Causal Inference via Difference-in-Differences

Term Paper

Erik Covarrubias[1] (3395844), Pietro Ducco[2] (3463167), Margarita Kaginian[3] (3465404), Norman Metzinger[4] (3501090), and Nebi Simsek[5] (3395869)

Department of Economics

Rheinische Friedrich-Wilhelms-Universität Bonn

submitted to: Prof. Dr. Christoph Breunig, Dr. Dennis Schroers

Degree Programme: Master of Science in Economics (M.Sc.)

submission date: Bonn, February 10, 2023

Abstract

This paper explores the use of **Difference-in-Difference** estimators for causal inference. First, we intuitively introduce the topic with a replication exercise of Card & Krueger (1994). Further, we provide a thorough theoretical background on panel data econometrics and fixed-effects estimators. Using Monte Carlo simulations, we also provide data-driven demonstrations of the estimators' properties. Then, we formally introduce the DiD model and discuss identification in a theoretical and applied approach using Di Tella & Schargrodsky (2004). Finally, we compare the fixed-effects and DiD estimators in more complex settings and conclude against using the former as a generalisation of DiD estimation.

[1]Section 5, Conclusions, Appendix C
[2]Section 6, Appendix D
[3]Section 3
[4]Introduction, Section 2, Appendix A
[5]Section 4, Appendix B

Contents

List of Figures

List of Tables

1 Introduction

The **Difference-in-Difference** (DiD) estimation compares the difference in outcomes before and after an exogenous intervention in the treatment to a control group. In the presence of unobserved confounders affecting the outcome of interest, the DiD strategy permits the estimation of the causal effect of treatment on the outcome in quasi-experimental settings.

In this paper, we review the classical DiD. First, we present our replication of the (Card & Krueger, 1994) paper, emphasising how homogenous treatment effects can be interpreted as causal. For that, we review the vital assumptions of common pre-trends and conditional independence in this setting. We argue that these assumptions might be violated in this application and see how implementing clustered standard errors can tackle the in-class correlation problems of treatment groups. Second, we investigate the theoretical aspects of Panel Data – the underlying data structure used in DiD. In this Panel Data structure, we discuss the fixed effects estimators, which help to control for unobserved confounders of fixed entities. Namely, we review the one-way and **Two-Way Fixed Effects** (TWFE) estimators. Third, we simulate these fixed effects estimators under various scenarios through a *Monte Carlo Simulation*. Our results show that the TWFE estimator performs well if the characteristics have enough variation across time. Notably, the two-way transformation eliminates the correlated errors, as we argued theoretically. Fourth, we derive the 2×2 DiD design based on the aforementioned fixed-effects models. We see that the DiD estimation circumvents the causal inference problem in quasi-experimental settings by using fixed effects and how the method identifies the effect of a conditionally independent intervention on the outcome of interest. Further, we replicate Di Tella & Schargrodsky (2004) to discuss the theoretical aspects of DiD estimation in practice. Fifth, we extend the 2×2 DiD estimator to a more general multiperiod case and compare it to the TWFE regression. We derive theoretically and show in simulations that the DiD and the TWFE regression are substantially different in this case. Lastly, we conclude our findings.

2 Card & Krueger (1994): DiD Application

Card & Krueger (1994) investigate whether the minimum wage introduction in New Jersey has a causal effect on employment. In this quasi-experimental setting, we discuss how the authors reason the underlying assumptions of DiD to hold and why the treatment effects can be interpreted as causal and homogenous. Further, we analyse how adding clustered standard errors could help conduct statistical inference in this framework. For this exercise, we use the actual data pulled from David Card's website for our analysis (Card & Krueger, 2014).

2.1 Set-up

Card & Krueger (1994) compare the treatment state of New Jersey to the control state of Pennsylvania to see whether the minimum wage introduction changed the employment in New Jersey[1]. The minimum wage is only introduced in New Jersey; therefore, we refer to New Jersey as the treatment state and Pennsylvania as the control state. The data is collected in two periods: pre and post-minimum wage introduction. The setup allows us to compare both states before and after the treatment. The data used is the employment data of fast food workers summarised on the store level; 420 stores across both states are observed. Also, data on opening hours or chain affiliation is collected. The selection of fast-food stores is randomised. This leads to comparable groups of fast-food stores in both states. Finally, the employment data is summarised at the state level. Thus, we have data of two states in two periods, resulting in the classical 2×2 DiD design.

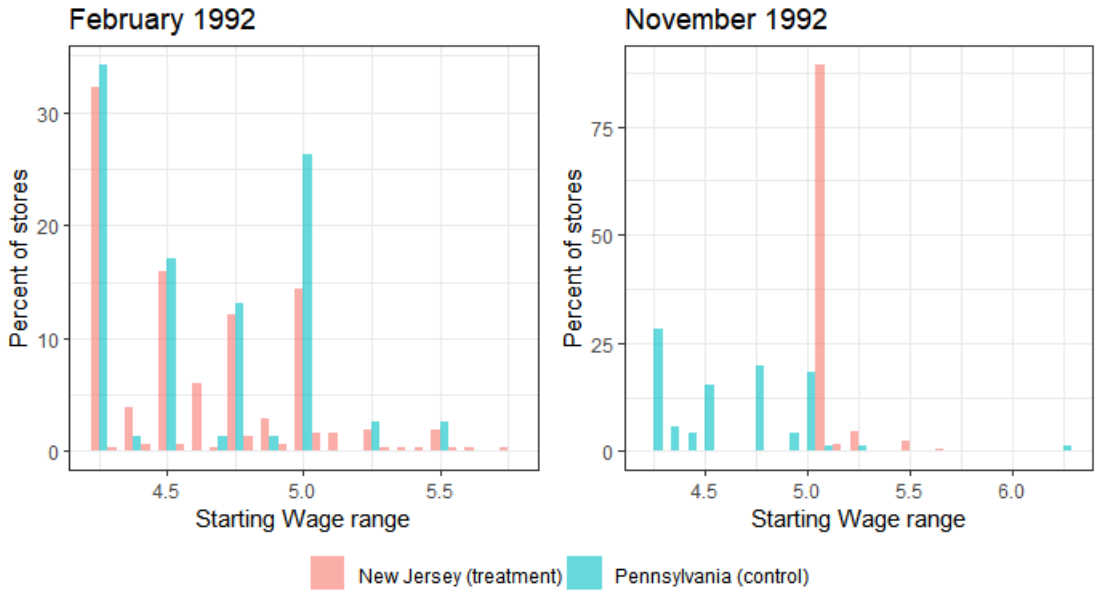
2.2 Assumptions

For our estimation to be unbiased and causal, we must control for confounding events and need our groups to be comparable. We need comparable groups because the control state of Pennsylvania functions as a counterfactual of New Jersey. This assumption is called the common pre-trends assumption. Therefore, Card & Krueger (1994) test for cross-state

[1]To be precise, the authors study New Jersey and eastern Pennsylvania, as they are directly neighbouring. This paper refers to these areas simply as New Jersey and Pennsylvania as it is shorter than the former.

differences, especially for starting wage, the fraction of full-time workers, and distribution of store types before treatment (see Appendix A.1). They find no significant results, so we can argue that the pre-trend assumption is not violated. Figure 1 shows how starting wages are similarly distributed across states before treatment and how the treatment affected the wage distribution. Another vital function is the conditional independence

Figure 1: Starting Wage Distribution Across Time and State



assumption. The idea is that employment is not correlated with any other unobservable event across the states. For example, Card & Krueger (1994) face the possibility that the announcement of the minimum wage leads to anticipations effects on restaurants within New Jersey. The management could fire people to cut costs before the treatment starts. Thus, the anticipation effect is correlated with the minimum wage introduction and the state of New Jersey. Card & Krueger (1994) consider this case and rule it out for their application. However, Angrist & Pischke (2009) argue that there is a confounding event. They observe that employment decreases before the introduction in Pennsylvania. That suggests a confounding event and that Pennsylvania is not a good counterfactual to New Jersey. For now, we consider the assumptions valid to interpret causal treatment effects. In Section 5 of this paper, we also review further assumptions next to these two key assumptions.

2.3 Application

We first replicate the simple specification of Hansen (2022). It allows us to display the causal effects investigated in Card & Krueger (1994). Let $D_{i,t}$ denote the treatment on the observed employment $FTE_{i,t}$[2]:

$$FTE_{i,t} = FTE(0)_{i,t} + (FTE(1)_{i,t} - FTE(0)_{i,t})D_{i,t}, \tag{1}$$

where $(FTE(1)_{i,t} - FTE(0)_{i,t})$ is the difference of employment between a treated state $FTE(1)_{i,t}$ and a perfect counterfactual that is untreated $FTE(0)_{i,t}$. This difference is the causal effect of the minimum wage if it is unequal to 0. Let us say there is a causal effect of minimum wage on employment; then we can write:

$$FTE(1)_{i,t} = FTE(0)_{i,t} + \theta, \tag{2}$$

where $\theta = (FTE(1)_{i,t} - FTE(0)_{i,t})$. θ describes the pure causal effect of the treatment. Further, we can describe the untreated potential employment with:

$$FTE(0)_{i,t} = \beta_0 + \beta_1 State_i + \beta_2 Time_t + \varepsilon_{i,t} \tag{3}$$

When we put eq. (2) and eq. (3) into eq. (1), we receive the specification of Hansen (2022):

$$FTE_{i,t} = \beta_0 + \beta_1 State_i + \beta_2 Time_t + \theta D_{i,t} + \varepsilon_{i,t} \tag{4}$$

State is a dummy variable with $State_i = 1$ indicating New Jersey and $State_i = 0$ defining Pennsylvania. $Time_t$ is also a dummy variable. When $Time_t = 0$, it is before the treatment, and $Time_t = 1$ is after the minimum wage is introduced. $D_{i,t}$ is an interaction term that can be rewritten as:

$$D_{i,t} = (State_i Time_t) \tag{5}$$

[2]To be precise in Card & Krueger (1994) $FTE_{i,t}$ is the full-time-equivalent of employment, measuring the number of employed workers per fast-food store at time t and state i. FTE is a unified measure of the employment status of part-time and full-time workers and management.

Table 1: Average Employment by State

	Pre	Post	Difference
Pennsylvania	23.33	21.17	2.17
New Jersey	20.44	21.03	0.59
Difference	2.89	-0.14	2.75
	(1st Difference)	(2nd Difference)	(DiD)

In this form, we can see that $D_{i,t}$ is the treatment dummy that measures the effect of the policy intervention in state i on a fixed time. The advantage of model 4 is that we can easily see and interpret the results displayed in Table 1. Suppose we want to know the average FTE of Pennsylvania in the pre-treatment period. We just set: $State_0, Time_0, D_{0,0} = 0$, and obtain $FTE_{i,t} = \beta_0$. To get the value of β_0, we have to calculate the average employment under eq. (4). We follow this approach and calculate the values for Table 1 where $\beta_0 = 23.33$. When taking the differences of the means from each period, we receive the first and second differences. Taking the difference between the two gives us the DiD estimand θ. Alternatively, we can get the first and second differences by taking the difference by state (horizontal in Table 1) and receive precisely the same θ. θ is the causal effect of the minimum wage on employment if the assumptions are valid. In other words, it is the change in New Jersey's employment relative to the change in Pennsylvania (Hansen, 2022). Appendix A.2 gives a visual representation of the causal effect.

A problem of model 4 is that it is harder to extend (Wing et al., 2018), e.g. add further controls. Therefore it would be easier to parametrise the model into TWFE. This is only possible for the 2×2 DiD design as Imai & Kim (2021) state. We discuss extensions of the 2×2 design in Section 5 and 6. Let us consider again eq. (1) but we define now equation eq. (3) as TWFE:

$$FTE(0)_{i,t} = v_t + u_i + X'_{i,t}\beta + \varepsilon_{i,t}. \tag{6}$$

v_t are time-varying but state-invariant characteristics and u_i are time-invariant but state-varying factors. Including them allows controlling for each type of unobserved characteristic. $X'_{i,t}$ is a set of control variables. Note that confounding effects varying at the

Table 2: Replicated and Original Regression Results

	Replication		Original Model	
	(i)	(ii)	(iii)	(iiii)
(Intercept)	$-1.88^{\cdot}$	-1.45		
	(1.07)	(1.21)		
New Jersey dummy	$2.28^{\cdot}$	$2.28^{\cdot}$	2.33^{*}	$2.30^{\cdot}$
	(1.19)	(1.20)	(1.19)	(1.20)
R^2	0.01	0.02		
Adj. R^2	0.01	0.01		
Num. obs.	351	351	357	357
Controls for chain and ownership	no	yes	no	yes

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: The original paper misses to report R^2 and the intercept. The standard errors are shown in brackets.

time-state level cannot be controlled for by fixed effects.[3] This is because the treatment $D_{i,t}$ is varying at the time-state level; any fixed effect at that level would capture all effects leading to no observable variation. If we include eq. (6) and eq. (2) into eq. (1), we receive:

$$FTE_{i,t} = v_t + u_i + X'_{i,t}\beta + \theta NJ_{i,t} + \varepsilon_{i,t} \tag{7}$$

eq. (7) is the same model that Card & Krueger (1994) uses and that we use in our replication. A more thorough analysis of this transformation is stated in Section 5. Our results can be seen in Table 2, where the estimands can be interpreted as the employment effect whether or not there is a treatment, i.e. being New Jersey or not. Thus, we can write $D_{i,t} = NJ_{i,t}$. Noteworthy are the different values for the estimates of our replication and the one of Card & Krueger (1994). We follow Card & Krueger (1994) and drop all observations with data missing in employment and starting wages in both periods. There is still a discrepancy in observations, though the standard errors are correctly replicated, and the model is slightly insignificant with a p-value just above 0.05.[4]

[3] An example in this context would be yearly wage negotiations of state-wide unions.

[4] We assume that Card & Krueger (1994) handle missing values differently as mentioned in their paper. However, the exact handling is not explained transparently.

Table 3: Average Employment by Region

	South NJ	Central NJ	North NJ	PA 1	PA 2
Pre Increase	16.74	22.56	22.17	25.65	22.27
Post Increase	17.59	21.77	22.74	21.73	21.89
Difference	0.85	-0.79	0.57	-3.92	-0.38

2.4 Homogeneous Treatment Effects

Card & Krueger (1994) implicitly assume homogenous treatment effects in their work. This means that θ affects all individuals in the same way. To be more precise, the minimum wage should affect every level of the unit below the state level the same. In Card & Krueger (1994) the data is first aggregated on the restaurant level, then on the regional level, and then on the state level we mainly analyse. In Table 3, we see the difference in employment in different regions within Pennsylvania and New Jersey. The question is whether those regions are differently affected by the treatment, essentially whether there are heterogeneous treatment effects. When Pennsylvania is heterogeneously affected, it is a bad control for New Jersey. If the treatment state New Jersey is heterogenous, the model of Card & Krueger (1994) is wrongly specified and an alternatively model like in de Chaisemartin & D'Haultfoeuille (2022) is advised. Hansen (2022) tests for the difference of treatment effects between states and find no significant difference; thus, we can assume the model of Card & Krueger (1994) is correctly specified. In Section 5, we discuss in detail the regression exclusion test to identify heterogenous treatment effects.

2.5 Inference using Clustered Standard Erros

Card & Krueger (1994) do not investigate whether employment is correlated within the state structure, i.e. between different regions or restaurants within a state, resulting in possibly misspecified standard errors as some state-specific influence, e.g. state legislation, could affect them. We can control with fixed effects for these time-invariant shocks in the estimation. Nevertheless, more is needed to fix the problem that individual fast food stores could be correlated with each other within the state. Thus, leading to correlated error terms. Clusters on the state level could solve this problem, especially because the variable of interest $NJ_{i,t}$ varies at the state level; it is recommended (see Angrist

& Pischke, 2009; Abadie et al., 2022). Bertrand et al. (2004) argue that DiD by design is prone to under-estimate the standard error of θ and therefore overreject estimations. Thus, clustering is advised.

First, we formalise the problem of correlated error terms following Angrist & Pischke (2009):

$$E[\varepsilon_{l,i}\varepsilon_{j,i}] = \rho\sigma_e^2 > 0 \tag{8}$$

Store l and j are in state i and have an intra-class correlation coefficient of ρ and a homoscedastic residual variance of σ_e^2.[5] As we are interested in how much the variance of our treatment effect $V(\theta)$ is over-estimated, we can formulate the Moulton factor (Moulton, 1986) for the case of different group sizes:

$$\frac{V_{cluster}(\theta)}{V(\theta)} = 1 + \left[\frac{V(n_i)}{\bar{n}} + \bar{n} - 1\right]\rho_d\rho \tag{9}$$

$\bar{n}$ is the average amount of restaurants. $V(n_i)$ is the variance of restaurants in each state. ρ_d is the intraclass correlation of our treatment variable $D_{i,t}$. If ρ_d increases, the difference to the regular standard error increases. If $\rho_d = 0$ no clustering is necessary, and the treatment within the state is truly randomly allocated. One can see that the intra-class correlation and the variance of each cluster are increasing the difference between $V_{cluster}(\theta)$ to $V(\theta)$.

For our application, we apply clusters to the three available unit levels in Card & Krueger (1994): state, region, and restaurant and compare the variance-bias tradeoff of these options. In Table 4, we can see that in columns (ii) and (iii), the error term has increased compared to no clustering. Implying $\rho_d \neq 0$, and thus there is some intra-class correlation. In columns (iii) and (iv), we see cluster sizes well below the often recommended 50 clusters (Bertrand et al., 2004; Angrist & Pischke, 2009). Especially the corner case of column (iv) provides highly significant results with a standard error close to 0. That is because the standard error is not calculated based on the number of observations but that of clusters. Consider eq. (9) where the outcome is calculated with two observations (and just one treated cluster). This way, the standard errors are algebraically becoming zero, and the treatment effect is highly significant while incorrectly

[5] Card & Krueger (1994) assume in their estimation homoscedastic error terms.

Table 4: Comparison of Different Levels of Clusters

Clusters	None	Store	Regions	State
	(i)	(ii)	(iii)	(iv)
(Intercept)	$-1.88^{\cdot}$	-1.88	-1.88	-1.88^{***}
	(1.07)	(1.37)	(1.37)	(0.00)
New Jersey Dummy	$2.28^{\cdot}$	2.28	2.28	2.28^{***}
	(1.19)	(1.45)	(1.40)	(0.00)
R^2	0.01	0.01	0.01	0.01
Adj. R^2	0.01	0.01	0.01	0.01
Num. obs.	351	351	351	351
RMSE		8.71	8.71	8.71
N Clusters		351	5	2

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: We use the naive Stata-clustered standard errors imitated by *R*.

estimated (Hansen, 2022).

We can see that the clustered standard error deviates massively from the original reported standard errors. Card & Krueger (1994) two state and time periods case is too small for controlling class dependencies within states (Hansen, 2022). An extension of the analysis with more control and treatment states, e.g. including all US States (see Card, 1992), would be an ideal setting for controlling for these class dependencies. Clustering is a helpful extension for causal inference following Abadie et al. (2022), especially where fixed effects cannot help and groups are not randomly assigned.

3 Panel Data

3.1 Theoretical Background

The Panel Data estimators are commonly used when data contains time series observations of some individuals. The unique features of such a data structure usually are the higher number of degrees of freedom and less multicollinearity than cross-sectional data, the availability to capture complex individual behaviour, and more accurate inference of

model parameters Hsiao (2007). When the observations are available for each individual and all time periods, the panel data is balanced; if there are different time periods for individuals, the panel data is unbalanced. In this paper, we will focus on balanced panels.

Panel data also can be useful when there is time-invariant endogeneity in the data. Unlike the cross-section data, using panel data, the consistent estimator can be achieved with the Fixed Effect (FE) Model without using instrumental variables. With Panel Data, it is also possible to model the dynamic relationships and effects and allow more heterogeneity forms (Hansen (2022)). One of the first articles applying panel data in econometrics are Mundlak (1961) and Balestra & Nerlove (1966). As of today, there are many articles and textbooks related to this topic. Especially lots of papers refer to computational methods, which require meticulous attention to detail. Whereas there are many interesting features related to Panel Data, we will deep dive into one of them - the practical application of the panel data model based on the Frisch-Waugh-Lovell Theorem. The theorem suggests an algorithm that can speed the calculation of the Ordinary Least Squares (OLS) estimator based on the original model form. In particular, it claims that the OLS estimator can be based on the OLS estimation of the demeaned model form (or what is called <<within transformation>> of the model).

In economics, many empirical analyses are interested in the causal, structural, or treatment effect of some variable on a variable of interest. One of the most popular examples here is the government policy effect. However, it is commonly known that economic policies are not randomly assigned, which forces economists to use some quasi-experimental techniques based on the observed data. The central assumption used for such analysis is the randomness of the variable of interest after controlling for some different factors. Another common approach is to use the instrumental variable methods in cases where treatment is not randomly assigned, but there is some external variable, such as a government program or service, that is either randomly assigned or the researcher is willing to take as exogenous conditional on the right set of control variables (or simply controls).

The most difficult part of such an analysis is to decide which control variables to include. This problem occurs both when the treatment/instrument is assigned randomly and not. In the first case, a researcher will be interested in including additional controls

to help absorb residual variation. In the second case, a researcher should choose what needs to be conditioned on to make the argument that the instrument or treatment is exogenous plausible. Here is where the Fixed Effects Model of Panel Data can be useful. Such models can control for both observed and stable unobserved confounders, which lends greater credibility to the assumption necessary to estimate unbiased causal effects. The unobserved confounders that FE models can account for are those characteristics that remain fixed over time within observed units (or over the units within observed time) of analysis. It is well known that estimation by the FE model is equal to the differenced estimator when there are only two periods in the data, however when the time-variant error is i.i.d. the FE estimator is more efficient (Gauss-Markov theorem, Hansen (2022)). In this section, we will construct the Fixed Effect estimator and discuss the properties needed to get a consistent estimator. Further in the paper, we will show the key changes for the DiD estimator and provide an application to see if the theoretical conclusion remains in practice.

Going forward, we will apply the following notation: Y_{it} as the target variable, where i are the observations on the individual and t refers to the time period, β is a $k \times 1$ coefficient vector, e_{it}. Furthermore, it is common to use (one-way) the error-components structure for regression error: $e_{it} = u_i + \varepsilon_{it}$, where u_i is the individual-specific effect (unobserved missing variable) and ε_{it} are idiosyncratic (i.i.d.) errors. Taking these notations, we can construct the one-way error component regression model:

$$Y_{it} = X_{it}^{'}\beta + u_i + \varepsilon_{it}$$

In econometrics literature u_i is called fixed effect if u_i is unknown and there is a correlation between $X_{it}^{'}$ and u_i. Due to this correlation, the pooled and random effects estimators are biased. Using cross-sectional data will not help us to get a consistent estimator for β. However, we still can get it by transforming the model with clever use of the panel data. Moreover, in this case, we will require the strict exogeneity for an error ε_{it}, meaning

Assumption 1

$$E(\varepsilon_{it}|X_{it}) = 0$$

3.2 One-way Fixed Effects Estimator

The one-way fixed effects (OWFE) regression model looks the same as the one-way component regression model where the source of endogeneity is u_i, which we want to eliminate:

$$Y_{it} = X_{it}^{'}\beta + u_i + \varepsilon_{it}$$

As u_i does not vary across time, we can use the difference between periods to eliminate it. Instead of using the set of equations for each time period (t) and each individual (i), we can use demeaned variables in the model, such as :

$$\dot{Y}_{it} = Y_{it} - \bar{Y}_i,$$

where $\bar{Y}_i = \frac{1}{T_i}\sum_t Y_{it}$. Same logic is used to get $\dot{X}_{it}$ and $\dot{\varepsilon}_{it}$. Then using the following two equations:

$$Y_{it} = X_{it}^{'}\beta + u_i + \varepsilon_{it}$$

$$\bar{Y}_i = \bar{X}^{'}{}_i\beta + u_i + \bar{\varepsilon}_i$$

We can take the difference between the two equations to get the following model:

$$\dot{Y}_{it} = \dot{X}^{'}{}_{it}\beta + \dot{\varepsilon}_{it}$$

The strict exogeneity **Assumption 1** implies: $E(\dot{\varepsilon}_{it}|\dot{X}_{it}) = 0$, so we can confirm that the OLS estimator for the demeaned equation is unbiased. To get the fixed effect estimator, we require the full rank assumption.

Assumption 2

The matrix $E(\sum_t^T \dot{x}_{it}\dot{x}_{it}^{'})$ should have a full rank.

The idea behind assumption 2 is to require enough variance over time between the model variables. As an example, in the simpler linear models, the estimated coefficient was always a scaled covariance, where the scaling was by variance term. Thus regressors must vary over time for at least some i and not be collinear. Without Assumption 2, the

estimation of the coefficients would be impossible. Based on the **Assumptions 1 and 2** above, the OLS estimator, known as the OWFE estimator, is

$$\hat{\beta}_{OWFE} = \left(\sum_{i}^{n} \sum_{t}^{T} \dot{X}_{it}^{'} \dot{X}_{it} \right)^{-1} \sum_{i}^{n} \sum_{t}^{T} \dot{X}_{it} \dot{Y}_{it}$$

The estimator is called a OWFE estimator because it can also be written in the form where we minimize the sum of squared residuals and treat u_i as parameters to be estimated:

$$(\hat{\beta}_{OWFE}, \hat{u}_1, ..., \hat{u}_n) = \arg \min_{b, u_1, ..., u_n} \sum_{i}^{n} \sum_{t}^{T} (y_{it} - x_{it}^{'} b - u_i)^2$$

Parameters u_i are called individual (one of the possible) fixed effects. Using Central Limit Theorem (CLT) for i.i.d. random vectors, we can get the variance for such an estimator:

$$\hat{V}_{\hat{\beta}_{OWFE}} = \left(\frac{1}{n} \sum_{i=1}^{n} \dot{X}_i^{'} \dot{X}_i \right)^{-1} \hat{\omega}_T \left(\frac{1}{n} \sum_{i}^{n} \dot{X}_i^{'} \dot{X}_i \right)^{-1},$$

where $\hat{\omega}_T = \frac{1}{n} \sum_i^n (\sum_t^T (X_{it} - \bar{X}_i) \hat{e}_{it}) \left(\sum_t^T (X_{it} - \bar{X}_i \right) \hat{e}_{it})'$.

It can also be shown that we could get the previous equation by applying GLS to the differenced estimator when $T = 2$. Using the fact that $V_{gls} \leq V_{pooled}$ (V_{pooled} is the variance based on OLS estimation) we will get for the differenced estimator, we can confirm that the fixed effects estimator is more efficient. We stop short of this discussion as it lies beyond the scope of this paper. For more insights, see Hansen (2022).

3.3 Two-way Fixed Effects estimator

Usually, more than individual fixed effects are needed to model the general population behaviour as it also can be that individuals experience some shocks from period to period. In this set up, we would require additional unobserved time-fixed effects in the model:

$$Y_{it} = X_{it}^{'} \beta + u_i + v_t + \varepsilon_{it}$$

The equation above represents the two-way error component model, where ε_{it} is an idiosyncratic error. To perform an estimation for coefficient β we would need to provide a within transformation for the model. The two-way within transformation subtracts both individual-specific means and time-specific means to eliminate both v_t and u_i from the two-way model. To do so, we would need the time-specific mean for each time t:

$$\tilde{Y}_t = \frac{1}{N_t} \sum_{i}^{n} Y_{it}$$

In the case of the balanced panels, the two-way within transformation is:

$$\ddot{Y}_{it} = Y_{it} - \bar{Y}_i - \tilde{Y}_t + \bar{Y}$$

where $\bar{Y} = \frac{1}{n} \sum_i \sum_t Y_{it}$ is the full sample mean. Hence, the error term in the transformed model will be:

$$\ddot{Y}_{it} = v_t + u_i + \varepsilon_{it} - (\bar{v} + u_i + \bar{\varepsilon}_i) - (v_t + \bar{u} + \tilde{\varepsilon}_t) + \bar{v} + \bar{u} + \bar{\varepsilon} = \varepsilon_{it} - \bar{\varepsilon}_i - \tilde{\varepsilon}_t + \bar{\varepsilon} = \ddot{\varepsilon}_{it}$$

The within-transformed model is stated below:

$$\ddot{Y}_{it} = \ddot{X}'_{it}\beta + \ddot{\varepsilon}_{it},$$

The estimator can be constructed applying the least squares under assumptions stated previously. **Assumption 1** implies the strict exogeneity: $E(\ddot{\varepsilon}_{it}|\ddot{X}_{it}) = 0$, **Assumption 2** implies the full rank for matrix $E(\sum_t^T \ddot{X}_{it}\ddot{X}'_{it})$. Once again, assumption 2 requires variety over time for within transformed features.

Using OLS, the estimated coefficient is equal to the following:

$$\hat{\beta}_{TWFE} = \left(\sum_{i}^{n} \sum_{t}^{T} \ddot{X}'_{it}\ddot{X}_{it} \right)^{-1} \sum_{i}^{n} \sum_{t}^{T} \ddot{X}_{it}\ddot{Y}_{it}$$

In the unbalanced case, the estimation method can be constructed based on including dummy variables for all time periods. We can define the dummy variable τ_t, where the t^{th} element is equal to 1, and the rest are equal to 0. For each individual i there can be

a different number of T periods (for one, we can observe only two periods, for another one, all T); that is why we can specify the time-fixed effect variable by multiplying the dummy variable with the time fixed effects as follows: $v_t = \tau_t' v$, where $v = (v_1, ..., v_T)'$. In this case, the individual with two time periods will have only two non-zero time-fixed effects, and the one with all T periods will have a whole vector v_t as time-fixed effects. Using the time dummy variables, the two-way model will look as follows:

$$Y_{it} = X_{it}'\beta + \tau_t' v + u_i + \varepsilon_{it}$$

This model can be estimated by OWFE with regressors X_{it} and τ_t and coefficient vectors β and v under **Assumptions 1 and 2**. It is essential to exclude one time dummy variable from τ_t to provide identification. This is one of the most popular methods for unbalanced panels.

In the following sections, we will provide the Monte-Carlo simulations for the fixed effects models and discuss the difference-in-difference estimator and its connection with panel data models.

4 Monte Carlo Simulations

In the previous section, we introduce **One-Way Fixed Effects** (OWFE) and TWFE estimators. These estimators are especially needed when the fixed errors are correlated with the covariates. Correlation with the error terms results in bias that may lead to wrong interpretations. Luckily, it is possible to eliminate the bias term with the transformations mentioned in the previous section if the model satisfies certain assumptions. In this section, we conducted some Monte Carlo Simulations to see how OWFE and TWFE estimators react to certain structural changes in the model.

4.1 Data Generating Process

For simulations, we used a multivariate case, and the dependent variable is explained by five different characteristics. For each unit at a specific time, the covariates are structured in three parts: Unit-specific randomly generated variables, time-specific randomly

generated variables and a variation over time.

$$X_{it}^{unit} \sim N(0, \Sigma_X); \quad X_{it}^{time} \sim N(0, \Sigma_X); \quad X_{it}^{var} \sim N(0, \Sigma_X)$$

For all three structures, we used the same randomly generated variance-covariance matrix for simplicity. N represents the number of units, and T represents the number of periods. X_{it}^{unit} is the same value for the unit i at each time $t = 1 \dots T$ while X_{it}^{time} is the same value for all the units $i = 1 \dots N$ at time t. Using the third structure X_{it}^{var}, we generated the variation for each unit across time. We sum up these three terms and obtain the covariates for unit i at time t.

$$X_{it} = X_{it}^{unit} + X_{it}^{time} + c_{var} X_{it}^{var}$$

Here it is easy to see that we can control for the variation across time with the constant c_{var}. Setting this value to 0, we will get unidentified TWFE estimates. Depending on the number of characteristics, we may also get an unidentified OWFE results since we may not satisfy the full rank assumption. Following, we generate the error terms and trend variables.

$$\epsilon_{it} \sim N(0, 1); \quad w_i \sim N(2, 0.5)$$

$$u_i = c_{unit} X_{(2)}^{unit} + \nu_i; \quad \nu_i \sim U(0, 1); \quad v_t = c_{time} X_{(2)}^{time} + \xi_t; \quad \xi_t \sim U(0, 1)$$

We generate time-invariant and unit-invariant errors correlated with the second dimension of X since we will be interested in estimates for the second parameter. The c_{unit} and c_{time} constants determine the correlation of error terms with this second dimension of X covariates. Assigning 0 to these values will eliminate the correlation. On the other hand, we can simply increase these constants to obtain higher correlations. We introduced the time trend variable for each unit with w_i. The parameters for w_i are set arbitrarily to obtain positive trend variables. Finally, we set $\beta = [1, 5, 3, 3, 3]$ and obtain the Y_{it}.

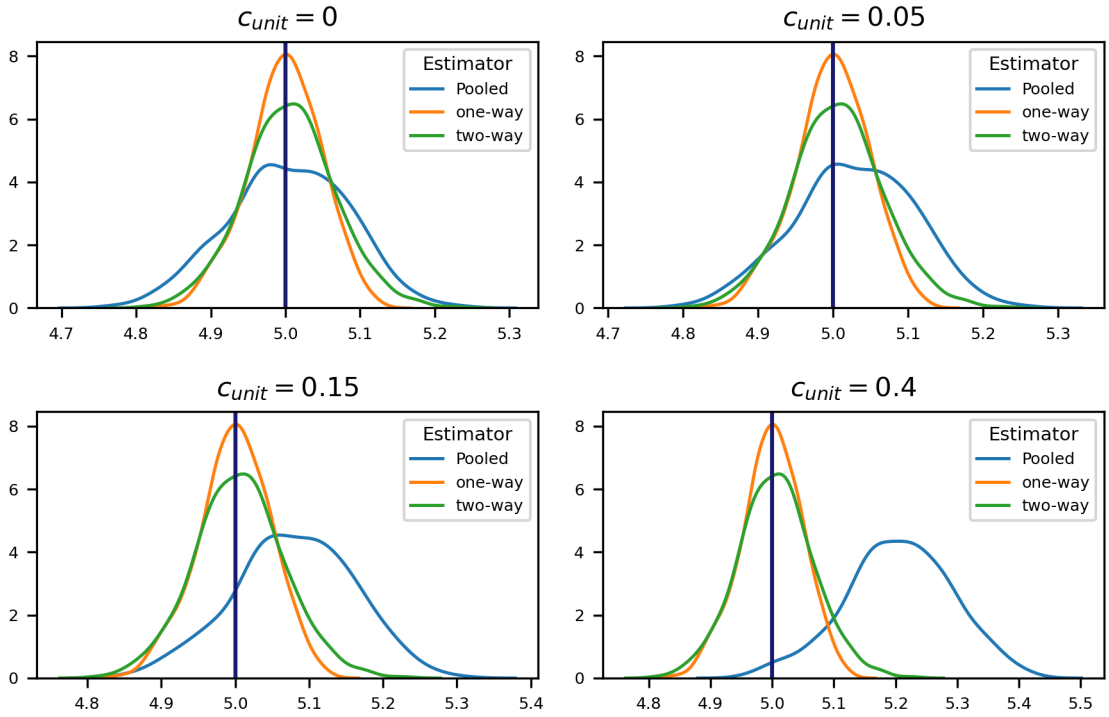
$$Y_{it} = [1, 5, 3, 3, 3] X_{it} + v_t + c_{trend} t w_i + u_i + \epsilon_{it}$$

Here t in front of w_i represents the periods, and with another constant c_{trend}, we will

control the intensity of the time trend. If we change this value to 0, we can eliminate the trended variables. In all our simulations, we set $N = 100$, $T = 20$ and simulate the results 500 times. In the next part, we will control the values of constants we introduced during DGP and show how OWFE and TWFE estimators react to these changes.

4.2 Results

Figure 2: Density Graph for β_2 Under Different Time Invariant Error Correlations



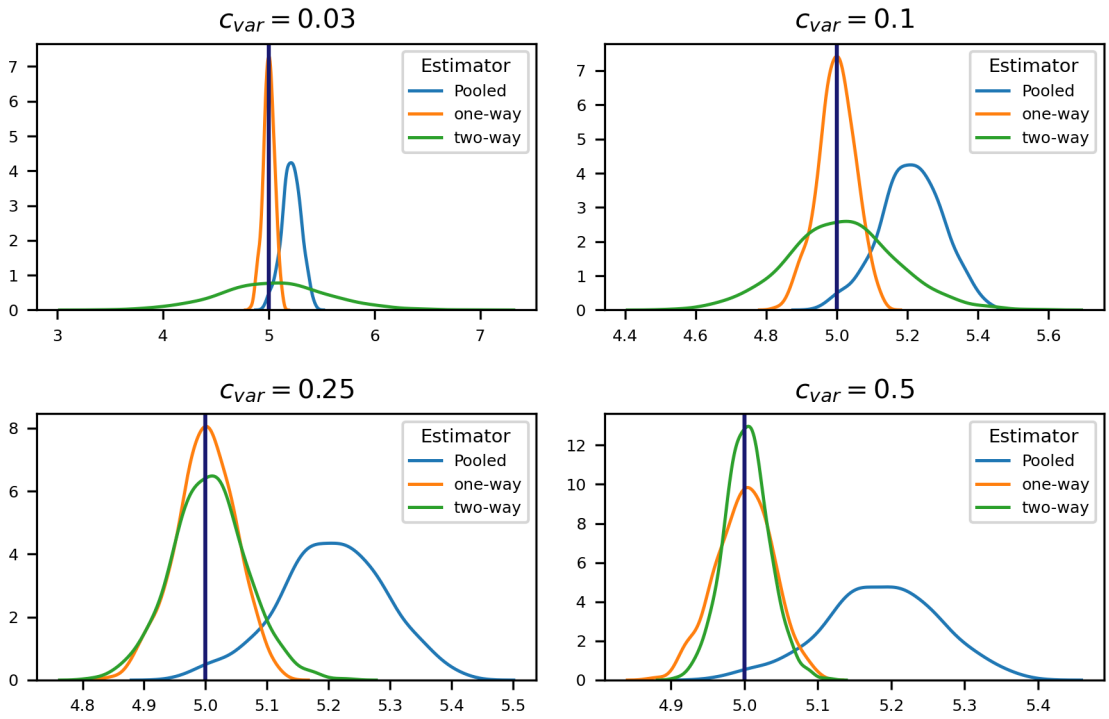
The figure plots the density graph of β_2 obtained from different estimators. We have controlled the performance of estimators for different values of c_{unit}. We set $c_{time} = 0$, $c_{trend} = 0$ and $c_{var} = 0.25$. We set $T = 20$, $N = 100$ and simulate the results for 500 times. The resulting density graphs visualize the performance of three different estimators: Pooled OLS estimator, OWFE estimator and TWFE estimator. RMSE values for estimators can be found in Appendix B Table 5(a).

It is important to understand the bias structure and justify not using Pooled OLS estimator without transformation when we have correlated error terms. To see that, we first look at the correlated time-invariant and set $c_{unit} > 0$ while $c_{time} = 0$. In particular, we will use $c_{unit} \in \{0, 0.05, 0.15, 0.4\}$. The resulting density graphs of estimates are shown in Figure 2.

Here we can see that as soon as we introduce the correlation with a positive c_{unit}, the

pooled OLS estimates become unreliable. On the other hand, the performance of one-way and TWFE estimators looks similar for all c_{unit} values. Since we don't lose much with a TWFE estimator, it looks like a safe choice for many analyses. However, when the variation over time is low, we may want to use OWFE instead of TWFE. To see that, we can control the variation of X covariates over time with c_{var} and observe the performance of the estimators. We will use $c_{var} \in \{0.03, 0.1, 0.25, 0.5\}$.

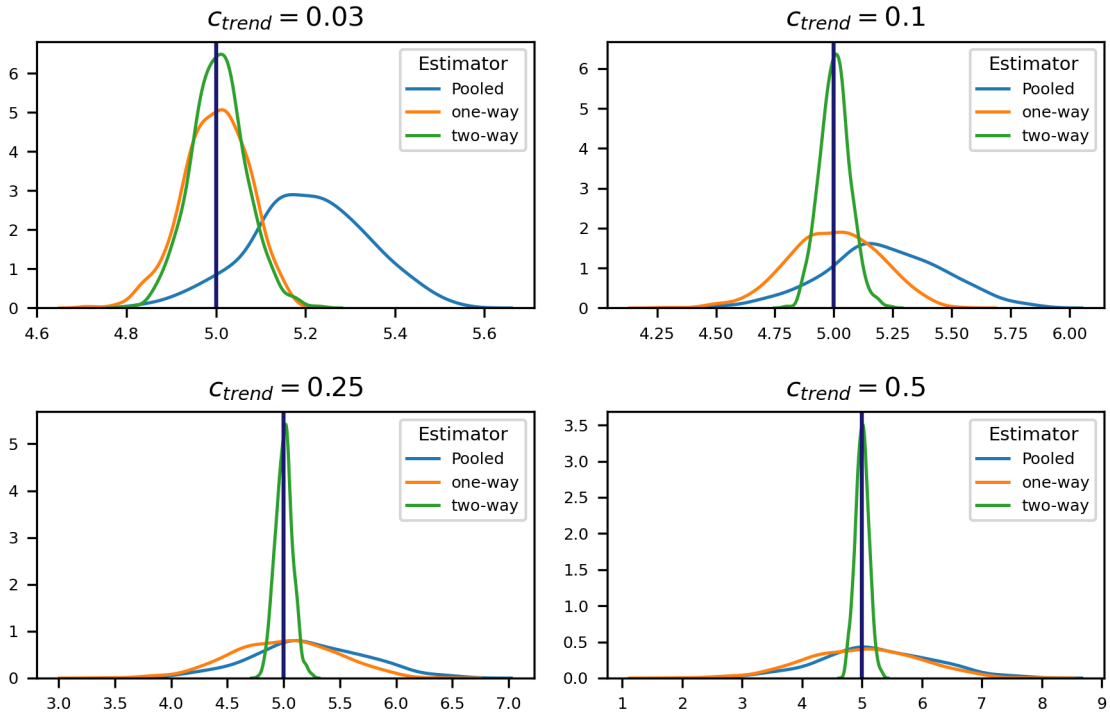
Figure 3: Density Graph for β_2 Under Low Variance Across Time



The figure plots the density graph of β_2 obtained from different estimators. We have controlled the performance of estimators for different values of c_{var}. We set $c_{time} = 0$, $c_{trend} = 0$ and $c_{unit} = 0.4$. We set $T = 20$, $N = 100$ and simulate the results for 500 times. Graphs visualize the performance of three different estimators: Pooled OLS, OWFE estimator and TWFE estimator. RMSE values for estimators can be found in Appendix B Table 5(b).

The OWFE estimator is a much better estimator when the variation across time is low since we eliminate the variation of covariates further with the two-way within transformation. However, after introducing sufficient variation, we observe that the TWFE estimator becomes even more efficient than the OWFE estimator since the decrease in the variation of residual terms dominates the decrease in the variation of X. Moreover, as we will see in Figure 4, introducing the trended variable on the right-hand side of the model will cause further efficiency loss for the OWFE estimator. Here we used

Figure 4: Density Graph for β_2 With Trended Variables

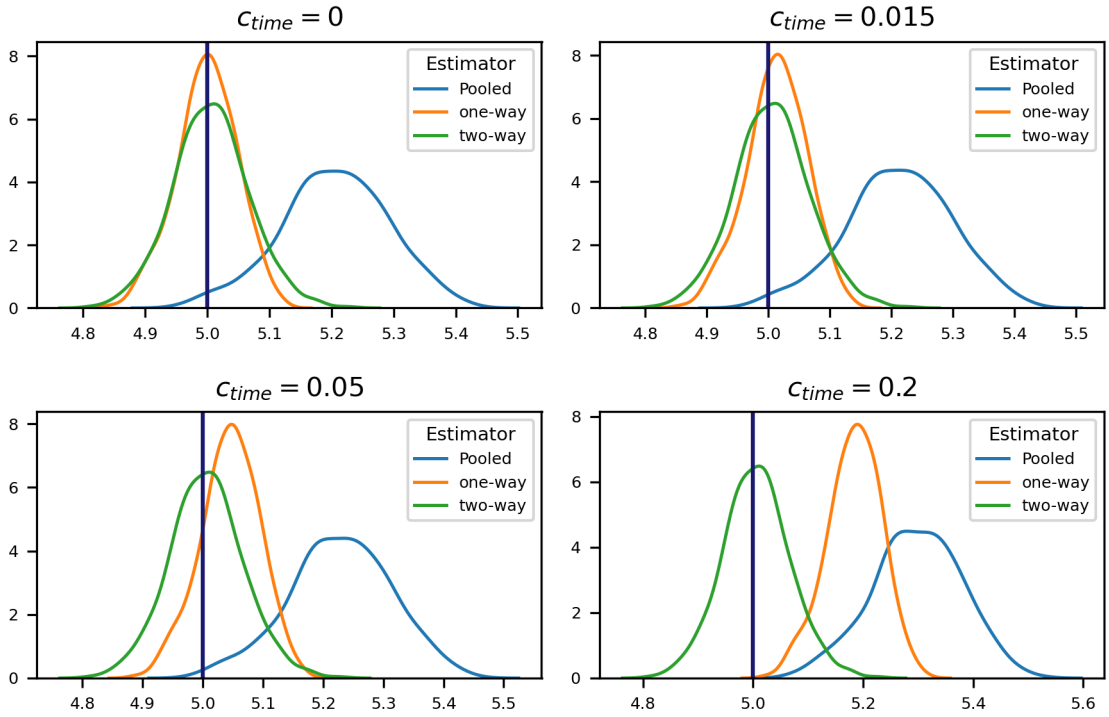


The figure plots the density graph of β_2 obtained from different estimators. We have controlled the performance of estimators for different values of c_{trend}. We set $c_{time} = 0$, $c_{var} = 0.25$ and $c_{unit} = 0.4$. We set $T = 20$, $N = 100$ and simulate the results for 500 times. Graphs visualize the performance of three different estimators: Pooled OLS, OWFE estimator and TWFE estimator. RMSE values for estimators can be found in Appendix B Table 5(c).

$c_{trend} \in \{0.03, 0.1, 0.25, 0.5\}$. As we increased the intensity of the trend, we saw that all estimators lost efficiency while TWFE estimator is performing much better than the OWFE estimator. Moreover, as expected, introducing a correlated unit invariant error term will cause bias for the one-way fixed effect estimator. To see that we will use $c_{time} \in \{0, 0.015, 0.05, 0.2\}$ and summarize the results in Figure 5.

In addition to our discussion about efficiency, the TWFE estimator eliminates the bias when there are correlated unit-invariant error terms. With that being said, of course, the dynamics of a real-world setting could be much more different and may not allow for a straightforward interpretation using the methods discussed here. We may need more assumptions or control for other potential bias possibilities like reverse causality or selection bias. In this section, we tried to show some properties of Pooled, OWFE and TWFE estimators and discuss some cases where we would like to use one over the other. Proper interpretation of estimated values is critical to reaching causal inferences, and in

Figure 5: Density Graph for β_2 Under Different Unit Invariant Correlations



The figure plots the density graph of β_2 obtained from different estimators. We have controlled the performance of estimators for different values of c_{time}. We set $c_{trend} = 0$, $c_{var} = 0.25$ and $c_{unit} = 0.4$. We set $T = 20$, $N = 100$ and simulate the results for 500 times. Graphs visualize the performance of three different estimators: Pooled OLS, one-way fixed effect estimator and TWFE estimator. RMSE values for estimators can be found in Appendix B Table 5(d).

our simulations, we hoped to give insights into such interpretations. Only using these methods with a proper discussion about the underlying model could lead us to sound causal interpretations.

In the next section, we will add a binary treatment variable to this model and introduce the generalized DiD model with a specific focus on the underlying assumptions. Later with an application, we will give an example of how the methods discussed in this section can be used to reach the so-called sound causal interpretations.

5 The Difference-in-Differences estimator

The one- and two-way fixed effects estimators reviewed in the previous section allow for unbiased estimation of the parameters of interest in the presence of unobserved confounders affecting the outcome of interest. Recall that our topic of interest remains causal

inference via statistical methods. However, the fixed effects estimators we studied before might not be suited for determining causality. For example, it turns out that fixed effects cannot solve the issue of reverse causality (Cunningham, 2021).

Under certain assumptions, the DiD strategy permits the estimation of the causal effect of treatment on the outcome using a variation of the fixed effect setup reviewed earlier (Bertrand et al., 2004; Angrist & Pischke, 2009; Cameron & Trivedi, 2005; Wing et al., 2018; Cunningham, 2021; Hansen, 2022).

Formally, let the binary intervention of interest be:

$$D_{it} = \begin{cases} 1 & \text{if individual } i \text{ recives treatment in period } t \\ 0 & \text{otherwise} \end{cases}$$

In the standard DiD estimation, focus is restricted to the 2×2 design. In this setting, there are two time periods, before and after the intervention, and two groups, control and treatment. Only for the second period does treatment occur for the latter group. (Cameron & Trivedi, 2005; Wing et al., 2018; Cunningham, 2021; Hansen, 2022). In this case, D_{it} can be expressed as an interaction term of a treatment group indicator and a period two indicator (Angrist & Pischke, 2009; Wing et al., 2018; Hansen, 2022).

We are interested in the casual effect of D_{it} on an outcome $Y = h(D, X, e)$, where X denotes a set of control variables and e represents a vector of unobserved confounders. Furthermore, let Y be described by a variation of the two-way error component model presented in the previous section as follows:

$$Y_{it} = \theta D_{it} + X'_{it}\beta + v_t + u_i + \varepsilon_{it} \tag{10}$$

where v_t is an unobserved time-specific effect, u_i is an unobserved individual-specific effect and ε_{it} is an idiosyncratic error. DiD estimation handles eq. (10) using a fixed-effects approach. The unobserved confounders are eliminated by applying a two-way within transformation to the model reviewed in the previous section (Cameron & Trivedi, 2005; Hansen, 2022):

$$\begin{aligned} \ddot{Y}_{it} &= Y_{it} - \tilde{Y}_t - \bar{Y}_i + \bar{Y} \\ &= \theta \ddot{D}_{it} + \ddot{X}'_{it}\beta + \ddot{\varepsilon}_{it} \end{aligned} \tag{11}$$

The DiD estimator is **Ordinary Least Squares** (OLS) applied to eq. (11) (Cunningham, 2021; Hansen, 2022; Hsiao, 2022). The most common way to implement the estimator is to transform the two-way model into its dummy variable representation discussed earlier in this document:

$$Y_{it} = \theta D_{it} + X'_{it}\beta + \tau'_t v + u_i + \varepsilon_{it} \tag{12}$$

Again, Equation (12) can then be estimated via a one-way fixed effects approach, with regressors D_{it}, X_{it} and τ_t and coefficient vectors θ, β and v (Hansen, 2022). Indeed, regressions of $\ddot{Y}_{it}$ on $\ddot{D}_{it}$ and $\ddot{X}_{it}$ produce the same residuals as the dummy variable regression, and hence, both procedures yield the same estimates (Wooldridge, 2021; Cunningham, 2021; Hansen, 2022). Estimates should always be interpreted in the dummy variable equation, as this allows comparing different units in the treatment and control groups at any point in time, while the transformed equations do not (Wooldridge, 2010).

5.1 Identification

We want to identify a set of conditions under which θ, as estimated via a DiD approach, can be identified as the causal effect of an intervention D_{it} on an observed outcome Y_{it}. Many of the necessary assumptions for our task can be carried over or adapted from the within estimator studied in the previous section. However, a critical addition must be made to ensure the interpretability of the parameter of interest as a casual effect.

Assuming the conditions presented below are fulfilled, then θ is the causal impact of D_{it} on Y_{it} (Hansen, 2022).

1. The data generating process follows the two-way error model described in eq. (10).

2. The regressor matrix has full rank in expectation:

$$\operatorname{rank}\left(E\left[\sum_{t=1}^{T}(\ddot{D}_{it}, \ddot{X}'_{it})(\ddot{D}_{it}, \ddot{X}'_{it})'\right]\right) = K + 1$$

3. Exogeneity of X_{it}.

$$E[X_{it}\varepsilon_{is}] = 0,\ \forall t, s$$

4. D_{it} is conditionally independent from ε_{it}.

$$D_{it} \perp\!\!\!\perp \varepsilon_{is} \mid X_{i1}, X_{i2}, ..., X_{iT}, \forall t, s$$

To see this, consider the average effect of the treated:

$$\begin{aligned} E[\ddot{Y}(1)_{it} - \ddot{Y}(0)_{it}] &= E\left[E(\ddot{Y}_{it} \mid \ddot{D}_{it} = 1, \ddot{X}_{it}) - E(\ddot{Y}_{it} \mid \ddot{D}_{it} = 0, \ddot{X}_{it})\right] \\ &= \theta + E\left[E(\ddot{\varepsilon}_{it} \mid \ddot{D}_{it} = 1, \ddot{X}_{it}) - E(\ddot{\varepsilon}_{it} \mid \ddot{D}_{it} = 0, \ddot{X}_{it})\right] \\ &= \theta + E\left[E(\ddot{Y}_{it} \mid \ddot{X}_{it}) - E(\ddot{Y}_{it} \mid \ddot{X}_{it})\right] \\ &= \theta \end{aligned} \tag{13}$$

The first equality follows from the law of total expectation and condition four, and the third follows from condition four again.

5.1.1 The assumptions in detail

The first assumption states that the outcome equals the specified two-way error component model, albeit it carries several critical implications. First, it requires that the unobserved confounders enter the data-generating process linearly and that v_t and u_i are indeed unit and time-invariant, respectively (Wing et al., 2018; Hansen, 2022). In this form, the model conveys that absent treatment, the control and treatment groups would, in expectation, evolve according to the sum of the unobserved confounders and the set of control variables. This is the common pre-trend assumption discussed in the first section (Angrist & Pischke, 2009, 2014; Wing et al., 2018).

Most importantly, the first assumption necessitates for treatment to have the same marginal impact on the outcome for all individuals in the treated and control groups. That is, the treatment effect θ is to be shared across all units in the panel (Cameron & Trivedi, 2005; Hansen, 2022; Hsiao, 2022). Although the presence of heterogeneous treatment effects does not violate the general treatment effect framework, its analysis lies beyond the scope of this work. A more severe issue arises whenever the control effect, i.e. the difference in the outcome variable across periods for the untreated group, is heterogeneous. If this were the case, the DiD estimation approach is misspecified, simply

because no credible control sample can be found (Hansen, 2022).

The third condition is the standard exogeneity assumption of fixed-effects estimation reviewed in the previous section. Similarly, the full rank condition presented above is akin to the one studied earlier, albeit with one more covariate.

Finally, condition four is the fundamental exogeneity assumption for DiD estimation (13). This seemingly simple assumption carries several critical implications with it. In essence, it is this condition that helps us get rid of the selection bias studied before (Hsiao, 2022). Under the conditional independence assumption, treatment is not enacted in response to knowledge about its effect on the outcome variable and said outcome variable does not change in anticipation of the policy instrumentation. That is, units with a (conditional) higher outcome potential are not selectively treated. This allows for an approximate experiment in the absence of random assignment (Angrist & Pischke, 2009, 2014; Wing et al., 2018).

Assumption four also requires that any coincident events with the intervention of interest D_{it} do not affect outcome Y_{it}. If this is not fulfilled, θ would capture the effect of the said coincident event and hence provide a biased estimate of the causal effect of the intervention (Angrist & Pischke, 2009, 2014; Hansen, 2022).

The conditional independence assumption cannot be formally tested. Instead, its credibility lies upon a well-constructed argument for it is existence (Lechner et al., 2011; Hansen, 2022)

5.2 Trended variables

So far, we have only allowed for time-invariant individual-specific effects. However, in applications which cover more extended time periods, trends might differ across units in a way which is not captured by the included controls X_{it} or the common time fixed-effect v_t (Angrist & Pischke, 2009; Wing et al., 2018; Hansen, 2022).

Consider a generalization of the two-way error component model introduced earlier (10), which included unit-specific linear time trends:

$$Y_{it} = \theta D_{it} + X'_{it}\beta + v_t + tw_i + u_i + \varepsilon_{it} \tag{14}$$

where t denotes the time trend and w_i the unit specific time trend fixed-effect.

As before, estimation of eq. (14) can be done through a dummy variable regression, with an interaction between a dummy indicating a specific unit and the time trend as an additional regressor. However, the large number of parameters to be estimated under this setup can lead to the problem of overidentification. As a rule, for the coefficients to be uniquely identified, we require that $T \geq 4$. Moreover, applications with few periods are generally inadequate to pin down the state-specific trends (Angrist & Pischke, 2009; Hansen, 2022).

If the unit-varying trends are not specified in the model, the estimated causal effect of the treatment can be inconsistent due to omitted variable bias (Angrist & Pischke, 2009; Hansen, 2022). For example, Besley & Burgess (2004) studied the effect of labour regulation on business output in states across Inda using a DiD setup. The identification strategy was based on exogenous changes in regulatory regimes in different states. Estimates from the DiD model excluding state-specific trends pointed out a significant detrimental effect of regulation on output. However, once these are included in the model, the effect vanished since labour regulation increased in Indian states where business output already had a downward trend.

5.3 Inference

The use of highly aggregated data in many DiD applications can lead to errors being correlated within groups and across time (Bertrand et al., 2004; Angrist & Pischke, 2009, 2014; Hansen, 2022). This dependence is referred to as clustered data. In Section 2, we covered how neglecting this can lead to the underestimation of standard errors and the effect this can have on the significance of estimations in the case of Card & Krueger (1994). Here, we dive back into the topic through a more theoretical approach.

With the use of the work by Moulton (1986, 1990) and Kloek (1981) we will set up the fundamental problem regarding correlated error terms. We return to a basic OLS framework and consider a general linear regression model with clustered data:

$$\mathbf{Y}_g = \mathbf{X}_g\beta + \mathbf{e}_g \tag{15}$$

Where $g = 1, .., G$ is the cluster-level index, and G represents the total number of clusters. Further, suppose each cluster has m observations. The error term is a $m \times 1$ vector described by $\mathbf{e}_g = (e_{1g}, ..., e_{mg})'$. We consider a model with homoscedastic disturbances where $E\left[e_{ig}^2 \mid X_g\right] = \sigma^2$ and $E\left[e_{ig}e_{lg} \mid X_g\right] = \sigma^2\rho$, $i \neq l$; that is, the error terms have an intracluster correlation of ρ.

Under this setup, a good approximation[6] for the variance of the OLS estimate is:

$$\mathbf{V}_{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2 (1 + \rho(m-1)) \tag{16}$$

Equation (16) shows that the presence of clustered data causes the OLS estimator variance to increase by $(1 + \rho(m-1))$. Ignoring this would thus lead to an underestimation of the estimates variance.

Several methods have surged over the years to account for this phenomenon when using DiD estimation. The most popular way to account for clustering data in DiD is clustered standard errors, which adjust the usual variance-covariance matrix by accounting for the correlation in the disturbance terms. Other methods include collapsing the data by averaging over the pre and post-intervention observations and assigning arbitrary covariance structures based on an order one autoregressive model. Bertrand et al. (2004) provide an overview of such techniques in their paper.

5.4 DiD in pratice: Di Tella & Schargrodsky (2004)

In their 2004 paper, Di Tella & Schargrodsky studied the causal effect of policing on crime rates. When studiying this phenomenon, one faces the issue of reverse casuality since police forces are usually allocated endogenously to areas where crime is rampant. DiD estimation could solve this problem via the conditional independence assumption reviewed earlier. Throughout the rest of this section, we conduct a replication exercise of the key results form Di Tella & Schargrodsky (2004) and further review some of the critical assumptions for identification in DiD estimation.

The terrorist attack in July 1994 on the main Jewish centre in Buenos Aires pushed

[6]For the formula to be exact, we would need all regressors to be fixed within clusters. Nonetheless, when regressors which do not fulfil this characteristic are incorporated into the model, the formula still provides for a good numerical approximation (Moulton, 1990).

the government to provide police protection to all Jewish institutions in Argentina. This translated to an exogenous allocation of police forces to certain city blocks within the country as forces were deployed to deter further terrorist attacks but not in response to crime rates in Argentinas neighbourhoods. Thus, the occurrence of the terrorist attack and the extended vigilance had no relation to street crime rates. Nonetheless, the policing efforts could have had a general deterrence effect. With this argument, Di Tella & Schargrodsky present the allocation of police forces following the terrorist attack as a valid treatment.

The authors set up a DiD approach comparing the average number of car thefts per a selection of blocks in Buenos Aires, before and after the terrorist attack, and between city blocks with and without a Jewish institution. The document at hand provides an excellent example of a coherent and credibly identified DiD application through which the theoretical aspects of causal inference can be further put to light and tested in practice; this is in contrast to Card & Krueger (1994), where the feasibility of some of the assumptions where put to the test earlier in this document.

5.4.1 The replication exercise

Di Tella & Schargrodsky originally estimated the following version of the model:

$$\begin{aligned} CarTheft_{it} =& \theta_1(SameBlock \times Post)_{it} + \theta_2(OneBlock \times Post)_{it} \\ &+ \theta_3(TwoBlock \times Post)_{it} + \tau_t' v + u_i + \varepsilon_{it} \end{aligned} \tag{17}$$

where v denotes the month fixed effects and u_i the block fixed effects. Again, $CarTheft_{it}$ represents the average monthly car thefts per block. Further, $(SameBlock \times Post)_{it}$ denotes city blocks with a Jewish institution and captures information after the terrorist attack. Similarly, $(OneBlock \times Post)_{it}$ and $(TwoBlock \times Post)_{it}$ stand for blocks one and two blocks away from a Jewish institution, respectively, after the terrorist attack.

Additionally to eq. (17), we follow Hansen (2022) and estimate the following version of the model for our exercise:

$$CarTheft_{it} = \beta_0 + \theta(SameBlock \times Post)_{it} + \tau_t' v + u_i + \varepsilon_{it} \tag{18}$$

The DiD estimations were carried out using block-clustered standard errors, and the rank assumption was tested before running the regressions.

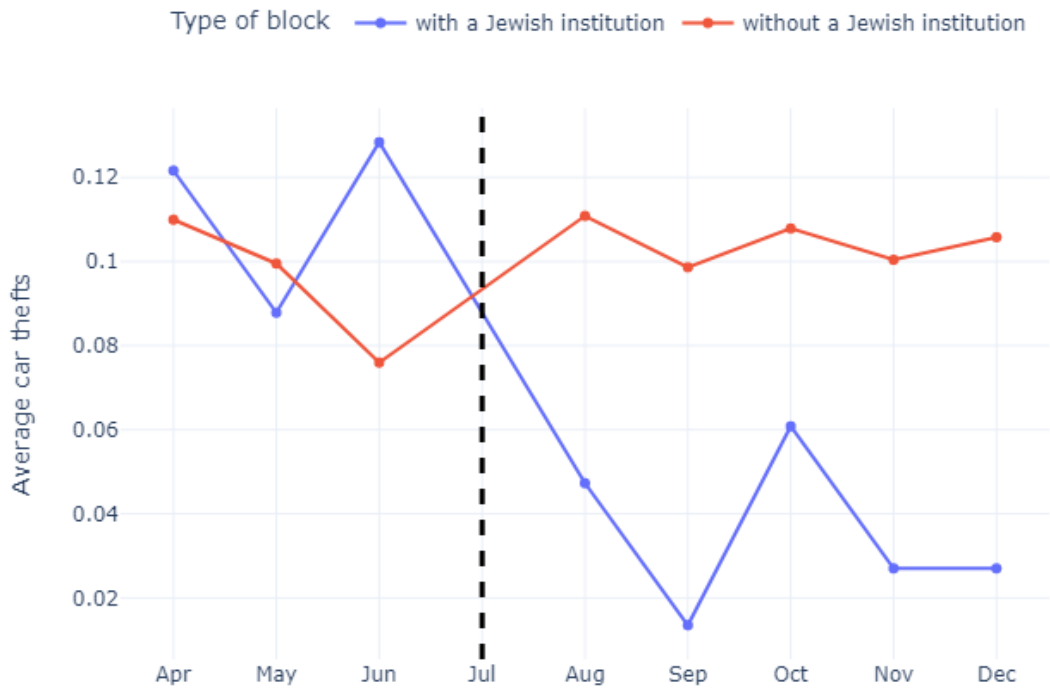
Results The summary of the obtained results can be found in Appendix C.1. City blocks with a Jewish institution perceived a significant decline in car theft after the terrorist attack and the subsequent allocation of police forces (Figure 6). The estimates in the second column suggest that policing caused a reduction of 0.087 average car thefts per month, a reduction of 77% with respect to the average before the attack (Appendix C.2). However, no significant effect was found for the areas one and two blocks away from a Jewish institution. In light of this, we stick to the modified version of the model for the rest of this section.

Testing the assumptions Recall that identification under eq. (18) requires homogeneous treatment effects. In our setup, this would imply that the treatment effect does not significantly vary over the months included in the analysis after the attack. We formally test for this via a regression exclusion analysis. The idea is to include dummies for the interactions of the post-month attacks and the blocks with a Jewish institution. The results are presented in Appendix C.1. Indeed, we cannot reject the hypothesis that the treatment effect is homogeneous, as the interactions included are insignificant.

Further, the identifying assumption in DiD estimation regarding the data-generating process requires common pre-trends. Figure 6 presents some evidence in favour of this. We can see that monthly average car thefts were similar across city blocks in the months before the terrorist attack, ranging between 0.080 and 0.128. Moreover, Di Tella & Schargrodsky (2004) empirically tested for the existence of different crime dynamics prior to the terrorist attack in city blocks with and without a Jewish institution and found no significant results.

Finally, the authors do not discuss the issue of coincident events; however, the nature of the attack makes it implausible for another phenomenon to explain away the large estimated treatment effect. Hence, we can confidently argue that the model is correctly identified and that police presence does have a causal effect on street crime.

Figure 6: Average Car Thefts in Buenos Aires per Block, 1994



6 Two-way fixed effect regression and difference in difference estimator

6.1 Overview

The TWFEr regression is often used in Economics and other social sciences to analyze causality from panel data. The main reasons to use this method are its ability to adjust for time, and unit effects and its resemblance to a generalization of the DiD estimator in the two groups and two periods design. This section aims to show that outside of that design, the two estimators (TWFEr and DiD estimator) are very different, each with its own strengths and weaknesses. A better understanding of how these two estimators work and why they could give different results is a useful tool for applied research. To draw a proper comparison between the two, we will follow what has been done by Imai &

Kim (2021) and proceed by using their equivalent matching estimators[7] as they excel in displaying counterfactuals for observation. Pictures from Imai & Kim (2021) will be used to show visually how the estimators use different observations for the counterfactual.

An observation's counterfactual is the value it would have reached if it belonged to the opposite treatment status. For example, if an observation is treated, its counterfactual is the value it would have reached if it was not treated. The opposite applies to control observations. By underlining the differences in how DiD and TWFEr compute counterfactuals, a clearer view of the behaviour of the two estimators will be given.

All of the analysis will be conducted in a different framework than before. The treatment variable is homogeneous, non-staggered (units can enter and leave the treatment at any time), all units can receive treatment, and our analysis will not be limited to two time periods.

The model that will be used in all of the paper will be:

$$Y_{i,t} = u_i + v_t + \theta X_{i,t} + e_{i,t} \tag{19}$$

Our model is composed of a time-invariant unit effect u_i, a unit invariant time effect v_t, an error term e_{it} and a dummy variable X_{it} (one if treated, zero if untreated, treatment is group independent and assigned casually to observations, all observations with $X_{it} = 1$ will be called "treated". In contrast, all the ones with $X_{it} = 0$ "untreated") this is the only covariate present in the model and the only one the matching estimators will care about when matching observations.

6.2 Two-way fixed effect regression

$$\begin{aligned}\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}[\{(Y_{i,t} - \bar{Y}) - (\bar{Y_i} - \bar{Y}) - (\bar{Y_t} - \bar{Y})\} \\ - \theta\{(X_{i,t} - \bar{X}) - (\bar{X_i} - \bar{X}) - (\bar{X_t} - \bar{X})\}]^2\end{aligned} \tag{20}$$

This estimator has been analyzed in Section 3.3. We already stressed how it can adjust for time-invariant unit effects and unit-invariant time effects and their limits. Instead,

[7]Part 2 and 3 are taken from the Imai & Kim (2021), while the idea of the leaving effect from Imai et al. (2021) its formalization, for what it counts, is my own while the later simulations and the combination they express are my own.

the focus will be on how the method estimates the treatment effect θ. To do so, we reach the equivalent matching method through a series of equations (the equivalence has been shown in Imai & Kim (2021) and is outside this paper's scope).

$$\hat{\theta} = \frac{1}{K}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(X_{i,t}(Y_{i,t} - \widehat{Y_{i,t}(0)}) + (1 - X_{i,t})(\widehat{Y_{i,t}(1)} - Y_{i,t})\Big)\right] \tag{21}$$

As shown above, matching methods try to compare each observation, given its treatment status, with its **potential outcome of opposite treatment status (counterfactual)**[8] "what the outcome variable would have been if everything was the same but the treatment". Treated observations will use the first difference inside the formula ($X_{i,t} = 1$) and untreated will use the second ($1 - X_{i,t} = 1$), the difference between the observation and its counterfactual represent the treatment effect. To estimate the counterfactual of an observation, we use the other observations available in our sample.

It is important to notice that the model object of our analysis is composed of unit effect, time effect, treatment and error. No other covariates are present. The only covariate available, as a consequence, the only one that the matching estimator cares about, is the treatment status. To build a proper counterfactual, we would need the potential outcome for that observation (hence, the same unit and time) but with the opposite treatment status. This is why it is possible to find an **equivalent matching method** without relying on propensity score matching or other techniques to make our counterfactual closer in covariates values to the original observation.

Here below, we present how the matching-TWFEr equivalent method computes the counterfactual of a specific observation, with x being the treatment status ($x \in \{0, 1\}$).

$$\widehat{Y_{it}(x)} = \frac{1}{T-1}\sum_{t' \neq t} Y_{i,t'} + \frac{1}{N-1}\sum_{i' \neq i} Y_{i',t} - \frac{1}{(T-1)(N-1)}\sum_{i' \neq i}\sum_{t' \neq t} Y_{i',t'} \tag{22}$$

Equation (22) computes the potential value of[9] of Y_{it} by averaging over all the observations with the same unit (excluding same time) plus all of the observations with the

[8] $\widehat{Y_{i,t}(x)}, x \epsilon \{0, 1\}$ is the potential outcome of unit i at time t with treatment status x, it is called counterfactual when for an observation the potential outcome has opposite treatment status compared to $X_{i,t}$.

[9] By using as an example a 2 by 2 design (easy to expand to a larger design) with $X_{1,1}, X_{1,2}, X_{2,1} = 0$ and $X_{2,2} = 1$ the counterfactual $\widehat{Y_{2,2}(0)} = t_2 + u_1 + u_2 + t_1 - u_1 - t_1$, if a treated unit were to be present, the treatment effect θ would have been inside the past formula.

same time (excluding same unit) and by subtracting the mean of all the observations that share neither unit nor time.

The model (19) is made of unit FE, time FE, treatment status (multiplied by θ), and the counterfactual of a treated observation would ideally contain all of the terms as above but with the opposite treatment status. A graphical illustration Figure 7 will make clearer which observations are used for the counterfactual estimate of $Y_{4,3}$.

Figure 7: Two-way Fixed Effect Estimator

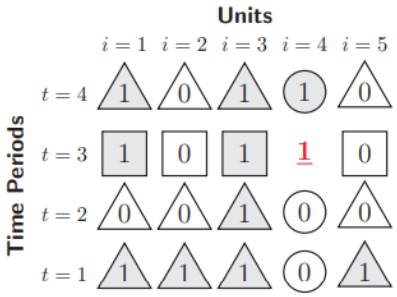
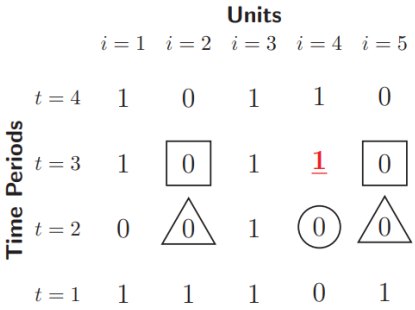


Figure 8: Differences in Differences Estimator



In the first figure one represent treated units while zero control units. The observation for which we want to compute the counterfactual is the red one. In this case, observations that share the same units are circles (first member in equation (22)), squares share the same time (second term in (22)), while triangles do not share either time or unit (last member in (22)). For example in this figure, $Y_{4,3}$ has many mismatches, one shares the same unit ($Y_{4,4}$), two with the same time $Y_{3,3}, Y_{1,3}$ and more in the adjustment set. On the other the second pictures does not have any mismatch, but relies on a smaller number of observations.

As shown in formula 22, TWFEr uses all but the same observation to compute its counterfactual, this implies that even units with the same treatment status (so-called mismatches[10]) do matter for our counterfactual. The treatment status of the counterfactual will not be the opposite, but the sum of the three averages[11]. This leads to a biased counterfactual.

Including observations with the same treatment status as the selected observation implies that the difference between a unit and its counterfactual will be smaller than expected. For this reason, Imai & Kim (2021) named it **"attenuation bias"**.

Even though the TWFEr counterfactual could be biased, this will not be the case for the estimate of the treatment effect ($\hat{\theta}$). In the $\hat{\theta}$ of the equivalent matching estimator,

[10] A mismatch is an observation that is used to estimate the counterfactual of another observation with which shares the same treatment status.

[11] In the example above the treatment status of $\widehat{Y_{4,3}(0)}$ is $\frac{1}{3} + \frac{2}{4} - \frac{7}{12}$

TWFEr adjusts for this bias through K[12], counting where and how many correct matches are present when computing $\hat{\theta}$ and rebalancing[13]. The following points are relevant to the analysis and are the object of a comparison with the DiD estimator. All the observations in the dataset contribute to the estimate of a single counterfactual, even the ones with the same treatment status.

Even if observations do not contribute with the same weight for a single counterfactual (sharing the same time or same unit implies a bigger weight), they will have the same weight since all the counterfactuals of all observations matter. The estimator's variance is mainly given by the error in the realized $Y_{i,t}$. The errors in the counterfactual are less relevant to our ends.

6.3 Differences in Differences

Since our dataset is made of more than two periods, the DiD estimator will be different from the one we have seen before in the paper (2×2 design), even though the idea will be the same.

With DiD, we also need to assume parallel trends. This assumption has already been investigated in Section 5 before. In the DiD estimator comparing each observation with its counterfactual is intuitive[14]. It is possible to estimate the treatment effect only on units that change their treatment status in the sample. Those units can be considered only if at least one unit had the same treatment status in t-1 and kept it in t. Without the presence of this unit, it would be impossible to build a counterfactual and to have a comparison to our unit.

We use control observation with the same time, with the same unit (but one time before) and other control observations that share unit and time with the one before. Since the DiD estimator computes a counterfactual only for units that received treatment, our method will compute the average treatment effect on the treated (ATT).

[12] formula is left to the appendix

[13] The following statement comes from OLS regression estimating a dummy variable without any bias

[14] Differences in Differences could be already considered as a matching estimator; by comparing a unit that goes from no treatment to treatment with one that stays untreated, we are building a counterfactual for that very same unit. Assume $X_{1,1}, X_{1,2}, X_{2,1} = 0 and X_{2,2} = 1$ then $Y_{2,2} - \widehat{Y_{2,2}(0)} = Y_{2,2} - (Y_{1,2} + Y_{2,1} - Y_{1,1}) = u_2 + v_2 + \theta - (u_1 + v_2 + u_2 - v_1 - u_1 + v_1) = \theta$

$$\tau = E[Y_{i,t}(1) - Y_{i,t}(0)|X_{i,t} = 1, X_{i,t-1} = 0]$$

To compute the counterfactual, we introduce some notation:
$M_{it}^{DiD} = \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 0\}$ the set containing the observation with the same unit, but that in a time period before and only if it was an untreated observation.
$N_{it}^{DiD} = \{(i', t') : i' \neq i, t' = t, X_{i't'} = 0, X_{i',t'-1} = 0\}$ the set containing observation with the same time, but different units, an observation belongs to this set if and only if it was untreated in time period t and t-1.
$A_{it}^{DiD} = \{(i', t') : i' \neq i, t' = t - 1, X_{i't'} = 0, X_{i',t'+1} = 0\}$ the set containing all the observation that are untreated in t-1 and t and have time t-1 and unit different from i.

$$\widehat{Y_{it}(0)} = Y_{it-1} + \frac{1}{|N_{it}^{DiD}|} \sum_{(i',t)\epsilon N_{it}^{DiD}} Y_{i't} - \frac{1}{|A_{it}^{DiD}|} \sum_{(i',t')\epsilon A_{it}^{DiD}} Y_{i't'} \tag{23}$$

The counterfactual of an observation (in (23) is for treated observations, as they are the only relevant for the DiD estimator to compute a potential outcome) in (23) is made of the outcome of the dependent variable of the same unit in a period before (the member of M_{it}^{DiD}) plus the average of the members of N_{it}^{DiD} minus the average of the members of A_{it}^{DiD}. It is coherent with the 2×2 design.

The variable $D_{i,t}$ tells us if an observation is suited to compute the ATT ($\hat{\theta}$) by taking into account if the observation is treated and if there exist other observations such that the two sets N_{it}^{DiD} and M_{it}^{DiD} are not empty.

$$t = 1, D_{it} = 0,$$

$$t \neq 1, D_{it} = X_{it} 1\left\{|N_{it}^{DiD}||M_{it}^{DiD}|\right\}$$

As a consequence, the ATT formula will be:

$$\hat{\tau} = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(Y_{it} - \widehat{Y_{it}(0)}\right) \tag{24}$$

The ATT formula sums over all observations (through unit and time sums) for the product between the variable D_{it} and the difference between the outcome variable and

its counterfactual. All is divided by the number of observations suitable for the ATT. From (23), it is possible to notice that not all the observations carry the same weight in computing the treatment effect. Some will not show up there ($D_{i,t} = 0$), while others will but with a smaller weight (used only in counterfactuals).

The following characteristics are of the DiD estimator. Not all observations matter for computing the counterfactual of an observation, only the one with the opposite treatment status, this makes the counterfactual unbiased. The number of observations whose counterfactual is available is always inferior to the full sample, this leads to a bigger variance in the estimator. It could happen that an observation cannot be found in computing any counterfactual; in that case, it will not carry any weight to compute the treatment effect.

6.4 Treatment effect

In the last section, the treatment effect on the treated has been estimated when units received the treatment moving from a no treatment status to a treatment status, and this is always the case if the treatment is staggered. But, by being in a context where units can enter and leave the treatment, it is possible to compute the treatment effect when units are leaving the treatment. In this way, it is possible to separate the **entering effect** and the **leaving effect**. What has been explained in the DiD section was the entering effect. Now it will be shown how to compute the leaving effect; the procedure will be symmetrical to what was explained before.

$$ATT_{leave} = E[Y_{i,t}(1) - Y_{i,t}(0)|X_{i,t} = 0, X_{i,t-1} = 1] \tag{25}$$

What we want to compute is the difference between a unit that goes from treatment to no treatment. The counterfactual for this is a treated unit that remained in the treatment. The same notation as before will be used. The observations object of the ATT_{leave} will be the ones that are now untreated but were treated in the period before. $M_{it}^{DiD_leave} = \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 1\}$the set containing the observation with the same unit, but that in a time period before and only if it was a treated observation. $N_{it}^{DiD\_leave} = \{(i', t') : i' \neq i, t' = t, X_{i',t'} = 1, X_{i',t'-1} = 1\}$ the set containing observations with the same time, but different units, an observation belongs to this set if and only if

it was treated in time period t and t-1
$A_{it}^{DiD\_leave} = \{(i', t') : i' \neq i, t' = t-1, X_{i't'} = 1, X_{i't'+1} = 1\}$ the set containing all the observation that are untreated in t-1 and t and have time t-1 and unit different from i.

$$\widehat{Y_{it}(1)} = Y_{it-1} + \frac{1}{|N_{it}^{DiD\_leave}|} \sum_{(i',t)\epsilon N_{it}^{DiD\_leave}} Y_{i't} - \frac{1}{|A_{it}^{DiD\_leave}|} \sum_{(i',t')\epsilon A_{it}^{DiD}} Y_{i't'} \tag{26}$$

The procedure for computing the counterfactual in (26) is the same as it was in (23) but with the new sets. The variable $D_{i,t}$ tells us if an observation is suited to compute the ATT ($\widehat{\theta_{leave}}$), by taking into account if the observation is untreated and if there exist other observations such that the two sets $N_{it}^{DiD\_leave}$ and $M_{it}^{DiD\_leave}$ are not empty.

$$t = 1, D_{it} = 0,$$

$$t \neq 1, D_{it} = (1 - X_{it}) 1\left\{|N_{it}^{DiD\_leave}||M_{it}^{DiD\_leave}|\right\}$$

As a consequence, the ATT_{leave} formula will be:

$$\widehat{ATT_{leave}} = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\widehat{Y_{it}(1)} - Y_{it}\right)$$

Notice that the ATT_{leave} computes the difference between being treated and untreated. A positive number implies that going from treated to untreated is a loss of value.

By using a TWFEr it is impossible to separate the entering and leaving effects. They are both present in the treatment effect estimate; this is due to the ordinary least square assumption of linearity and leads to assuming symmetric effects for entering and leaving the treatment. It could be the case that the magnitude and significance of the treatment are due to just one of the two. An interesting view on this topic has been given in a replication study in Imai et al. (2021) where they used this very same method compared to TWFEr replicating the work of Acemoglu et al. (2019) and more general example will be given later.

A note on the side should be made: the leaving effect can be computed if and only if there are observations leaving the treatment. This cannot happen in a non-staggered design.

6.5 Simulations

The main difference between the two estimators is the observations used to estimate the treatment effect and how this exposes the two methods to specific weaknesses.

All the simulations will be conducted in the following way. A dataset is generated following a specific model shown below. First, DiD and TWFEr models are estimated on the standard model present in the Overview (19), and all of it is repeated 30 times. Then we look at all the DiD and TWFEr estimates and compare them. All the parameters to generate the data will be shown. The number of units and time periods are always set to 20. The probability of treatment is set at 40%. The error is normally distributed (mean is zero, and variance amounts to one).

The first simulation study regards the existence of a difference in entering the treatment (θ_{enter}) and leaving the treatment (θ_{leave}). Our model (19) does not specify a difference between the two and is missing these parameters. We expect TWFEr to generate an estimate between the two θs, while DiD to generate a precise estimate of θ_{enter}. We will also use DiD to compute an unbiased ATT_{leave} to show the opportunities of this powerful estimator.

The model (19) is assumed when computing the estimates, but the true one is presented below.

$$te_{i,t} = |\{X_{i',t'}|i' = i, t'\epsilon[2,t]X_{i',t'} = 1, X_{i',t'-1} = 0\}|\text{times the treatment was entered}$$
$$tl_{i,t} = |\{X_{i',t'}|i' = i, t'\epsilon[2,t]X_{i',t'} = 0, X_{i',t'-1} = 1\}|\text{times the treatment was left}$$

$$t \neq 1, Y_{i,t} = u_i + v_t + \theta_{enter}te_{i,t} - \theta_{leave}tl_{i,t} + e_{i,t}$$

$$t = 1, Y_{i,t} = u_i + v_t + \theta_{enter}X_{i,t} + e_{i,t}$$

Unit, time fixed effects and treatment are the same as in the first model introduce (19). However, every time a unit i enters the treatment, a factor θ_{enter} is added, and every time it leaves the treatment, a factor θ_{leave} is subtracted.

How is TWFE behaving with this model? Let us simplify and ignore time and unit FE for a moment[15]. For example, assume T=4 and focus on a single unit and $X_1 =$

[15]In this model misspecification, we are not stressing the ability to adjust for time and unit effects, but the linear regression assumptions of linearity and its consequence of assuming symmetry when receiving and leaving treatment if treatment is specified as dummy variable

$1, X_2 = 0, X_3 = 1, X_4 = 0$. Then, the treatment effect would be:

$$\widehat{\theta} = \bar{Y(1)} - \bar{Y(0)} = \frac{Y_1 + Y_3}{2} - \frac{Y_2 + Y_4}{2} =$$

$$= \frac{(\theta_{enter}) + (2\theta_{enter} - \theta_{leave})}{2} - \frac{(\theta_{enter} - \theta_{leave}) + 2(\theta_{enter} - \theta_{leave})}{2} = \theta_{leave}$$

If units never left treatment, then there would be no θ_{leave} in the equation and $\theta = \theta_{enter}$. The estimate is dependent on how many untreated units are before being treated for the first time and how many are after[16]. If they are all after, θ_{leave} will dominate; in the opposite case θ_{enter} will. In the first case, the difference between treated and untreated amounts up to θ_{enter}, while in the second one is only θ_{leave}. The estimate of the linear regression will always be between the two. In the simulations, it will be the average (random designs are repeated)[17]. In the DiD estimator:

$$\theta_{leave} \text{ will be perfectly computed by } ATT_{leave}$$
$$\theta_{enter} \text{ will be perfectly computed by } ATT_{enter}$$

Assuming a bigger θ_{leave} than θ_{enter} will imply that all observations after they left the treatment will be smaller than if they never received it. This will make our treatment effect estimates bigger. On the other hand, a smaller θ_{leave} will make all of our untreated observations after leaving treatment bigger than if they never received it.

These two simulations showed how varying both of them will always change values in the TWFEr estimate.

It is interesting in the first picture that if our θ_{leave} is smaller (bigger) than θ_{enter} effect, the TWFEr will compute a smaller (bigger) treatment estimate even if adopting the policy is better (worse) compared to a case with a bigger (smaller) θ_{leave} that will make our TWFEr estimate bigger (smaller). It is better (worse) because if tomorrow the policy cannot continue anymore, leaving it will result in smaller (bigger) damage compared to the case with symmetric treatment effects (OLS assumption), implying that overall the dependent untreated variable will be higher (lower) by $\theta_{enter} - \theta_{leave}$. Therefore, it is

[16]Imagine 2 observations of the same unit, $X_1 = 0$ and $X_2 = 1$, the treatment would be $Y_1 - Y_0 = \theta_{enter}$, instead if it was to be the opposite case, first a treated observation and then untreated $Y_0 - Y_1 = \theta_{enter} - (\theta_{enter} - \theta_{leave}) = \theta_l eave$

[17]In a TWFE environment, our treatment effect estimate would also be influenced by how many observations have never been in treated in units other than the one the counterfactual is estimated

Figure 9: Leaving Treatment

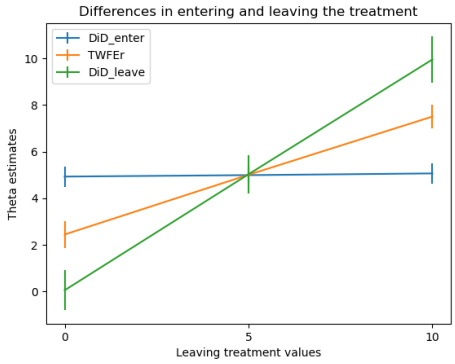
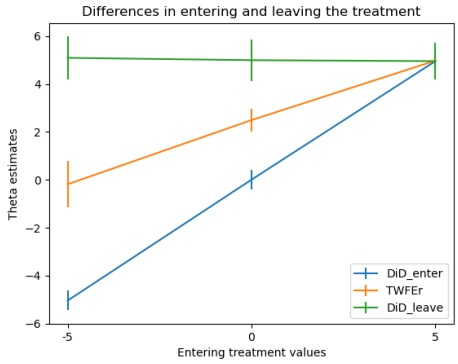


Figure 10: Entering Treatment



The simulations are run with the values used on the x axis, in this case three for each figure.

really important to detach the view of the treatment effect (in non-staggered contexts) as only entering effect or as treatment will make our variable bigger by θ.

In the second figure, it should be noticed how TWFEr has significant estimates even if adopting the treatment is not significant at all ($\theta_{enter} = 0$). If the average of the two treatment effects is zero, TWFEr is estimating a non-significant treatment effect while both of them are significant ($\theta_{enter} - \theta_{leave} = -10$), meaning that deciding to adopt the policy will make us worse by θ_{enter} and leaving it would do it by $-\theta_{leave}$, it would be better not to adopt the policy, but TWFEr suggests that is not going to be significantly different from zero and same applies for positive values. If the act of changing policies leads to a positive or negative effect, this cannot be always identified by TWFEr. A second simulation study will be left in the Appendix. .

7 Conclusions

In this paper, we explore the topic of causal inference via DiD. We provide a replication of (Card & Krueger, 1994) to deliver an intuitive base for understanding the method. In addition, we discuss the critical assumptions of common pre-trends and conditional independence and argue how they might be violated in this setting. Further, get a first glance at inference when using clustered data.

We then provide the necessary theoretical background in panel data econometrics for the formal discussion of the DiD setup while diving deep into the one- and two-way

fixed effects estimators. Next, we formally show how these methods allow for unbiased estimation under the presence of time and unit-fixed unobserved confounders through the demeaning of the regression model. Lastly, we cover the critical assumption for the validity of the one- and two-way fixed effects estimators.

Further, we compare these estimators by applying a series of Monte Carlo simulations. First, we visualise how TWFE estimator performs well when there is enough variation in the covariates across time. Further, we show how TWFE outperforms OWFE when the data exhibits trended variables; however, controlling for trends becomes necessary when the drift is strong; else, the TWFE exhibits efficiency loss. Then, we demonstrate how double demeaning eliminates the bias otherwise generated by unobserved confounders, which are correlated with the covariates.

Next, we set up the DiD model with the theoretical tools discussed earlier. We show how identifying a treatment conditionally independent from the idiosyncratic errors allows for estimating the former's causal effect on an outcome of interest, thus solving the problem of causal inference, and emphasise the importance of homogeneous treatment effects under the classical DiD setup. After that, we reassume the discussion of trended variables by providing an extension of the DiD framework. We also provide a theoretical discussion of the effects of clustering on the efficiency of linear estimators. Finally, we show how the validity of the assumptions discussed earlier is critical for DiD estimation in a replication exercise of Di Tella & Schargrodsky (2004).

We extend the discussion by comparing the DiD estimator and the TWFEr under more complex setups. The two estimators behave variously regarding different model misspecifications. While TWFEr has a lower variance, it is also more vulnerable to misspecifications. We show how the estimates of the causal effect from TWFEr could be insignificant if leaving and entering the treatment have different values. We advise against utilising the TWFE model as a generalisation of DiD estimation.

Bibliography

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2022). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*(1), 1–35.

Acemoglu, D., Naidu, S., Restrepo, P., & Robinson, J. A. (2019). Democracy does cause growth. *Journal of political economy*, *127*(1), 47–100.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton university press.

Balestra, P., & Nerlove, M. (1966). Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica: Journal of the econometric society*, (pp. 585–612).

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, *119*(1), 249–275.

Besley, T., & Burgess, R. (2004). Can labor regulation hinder economic performance? evidence from india. *The Quarterly journal of economics*, *119*(1), 91–134.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Card, D. (1992). Using regional variation in wages to measure the effects of the federal minimum wage. *Ilr Review*, *46*(1), 22–37.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, *84*(4), 772–793.
URL `http://www.jstor.org/stable/2118030`

Card, D., & Krueger, A. B. (2014). *Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania.* https://davidcard.berkeley.edu/data_sets.html [Accessed: 03.12.2022].

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.
URL https://books.google.de/books?id=PSEMEAAAQBAJ

de Chaisemartin, C., & D'Haultfoeuille, X. (2022). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. Working Paper 29691, National Bureau of Economic Research.
URL http://www.nber.org/papers/w29691

Di Tella, R., & Schargrodsky, E. (2004). Do police reduce crime? estimates using the allocation of police forces after a terrorist attack. *American Economic Review*, *94*(1), 115–133.

Hansen, B. (2022). *Econometrics*. Princeton University Press.
URL https://books.google.de/books?id=Pte7zgEACAAJ

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *Test*, *16*(1), 1–22.

Hsiao, C. (2022). *Analysis of panel data*. 64. Cambridge university press.

Imai, K., & Kim, I. S. (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, *29*(3), 405–415.

Imai, K., Kim, I. S., & Wang, E. H. (2021). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*.

Kloek, T. (1981). Ols estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica: Journal of the Econometric Society*, (pp. 205–207).

Lechner, M., et al. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, *4*(3), 165–224.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, *32*(3), 385–397.
URL https://www.sciencedirect.com/science/article/pii/0304407686900217

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, (pp. 334–338).

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics*, *43*(1), 44–56.

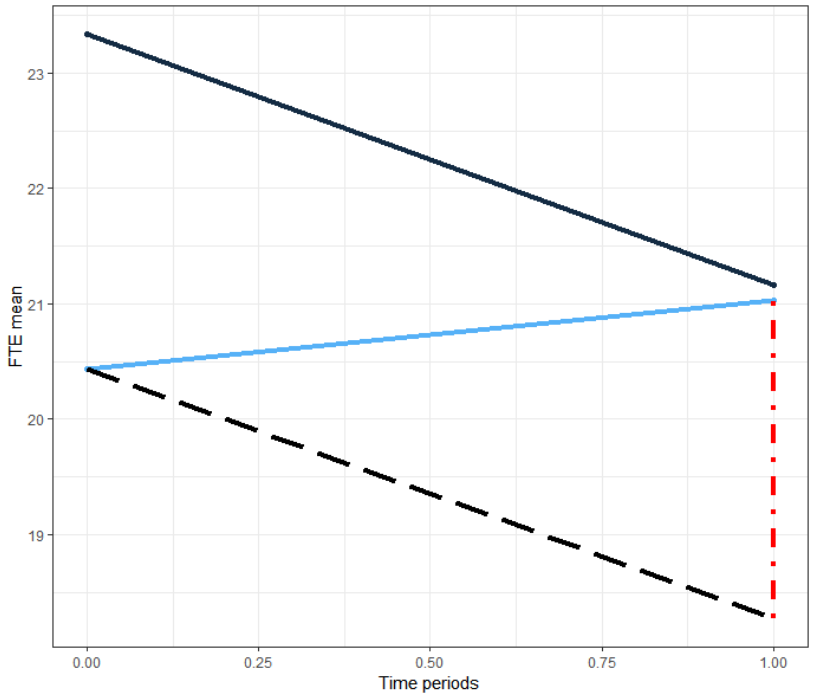
Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health*, *39*(1), 453–469.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.

A Appendix

A.1 Display of the Causal Effect in Card & Krueger (1994)



Note: The graph uses the results of table 1. The lower blue straight line is the change in employment in New Jersey. The straight line with the negative slope is the change in employment in Pennsylvania. The line with stripes shows Pennsylvania as a counterfactual for New Jersey. I.e. how New Jersey would have behaved without the treatment. The orthogonal line is the difference between the counterfacutal and New Jersey. Thus, it is the causal effect of the minimum wage introduction, given the model is correctly specified.

A.2 Distribution of store types

	Pennsylvania	New Jersey
Burger King	44.3%	41.1%
KFC	15.2%	20.5%
Roys	21.5%	24.8%
Wendys	19.0%	13.6%

Note: Card & Krueger (1994) report no significant differences of the store distribution.

B Appendix

Table 5: Root Mean Squared Error of β_2 Estimates for the Simulations

(a) Time Invariant Error Correlations c_{unit}

	0.00	0.05	0.15	0.40
Pooled	0.0065	0.0073	0.0128	0.0496
one-way	0.0024	0.0024	0.0024	0.0024
two-way	0.0038	0.0038	0.0038	0.0038

(b) Variance Across Time c_{var}

	0.03	0.10	0.25	0.50
Pooled	0.0531	0.0525	0.0496	0.0409
one-way	0.0030	0.0029	0.0024	0.0015
two-way	0.2626	0.0236	0.0038	0.0009

(c) Trended Variables c_{trend}

	0.03	0.10	0.25	0.50
Pooled	0.0603	0.1069	0.3082	0.9515
one-way	0.0053	0.0360	0.2141	0.8513
two-way	0.0038	0.0040	0.0058	0.0123

(d) Unit Invariant Error Correlations c_{time}

	0.000	0.015	0.050	0.200
Pooled	0.0496	0.0524	0.0594	0.0954
one-way	0.0024	0.0026	0.0045	0.0362
two-way	0.0038	0.0038	0.0038	0.0038

This table summarizes the root mean squared errors of the estimates in different scenarios. For each scenario, the resulting RMSE of three different estimators is included. Part(a) controls for time-invariant error correlations which are represented by c_{unit}. We set $c_{time} = 0$, $c_{trend} = 0$ and $c_{var} = 0.25$. Part(b) controls for the variation of covariates across time using c_{var}. We set $c_{time} = 0$, $c_{trend} = 0$ and $c_{unit} = 0.4$. Part(c) includes the trended variables into the model and control the intensity with c_{trend}. We set $c_{time} = 0$, $c_{var} = 0.25$ and $c_{unit} = 0.4$ in this part. Finally, Part(d) controls for the correlation of unit-invariant error terms and we set $c_{trend} = 0$, $c_{var} = 0.25$, $c_{unit} = 0.4$. We take the average of 500 simulations where we used 20 time periods, 100 units for each simulation.

C Appendix

C.1 Summary of resuts (Di Tella & Schargrodsky, 2004)

Table 6: The Effect of Police Presence on Car Theft

	Original model specification	Modified version	Regression exclusion test
	(i)	(ii)	(iii)
(Intercept)		0.1104***	0.1104***
		(0.0083)	(0.0083)
Same Block	−0.0898**	−0.0870**	−0.0962**
	(0.0306)	(0.0302)	(0.0313)
One Block Away	−0.0082		
	(0.0163)		
Two Blocks Away	−0.0045		
	(0.0137)		
Same Block (August)			0.0152
			(0.0319)
Same Block (September)			−0.0064
			(0.0182)
Same Block (October)			0.0317
			(0.0402)
Same Block (November)			0.0054
			(0.0202)
April	0.1104		
	(0.0083)		
May	0.099	−0.0114	−0.0114
	(0.0068)	(0.0113)	(0.0113)
June	0.0782	−0.0322	−0.0322
	(0.0064)	(0.0113)	(0.0113)
August	0.1146	0.0014	0.0011
	(0.0096)	(0.0122)	(0.0124)
September	0.1015	−0.0117	−0.0111
	(0.0096)	(0.012)	(0.0122)
October	0.1124	−0.0009	−0.0018
	(0.0093)	(0.0123)	(0.0124)
November	0.1038	−0.0095	−0.0093
	(0.009)	(0.0122)	(0.0123)
December	0.1089	−0.0043	−0.0039
	(0.0095)	(0.0126)	(0.0128)
Month Fixed Effects	Yes	Yes	Yes
Block Fixed Effects	Yes	Yes	Yes
R^2	0.0034	0.0033	0.0034
Num. entities	876	876	876
Time periods	8	8	8

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: Ellaborated with data from Di Tella & Schargrodsky (2004); Hansen (2022). The models were estimated using the *linearmodels* package for Python. Results correspond to the Stata replication file in Hansen (2022). July data is dropped for the estimations as it is incomplete; however, the results are robust to the alternative setup.

C.2 Descriptive statistics (Di Tella & Schargrodsky, 2004)

Table 7: Average Car Thefts per Month in City Blocks in Buenos Aires

	with a Jewish Institution	Without a Jewish Institution
April	0.122	0.110
May	0.088	0.100
June	0.128	0.076
August	0.047	0.111
September	0.014	0.099
October	0.061	0.108
November	0.027	0.100
December	0.027	0.106
Before the attack	0.113	0.095
After the attack	0.032	0.103

Note: Ellaborated with data from Di Tella & Schargrodsky (2004); Hansen (2022).

D Appendix

D.1 K-factor TWFEr

$$K = \frac{1}{NT}\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T} X_{i,t}\left(\frac{\sum_{t'\neq t}(1-X_{it'})}{T-1} + \frac{\sum_{i'\neq i}(1-X_{i',t})}{N-1} - \frac{\sum_{i'\neq i}\sum_{t'\neq t}(1-X_{i',t'})}{(T-1)(N-1)}\right) \\ +(1-X_{i,t})\left(\frac{\sum_{t'\neq t}(X_{it'})}{T-1} + \frac{\sum_{i'\neq i}(X_{i',t})}{N-1} - \frac{\sum_{i'\neq i}\sum_{t'\neq t}(X_{i',t'})}{(T-1)(N-1)}\right) \tag{27}$$

D.2 Carry-over effect

The carry-over effect is an effect that is manifested if a treated unit in time t stays in treatment in t+1, however this effect will be lost if the unit leaves the treatment.

$$Y_{i,t} = u_i + v_t + \theta X_{i,t} + carry\_over Z_{i,t} + e_{i,t}$$

The model is the same as (19) with the addition of the carry over effect explained below.

$$t = 1 \rightarrow Z_{i,t} = 0$$

$$t \neq 1 \rightarrow Z_{i,t} = |\{X_{i,t}|X_{i,t} = 1, X_{i,t-1} = 1\}|$$

The variable $Z_{i,t}$ is equal to one if a treated observations was treated in the period before, in any other case the variable will be zero. When this variable is equal to one a "carry over" effect is manifested, a premium for staying in treatment, this effect disappear once the treatment is left. While carry over effect itself has not been object of your analysis this is a recurrent case of model misspecification and it is useful to show how the different estimators reacts, an example is given.

$$X_{i,t} = 1, Z_{i,t} = 1, X_{i,t+1} = 0, Z_{i,t+1} = 0 \tag{28}$$

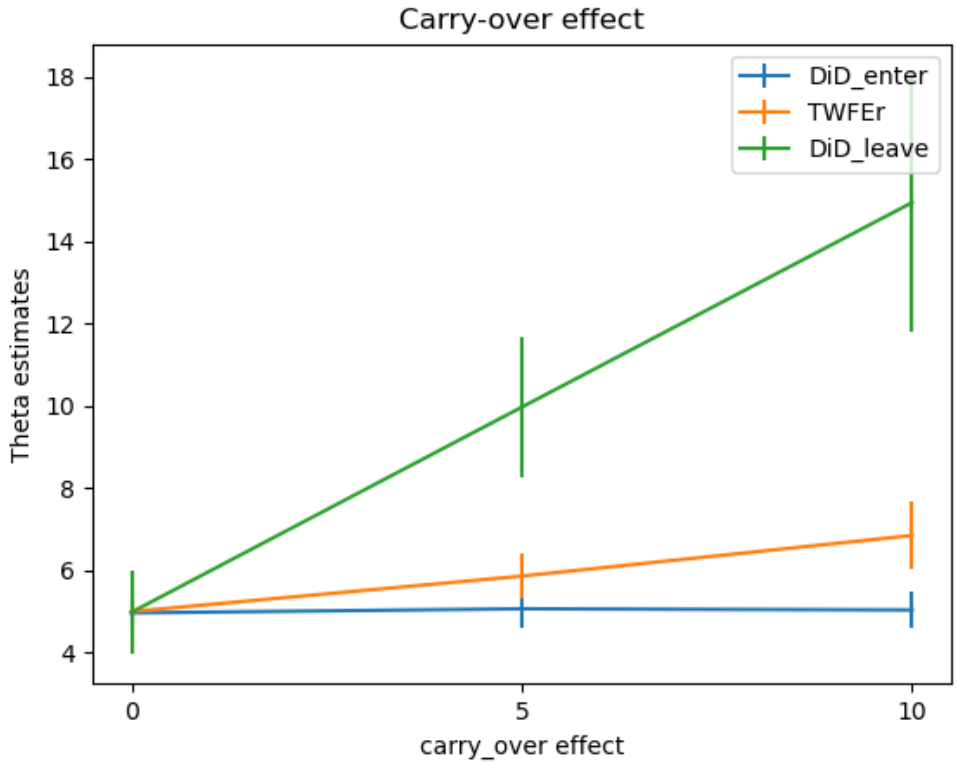
$$Y_{i,t} - Y_{i,t+1} = v_t - v_{t+1} + carry\_over + \theta \tag{29}$$

The estimates of the TWFE will be affected since the treated units are on average larger (if the carry over effect is positive, if not it would be smaller) than what it is in the estimator model (19), however the magnitude of the effect depends on how many units are recipients of the carry over effect. It is clear that this effect will not affect the ATT_{enter}

since the counterfactual is not changing compared to the standard setting. However, it will affect the counterfactual of units leaving the treatment (they are compared with units that stayed there) ATT_{leave} will be biased. With the increase of the carry-over effect also the variance of ATT_{leave} increases.By looking at what an observation and its counterfactual are it will be clearer:

$$\widehat{\theta_{leave}} = (Y_{c,t+1} - Y_{c,t}) - (Y_{i,t+1} - Y_{i,t})$$

The first difference belongs to $0, carry\_over$ while the second to $-\theta, -(\theta + carry\_over)$, the ATT_{leave} can range between $[\min\{carry\_over, \theta\}, 2carry\_over + \theta]$, this ranges only become wider with an increase in $carry\_over$,hence more variant.



It is possible to notice that with a $carry\_over$ effect of 5, the TWFE is already significantly different than the treatment effect value of 5.