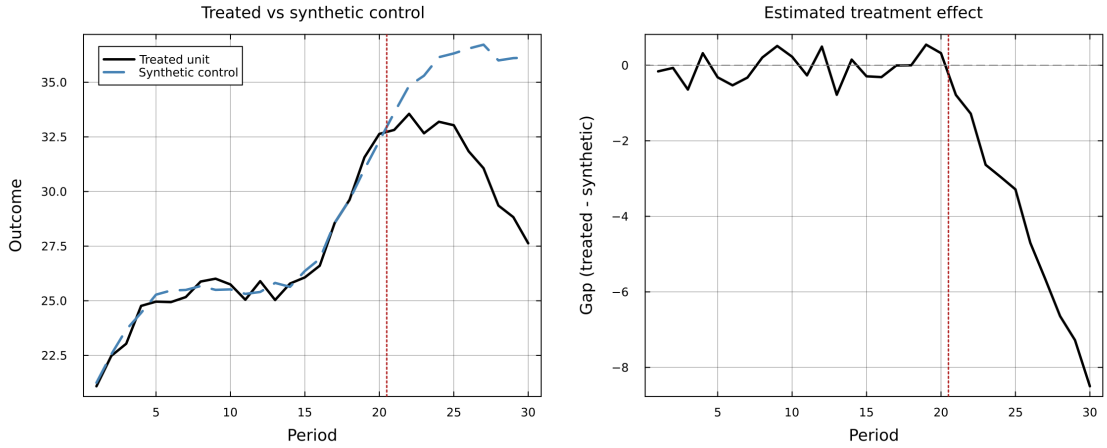

Figure 5: Synthetic control in a simulated example. Left: outcome paths of the treated unit and its synthetic control, with the treatment date marked by the dotted vertical line. The synthetic control matches the treated unit closely before treatment. Right: the gap between treated and synthetic outcomes, which estimates the (here negative and widening) treatment effect.

Things to Remember for Synthetic Control

- Post-treatment outcomes cannot be used in X. Pre-treatment outcomes can.
- There can be bias if there are unobserved endogenous characteristics that cannot be matched.
- When the distance $\|X_1 - X_0 W^*\|$ is large, the pre-treatment fit is bad: do not use synthetic control in that case.
- If the donor pool is too large, there can be over-fitting. Choose the donor pool judiciously.
- All unit weights must lie within $[0, 1]$.

7 Triple Differences (DDD)

7.1 Main Idea

Another method to extend DID is triple differences (DDD). Return to Card and Krueger (1994) and suppose we worry that Pennsylvania and New Jersey have genuinely different time trends in average wages, driven by some common shock that hits the two states differently. Then the simple two state DID is contaminated.

But notice what the minimum wage does and does not touch. A minimum wage increase affects the cashier at McDonald's, whose wage sits at the minimum, but it does not affect teachers, whose wages are far above it. If whatever differential state level shock we worry about moves cashiers and teachers alike, then the *gap* between cashier and teacher wages has the same trend in Pennsylvania and New Jersey absent the policy. The only genuinely treated group is cashiers in New Jersey after the policy.

7.2 The Three Way Fixed Effect Model

Denote i as the individual, g as the occupation group ($g = 1$ for cashiers), s as the state ($s = 1$ for New Jersey), and t as time ($t = 1$ for November, the post period). Extend the TWFE model to

$$Y_{it} = \rho D_{it} + \lambda_{g,t} + \delta_{s,t} + \alpha_i + \epsilon_{it}. \tag{14}$$

This is a **three way fixed effect (3WFE) model**. We control for individual level fixed effects α_i (which absorb group and state levels, since occupation and state are fixed within individual), occupation by time effects $\lambda_{g,t}$, and state by time effects $\delta_{s,t}$. The treatment dummy switches on only for the triple intersection: $D_{it} = 1$ only if $t = 1$, $g = 1$, and $s = 1$.

The state by time effects $\delta_{s,t}$ are exactly what the two state DID could not control: they allow New Jersey and Pennsylvania to follow different aggregate trends. The occupation by time effects $\lambda_{g,t}$ allow cashiers and teachers nationwide to trend differently. What must be excluded is a shock specific to the triple cell (cashiers in New Jersey after the policy) other than the policy itself.

7.3 Identification

The extended parallel trend assumption is that the error is orthogonal to the treatment given the fixed effects, $E[\epsilon_{it} \mid \lambda_{g,t}, \delta_{s,t}, \alpha_i, D_{it}] = 0$. Under this assumption, the policy effect ρ is recovered by three nested differences. Write $E[Y_{ist} \mid s, g, t]$ for the cell means implied by (14).

First difference (same state, same occupation, across time; the individual effect α_i cancels):

$$E[Y \mid s{=}0, g{=}0, t{=}1] - E[Y \mid s{=}0, g{=}0, t{=}0] = \lambda_{0,1} - \lambda_{0,0} + \delta_{0,1} - \delta_{0,0}, \tag{15}$$

$$E[Y \mid s{=}0, g{=}1, t{=}1] - E[Y \mid s{=}0, g{=}1, t{=}0] = \lambda_{1,1} - \lambda_{1,0} + \delta_{0,1} - \delta_{0,0}, \tag{16}$$

$$E[Y \mid s{=}1, g{=}0, t{=}1] - E[Y \mid s{=}1, g{=}0, t{=}0] = \lambda_{0,1} - \lambda_{0,0} + \delta_{1,1} - \delta_{1,0}, \tag{17}$$

$$E[Y \mid s{=}1, g{=}1, t{=}1] - E[Y \mid s{=}1, g{=}1, t{=}0] = \lambda_{1,1} - \lambda_{1,0} + \delta_{1,1} - \delta_{1,0} + \rho. \tag{18}$$

Second difference (same state, across occupations; the state by time trend cancels):

$$(16) - (15) = \lambda_{1,1} - \lambda_{1,0} - (\lambda_{0,1} - \lambda_{0,0}), \tag{19}$$

$$(18) - (17) = \lambda_{1,1} - \lambda_{1,0} - (\lambda_{0,1} - \lambda_{0,0}) + \rho. \tag{20}$$

Each second difference is itself a DID: the cashier versus teacher gap change within a state. The occupation by time trend survives, but it is the *same* in both states.

Third difference (across states):

$$(20) - (19) = \rho. \tag{21}$$

The occupation trend cancels, and only the policy effect remains. DDD is easily understood as a difference between two DIDs: the cashier-minus-teacher DID in New Jersey minus the cashier-minus-teacher DID in Pennsylvania.

8 Conclusion

When we observe the same units across time, we have panel data, and the fixed effect toolkit lets us cancel out all time invariant confounders. The FE, dummy, and FD regressions all do this; FE and dummy regressions are numerically identical, FE and

FD coincide with two periods but differ beyond that, and the choice between FE and FD depends on the serial correlation structure of the errors.

The main assumption for DID is parallel trends. Traditionally we assess it in two steps: draw the picture, and run the event study. But a statistical pre-test that minimizes the type I error rate mechanically inflates the type II error rate, and Roth (2022) shows that in realistic settings the power of pre-trend tests is low, while conditioning on passing them distorts both variance and bias. Be careful using them: check the power, do sensitivity analysis, and above all validate the assumption with economic knowledge of the setting.

When it is hard to find any control group, you can create one with the synthetic control method: two sets of weights, one over characteristics (v) and one over donor units (w), combine the donor pool into an artificial control that mimics the treated unit's pre-treatment behavior. Alternatively, if there is a further layer of difference that gives you exogeneity, such as an occupation group untouched by the policy, triple differences absorbs state specific trends with state by time fixed effects and identifies the effect from the triple interaction.

Throughout, keep asking the identification source question: which variation in the data is identifying the parameter? The answer determines the assumption you need, the interpretation you may claim, and the defense you must give. In the next chapter we stay with panel data and confront the staggered adoption case, where treatment starts at different times for different units and the TWFE regression turns out to hide serious problems.

A Homework Problems

1. **FE, Dummy, and FD Estimators.**

   Consider the fixed effect model $Y_{it} = \alpha + \rho D_{it} + A_i'\gamma + \epsilon_{it}$ with $T = 2$ periods (no additional covariates).

   (a) Write down the within (FE) transformation of the model and the first difference (FD) transformation of the model.

   (b) Show that with $T = 2$ the two transformations are identical up to a constant scaling, so the FE and FD estimators of ρ coincide.

(c) Now suppose $T > 2$. Suppose the errors ϵ_{it} are i.i.d. across t. Show that the first differenced errors $\Delta\epsilon_{it}$ are serially correlated, and compute $\text{Corr}(\Delta\epsilon_{it}, \Delta\epsilon_{i,t-1})$.

(d) Suppose instead that ϵ_{it} follows a random walk, $\epsilon_{it} = \epsilon_{i,t-1} + u_{it}$ with u_{it} i.i.d. Which estimator, FE or FD, would you prefer now, and why?

2. **Triple Differences Identification.**

Consider the 3WFE model (14) with two states ($s \in \{0, 1\}$), two occupation groups ($g \in \{0, 1\}$), and two periods ($t \in \{0, 1\}$), where $D_{it} = 1$ only if $s = 1$, $g = 1$, $t = 1$.

(a) Derive the four within-cell time differences (15)–(18) from the model, explaining why the individual effect α_i cancels in each.

(b) Form the two second differences and explain, in words, what each one holds fixed and what remains.

(c) Show that the third difference recovers ρ, and state precisely which shocks the DDD design allows for that the simple two state DID does not.

(d) Give one example of a shock that would still bias the DDD estimate.

3. **Data Splitting in Synthetic Control.**

The cross-validation procedure of the synthetic control method divides the pre-treatment sample into two parts: a training period and a validation period. Explain the reason why we split the data into these two parts. No math!

Practice Example: A DID Simulation with Event Study

In this practice example we simulate a panel with a policy that satisfies parallel trends, estimate the policy effect by TWFE DID, and then run the event study regression to verify the flat pre-trend and trace the dynamic effect.

Data generating process. Consider $J = 40$ states observed over $T = 12$ years. States $1, \ldots, 20$ are treated: a policy takes effect in year 8 and stays on. The outcome is

$$Y_{it} = \alpha_i + \lambda_t + \delta_{\tau(t)} D_i \cdot \mathbf{1}(t \geq 8) + \epsilon_{it}, \qquad \epsilon_{it} \sim \mathcal{N}(0, 1),$$

where $\alpha_i \sim \mathcal{N}(0, 4)$ are state effects, λ_t is a common trend, D_i indicates the treated group, and the dynamic effects at event time $\tau = t - 8$ are

$$(\delta_0, \delta_1, \delta_2, \delta_3, \delta_4) = (2.0,\ 3.0,\ 3.5,\ 4.0,\ 4.0),$$

with $\delta_\tau = 0$ for $\tau < 0$ (parallel trends holds by construction). The static DID target is the average post-period effect, $\bar{\delta} = 3.3$.

Analysis steps. The do-file (`src/chapter 8/ch8_did.do`) performs:

1. Simulate the panel and save the dataset.
2. Descriptive step: compute and list group-by-year mean outcomes, the raw ingredients of the trends plot.
3. Static DID: regress Y on the treatment dummy with state and year fixed effects; compare $\hat{\rho}$ to $\bar{\delta} = 3.3$.
4. Manual double difference: compute the four cell means (treated/control, pre/-post) and verify that the double difference matches the regression estimate.
5. Event study: regress Y on event time dummies interacted with the treated indicator, omitting $\tau = -1$; verify the pre-period coefficients are near zero and the post-period coefficients track the true $(\delta_0, \ldots, \delta_4)$.
6. Cluster standard errors at the state level and compare with unclustered ones.
7. Key lessons summary.

Key lessons.

- Under parallel trends, TWFE DID recovers the (average) policy effect; the manual double difference and the regression coefficient agree exactly in the two-by-two collapse.
- The event study shows both halves of the design's credibility: flat, insignificant pre-period coefficients, and interpretable dynamic effects afterward.
- Clustering matters: within-state correlation of errors makes unclustered standard errors misleadingly small. Chapter 11 returns to this at length.

The do-file is located at `src/chapter 8/ch8_did.do` and the simulated dataset at `data/chapter 8/ch8_did.dta`.

References

Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature*, 59(2): 391–425.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science*, 59(2): 495–510.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Card, David and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review*, 84(4): 772–793.

Rambachan, Ashesh and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *Review of Economic Studies*, 90(5): 2555–2591.

Roth, Jonathan. 2022. "Pretest with Caution: Event-study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.