econometrics

IMPACT FACTOR 1.4

CITESCORE 2.2

Review

Vis Inertiae and Statistical Inference: A Review of Difference-in-Differences Methods Employed in Economics and Other Subjects

Bruno Paolo Bosco and Paolo Maranzano

MDPI

https://doi.org/10.3390/econometrics13040038

econometrics

MDPI

Review

Vis Inertiae and Statistical Inference: A Review of Difference-in-Differences Methods Employed in Economics and Other Subjects

Bruno Paolo Bosco [1,*] and Paolo Maranzano [1,2]

[1] Department of Economics, Management and Statistics (DEMS), University of Milan-Bicocca, Piazza Ateneo Nuovo n.1, 20126 Milan, Italy; paolo.maranzano@unimib.it or paolo.maranzano@feem.it
[2] Fondazione Eni Enrico Mattei (FEEM), Corso Magenta n.63, 20123 Milan, Italy
* Correspondence: bruno.bosco@unimib.it

Abstract

Difference in Differences (DiD) is a useful statistical technique employed by researchers to estimate the effects of exogenous events on the outcome of some response variables in random samples of treated units (i.e., units exposed to the event) ideally drawn from an infinite population. The term "effect" should be understood as the discrepancy between the post-event realisation of the response and the hypothetical realisation of that same outcome for the same treated units in the absence of the event. This theoretical discrepancy is clearly unobservable. To circumvent the implicit missing variable problem, DiD methods utilise the realisations of the response variable observed in comparable random samples of untreated units. The latter are samples of units drawn from the same population, but they are not exposed to the event under investigation. They function as the control or comparison group and serve as proxies for the non-existent untreated realisations of the responses in treated units during post-treatment periods. In summary, the DiD model posits that, in the absence of intervention and under specific conditions, treated units would exhibit behaviours that are indistinguishable from those of control or untreated units during the post-treatment periods. For the purpose of estimation, the method employs a combination of before–after and treatment–control group comparisons. The event that affects the response variables is referred to as "treatment." However, it could also be referred to as "causal factor" to emphasise that, in the DiD approach, the objective is not to estimate a mere statistical association among variables. This review introduces the DiD techniques for researchers in economics, public policy, health research, management, environmental analysis, and other fields. It commences with the rudimentary methods employed to estimate the so-called Average Treatment Effect upon Treated (ATET) in a two-period and two-group case and subsequently addresses numerous issues that arise in a multi-unit and multi-period context. A particular focus is placed on the statistical assumptions necessary for a precise delineation of the identification process of the cause–effect relationship in the multi-period case. These assumptions include the parallel trend hypothesis, the no-anticipation assumption, and the SUTVA assumption. In the multi-period case, both the homogeneous and heterogeneous scenarios are taken into consideration. The homogeneous scenario refers to the situation in which the treated units are initially treated in the same periods. In contrast, the heterogeneous scenario involves the treatment of treated units in different periods. A portion of the presentation will be allocated to the developments associated with the DiD techniques that can be employed in the context of data clustering or spatio-temporal dependence. The present review includes a concise exposition of some policy-oriented papers that incorporate applications of DiD. The areas of focus encompass income taxation, migration, regulation, and environmental management.

check for updates

Academic Editor: Guglielmo Maria Caporale

Received: 11 August 2025
Revised: 19 September 2025
Accepted: 24 September 2025
Published: 30 September 2025

Citation: Bosco, B. P., & Maranzano, P. (2025). *Vis Inertiae* and Statistical Inference: A Review of Difference-in-Differences Methods Employed in Economics and Other Subjects. *Econometrics*, *13*(4), 38. https://doi.org/10.3390/econometrics13040038

Keywords: difference-in-differences (DID); review for causal inference; applied and empirical economics; treatment and control; extensions of the DiD estimator to heterogenous treatment framework

JEL Classification: C23 (single equation panel data models and spatial-temporal models); C50 (general econometric modelling); C54 (quantitative policy modelling); D04 (microeconomic policy: formulation; implementation; and evaluation); E6 (macroeconomic policy; macroeconomic aspects of public finance)

1. Introduction to DiD

With a Difference-in-Differences (DiD) analysis, we try to estimate whether a response variable (i.e., a variable exposed to a *treatment*) will achieve a mean value that, computed for the set of all treated units (treated group), is statistically different than the mean value computed for the set of some comparable untreated units (untreated or control group), once any factors affecting the link between the treatment and the effect (*confounders*) are ruled out. Therefore, the DiD analysis aims at "discovering" if a time-contingent cause–effect relationship (a post hoc, ergo propter hoc relationship) between the response variable and the treatment is statistically consistent with the data. Therefore, with DiD, we cultivate the ambition of evaluating whether a precise causative link between causes and effects—defined according to a model based on a proper identification of the relationship among variables—is consistent with the data and estimate how intensive and statistically robust the cause–effect link is.

DiD analysis is widely employed in economics, public policy, health research, management, environment analysis, and other fields. Examples of DiD analysis of changes in response variables after a treatment are numerous and encompass various research fields. They may be the human mortality rate, unemployment, or the quantity of corn harvested, to mention a few. The treatment or causal factor may include the use of a new pharmaceutical drug, the implementation of new training program for unemployed workers, the application of a new agricultural technique, etc. Other known examples of a response variable are SAT (*Site Acceptance Test*) scores of equipment under quality check, the level of pollution in a county before and after the adoption of environmental measures, or the tree cover density in a region subjected to reforestation. Owing to its flexibility, DiD has been widely used in economics, public policy, health research, management and numerous other fields.

To conduct the aforementioned causal analysis, DiD relies on a combination of before–after and treated–untreated group comparisons. However, it is essential to emphasise from the beginning that the variable corresponding to treatment must be expressed as a dichotomous variable (zero vs. one; yes vs. no) and not as a continuous variable. As will be elucidated subsequently, treatment variables will assume a role analogous to that of dummy variables in more conventional regression analysis. As would be expected, the term "treatment dummy" is frequently utilised in DiD studies and in studies that are structured as DiD models. The term dummy, for example, appears already in Ashenfelter and Card's (1985) study of the effect of some training program in the USA where the longitudinal structure of earnings of trainees and group comparisons are used to estimate the effectiveness of the program for participants. In their study of the relationship between casual factor and response variable, Card and Krueger (1994) provide another early clear illustration of how the treatment is included in the analysis. They investigated whether an increase in minimum wage by New Jersey in 1992 from USD 4.25 to USD 5.05 (treatment) resulted in a

statistically significant change in employment level amongst fast food restaurant workers in New Jersey (treated units) from that in neighbouring Pennsylvania, which did not change its minimum wage (untreated units). The treatment was not the *amount* of the minimum wage increase (a possible continuous variable) but the "mere" asymmetric implementation (e.g., Yes in New Jersey and No in Pennsylvania) of the policy measure.

There is a discussion about the true "fatherhood" of the DiD method and, not surprisingly, there are clear pioneering antecedents of DiD applications outside economics. Examples include medicine (the study of the causes of London' worst cholera epidemics of 1849 with 14,137 victims) and agriculture (studies of changes in soil productivity enhanced by new cultivation techniques in neighbouring areas of Africa in the 1980s conducted by revolutionary governments after the victory of their anti-colonialist movements in the second half of the 1970s). A recognised common methodological basis is R. A. Fisher's analysis of variance (ANOVA).

In general, the DiD analysis aims at estimating the mean difference between actual and *potential realisations* of a response variable in treated units. The latter can be defined as the unobserved (and unobservable) realisations in the treated units, and after the occurrence of an asymmetric exogenous event, of the response variable *as if the event/treatment never happened*. Therefore, the ultimate goal of DiD is the estimation of the "deviation" of the actual realisations of the response variable in the post-event period from its potential *event-absent hypothetical* values. That deviation will be interpreted as the effect of the event/treatment. Yet, since the above potential realisations do not exist, DiD uses instrumentally the actual response realisations recorded in untreated units provided that, as we shall see in the next sections, some conditions are satisfied.

Notice that the above-mentioned modification in mean differences generally occurs over time, and the passage of time represents a complicating challenge in a DiD study. Whatever the response variable we study, the passage of time may affect in a potentially significant way the actual realisation of the response as the data generation process proceeds through a possibly long-time span and encompasses several pre-treatment and post-treatment periods. Hence, the specific effect attributable to the passage of time (and the numerous factors potentially concealed in the passage of time) on the mean value of the response variable in both the control group and the treatment group must be properly considered. In other words, the researcher must determine if it was the treatment itself the cause of any change in the mean value of the response variable within the treatment group *over and above* what was caused by the pure passage of time or by time-conditioned factors. In summary, the main ingredients of the DiD approach to the estimation of the cause–effect relationship are the existence of a treatment administered to treated units only, the behaviour of the treated and untreated response variables before and after the moment the treatment was implemented, and an appropriate consideration for the passage of time.

In this review, we present the DiD estimation approach to the cause–effect relationship. We will try to highlight how the effect of the treatment can be estimated separately from the effect of the passage of time with the DiD method. To do so, in Section 2, we first present the simplest DiD framework in which the treatment status of each unit can vary over time according to the following dynamics: an initial time period (e.g., months or years) in which there is no treatment is followed by a time period with treatment administered to some units only. The moment in which the treatment is introduced represents the temporal turning point of the entire period under examination. The units under investigation can, in turn, be assigned to two groups: those classified as *never treated* (the control group) because they are never subjected to the treatment during the entire sample period and those units that are *treated in the post-intervention period only* (the treated group). We will assume that the latter are uninterruptedly treated from the introduction of the treatment until the end of the

observed periods (staggered variables case). In the initial simplest DiD framework, we will assume that the treatment is the only relevant independent variable affecting the outcome of the response dependent variable. Then, in Section 3, we discuss the OLS way to estimate the effect of the treatment, as well as the identification problems related to the estimation process. We call this initial framework a *homogeneous* case with staggered variables but no cofactors: the same units are treated during the same periods and uninterruptedly until the end of the observational period. Homogeneity means that all the treated units will start to receive the treatment in the same moment and since they are staggered, they terminate the treatment in the same moment. The presence of cofactors under the homogeneous case is analysed in Section 7. Analogously, the model structure in which treatments are administered in different periods to different treated units and never administered to some other units is considered later in Section 8 and it is termed a *heterogeneous* case (with or without cofactors): treated units will start to receive the treatment in different moments but since they are staggered, they terminate the treatment in the same moment at the end of the observed sample period. In this review we will assume that units are always staggered, whenever they initiated with the treatment. DiD techniques to be used under a more complicated data structure (e.g., data generating clustering phenomenon or spatial-temporal relations) are introduced in Section 9. Eventually, in Section 10, we provide a summary of the main topics discussed in this review, and in Section 11, we survey some applications taken from the literature discussing the specific empirical settings employed and the main results.

A particular aspect of DiD on which we decided to focus is the exogeneity character of the treatment and the so-called parallel trend assumption (see Sections 3.1 and 3.2). They represent fundamental elements of the method. As some of the papers discussed at the end of this review will clearly show, in many cases, DiD represents the statistical approach needed to overcome the simultaneity and endogeneity difficulties inherent in many circumstances in more traditional OLS estimation techniques. Yet, this advantage of DiD over alternative techniques requires that some crucial assumptions about the data generation process are satisfied.

This review also briefly surveys other aspects of the DiD methods that, in our opinion, are not sufficiently considered by the DiD literature. For instance, Section 2.2 discusses the issue represented by the hypothesis of no inter-unit interference (a unit's outcome must be unaffected by the treatment status of other units). The literature defines this condition as Stable Unit Treatment Value Assumption (SUTVA) and shows why the DiD identification process requires that the treatment applied to some units should not affect the outcome for other units. In other words, the potential outcome of a generic unit in the analysed sample should not depend on the treatment status of some other units in the same sample or on the mechanism by which units are assigned to the control or treatment groups. We also pay special attention to the role that confounding factors have in DiD (Section 6).

In Appendices A and B of this review, we include a few ad hoc datasets made by the authors to be used as examples of the DiD estimation techniques analysed in this review and to conduct exercises.[1]

Although this simple review is conceived for applied economists, readers should keep in mind that DiD's most attractive features are its (relative) simplicity and wide applicability. After all, to carry out a basic DiD study, we just require observations from a treated group and an untreated (comparison) group both before and after the intervention is enacted. Accordingly, in the last sections, we discuss some papers that have applied DiD techniques in various research areas relevant in a public economics or public policy perspective.[2]

We stress that this review covers the basic (almost intuitive) DiD techniques. There are other, more advanced reviews (Callaway, 2022; Roth et al., 2023; Baker et al., 2025 to mention three important papers), as well as Chapter 5 of Angrist and Pischke (2009) and Chapter 21 of Wooldridge (2010) that should be consulted by more advanced users.

2. Fundamentals of the Two-Group and Two-Period Homogenous DiD

This review surveys the basic methods and recent developments introduced in the DiD literature in the last 30 years. Clearly, the fundamental notions of DiD could be assumed to be almost common knowledge, and, in theory, they should not require a new basic review to be added to the many that already exist. Yet, since we want to offer a (maybe incomplete, but) self-contained treatment of the subject, we start with the basic framework need to identify a DiD model.

Assume that we have randomly drawn from an infinite population two samples of individuals (with or without the same numerousness of units), denoted as G1 and G2. We call i an individual belonging to G1 and j an individual belonging to G2. Assume that the two periods under study are two years, each divided for expositional convenience into 12 months. We observe in each month of the first of the two years under study, the realisation of a random variable y, representing the relevant variable under our investigation (income, unemployment, indebtedness, hours of work, rate of financial criminality, level of fever, etc.). For reasons that will become clear very shortly, we call y the response variable. If y_{it} is the realisation of y for an individual i in G1 recorded during each month $t = (1, \ldots, 24)$, then $Y_{it\in(1,\ \ldots,24)} = N^{-1}\sum_{i=1}^{N} y_{it}$ is the month t mean value of y generated by data of individuals belonging to G1, with N representing the total number of individuals in G1. Correspondingly, $Y_{jt\in(1,\ldots,24)} = M^{-1}\sum_{j=1}^{M} y_{it}$ is the month t mean value for group G2, generated by all j individuals of that group formed by M individuals. As a result, for each year, we record 12 mean monthly values for each group. Altogether, we have 48 mean observations in the 2-group $\times$ 2-year dataset.

We assume that the monthly evolution of Y_{it} and Y_{jt} during the first 12 months is linearly parallel. In other words, we assume that the time evolution of the two series of mean values follows a parallel path with the same time slope so that the two paths are separated only by a group-specific constant (a sort of individual fixed effect used in the fixed-effect least squares with dummy variables panel data analysis). Then, the plot of the time behaviour of the 24 mean values during the first year (first 12 months) corresponds to the left portion of Figure 1 reproduced below (the first 12 months to the left of the vertical line).

Assume now that at the end of the first year (i.e., in correspondence to the vertical line in Figure 1), "something" affecting only G2 happened, ceteris paribus. That *something* is generally assumed to correspond to an exogenous event and it is called treatment (e.g., a new regulation, a more or less exogenous change in tax rates implemented in G2, some new subsidies paid to firms of that group, higher interest rates, a natural event, a new pharmaceutical therapy, etc.). This way of introducing treatments should make clear that the word "period" used in this DiD review is not synonymous with a calendar unit of time but of "temporal phase". In the 2 $\times$ 2 case, we have two periods/phases: the first one (lasting 12 months) with no event and the second (lasting 12 months) with an event affecting the units of a group (G2 in our example) right from the *arrival* of the event and continuously *until the end* of our sample time.

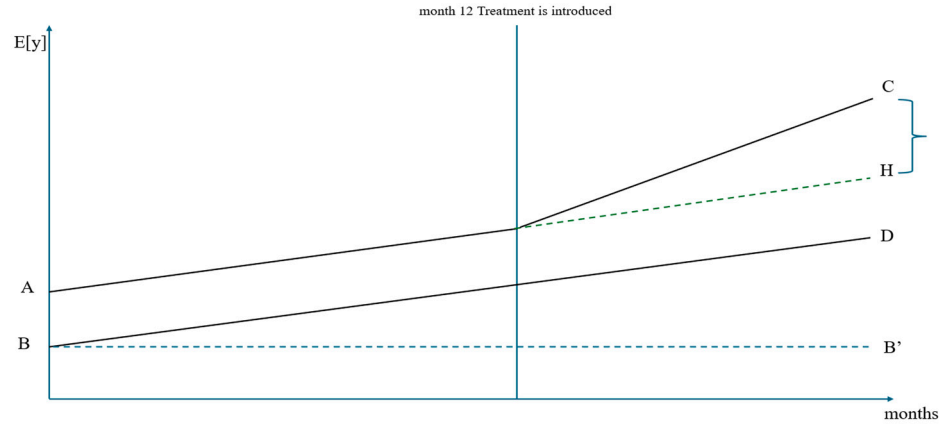


Figure 1. *Effects of the treatment on E*[*y*] *in treated units. Notes* The two solid lines are the mean values of the response *y*. Line *BD* refers to the control units (not subjected to treatment) and the upper broken line refers to the treated units. The vertical line indicates when the treatment was introduced. *BB'* is drawn for pure graphical reference. *CH* is the mean effect of the treatment. By observing the broken line alone, one cannot be sure that the new path is due to the treatment or to something else. It is the comparison with the control group that *might* give us the *perception* (possibly just the optical illusion) that the treatment can be the cause of the change in the broken line path.

We assume that individuals in G2 cannot anticipate the introduction of the treatment (and therefore cannot react in advance to its occurrence). Now, the right part of Figure 1 (the part to the right of the vertical line) becomes relevant. Inspection shows the time path of the expected values of y for the treated group G2, which, after the treatment, can be specified as follows:

$$E\left[y_j \middle| Treatment = 1\right]$$
$$\equiv \text{Expected value of } y_j \text{ in G2 conditional upon the realisation of the event}$$

where *Treatment* = 1 means that the treatment is operative.

The G2 path has been twisted upward about the point in the plot corresponding to the last month of the first year (i.e., the last pre-treatment or pre-event month), while the time path of G1 proceeds according to the previous linear trend, and it is as follows:

$$E\left[y_j \middle| Treatment = 0\right]$$
$$\equiv \text{Expected value of } y_j \text{ in G1 conditional upon the absence of the event}$$

where *Treatment* = 0 means that the treatment is not operative.

Hence, we assume that $E[y]$ in G1 is unaffected by the treatment that is implemented with respect to G2 units only, and that the treatment affecting G2 has no spillover effects on G1, i.e., that the treatment affects G2 only and has no unplanned impacts on the untreated units in G1 (no *externalities*, to use current microeconomics parlance). Then, the dashed line represents the possible realisations of the expected values of the mean values of y for G2 in the absence of treatment but under the linear parallel trend hypothesis discussed above. In other words, the dashed line indicates what the path of the expected realisations of y in G2 would have been expected *were the "perturbing" event* (*the treatment*) *absent*, as if the Galilean *inertia principle* for uniform linear motion of corps was at work (no intervening external forces). Clearly, these realisations are not observed: actually, *they do not exist*. For that reason, we name them "potential realizations" as if they were the effect resulting from the application of a "*vis inertiae*, or force of inactivity" to use the terms employed by Newton in his *Philosophiae Naturalis Principia Mathematica* of 1687 because their force depends only upon their position (potential energy).

We now have all the ingredients useful to measure the average effect that the treatment had on G2 (the treated group). The mean effect of the treatment is the difference between the value assumed by the mean y in G2 after the treatment (solid line) and the value that it would have potentially assumed by pure *vis inertiae* in the absence of the treatment (dashed line). That effect corresponds to the segment CH in Figure 1 whose length is the difference between the abscissa of point C and the abscissas of point H.

The measure of the length of the segment $CH = C - H$ can be recovered as follows:

$$CH \equiv Average\ Effect\ of\ the\ Treatment\ upon\ Treated$$

$$= \left(C - B'\right) - (H - D) - \left(D - B'\right)$$

$$= (C - B) - (A - B) - (D - B)$$

$$= (C - A) - (D - B)$$

By manipulating the last equation, we express CH as a difference between two differences, namely:

$$Average\ Effect\ of\ the\ Treatment\ upon\ Treated = (C - D) - (A - B)$$

In the latter version, the measure of the treatment effect corresponds to the difference between two terms. The first term (included in the first parenthesis) measures the difference in the realisations of the expected value of y for both groups (treated and untreated) in the post-treatment period, i.e., when *Treatment* = 1 in the above expected values formulas. The second term (included in the second parenthesis) measures the difference between the initial intercepts for the two groups, i.e., when *Treatment* = 0. The latter corresponds to the constant vertical distance of the two lines in the pre-treatment period and (under the hypothesis that the time trends are initially parallel and would have remained parallel in the absence of treatment). In other words, it is assumed that the constant initial difference (segment AB) is constant during the entire period 1 because it depends only upon the above-mentioned idiosyncratic constant elements and then, because of the Galilean inertia principle, it is bound to remain constant in the absence of treatment (the only new intervening force), even in period 2. This motivates the plot dashed line in period 2 in Figure 1 and the previous description of those (unobserved) values as "potential".

Therefore, a measure of the treatment effect, defined as the *Average Effect of the Treatment upon Treated* (*ATET* from now on), is obtained by differencing, on the one hand, the mean response for the treatment and control units over time to eliminate time-invariant unobserved characteristics and, on the other hand, by differencing the mean response of the groups (treated and untreated) to eliminate time-varying unobserved effects common to both groups. In other words, the DiD technique eliminates the influx of time-varying factors (confounders) by comparing the treatment group with a control group that is subjected to the same time-varying factors (confounders) as the treatment-receiving group.

For example, we may consider y to be the employment rate, and the treatment to be a subsidy paid exclusively to firms in G2 (e.g., a specific region of the country) for every new employee. If the expected unemployment in G1 and G2 follows a parallel trend in period 1 (when no subsidy was paid to firms), expected unemployment in G2 should adhere to the linear trend of period 1 and remain parallel to that of G1. The dashed line indicates the potential expected value of unemployment in G2 in the absence of a subsidy for firms in G2 during period 2.

Figure 1 shows that the untreated group G1 has a role of paramount importance in the measurement procedure depicted above. G1 (untreated) acts as a control group and

supplies (loosely speaking) the substitutes for the unobservable (because they are never realised) counterfactual observations of G2 to be used when studying the effect of the treatment. To be specific, the hypothesis is that in the absence of treatment, reality would have evolved in G2 as described by the $E[y]$ recorded in G1, with the obvious consequence that the right curly bracket in Figure 1 would not exist because C $\rightarrow$ H. In other words, it would be $E\left[y_j \middle| Treatment = 1\right] = E\left[y_j \middle| Treatment = 0\right]$.

We can now proceed to estimate the effect of the event using all the expected values as follows, recalling that in Figure 1 each point can be interpreted as follows:

- C is the expected value of y for the treated group conditional upon the application of the treatment on that group;
- D is the expected value of y for the untreated group conditional upon the absence of the treatment for that group;
- A is the expected value of y for the treated group conditional upon the absence of the treatment;
- B is the expected value of y for the untreated group conditional upon the absence of the treatment.

Then, after calling $h = (i, j)$ a generic individual in the *population* (either treated or untreated, i.e., G1 + G2), we may write a linear regression model as follows:

$$y_{ht} = \beta_0 + \beta_1 \times D1 + \beta_2 \times D2 + \beta_3 \times [D1 \times D2] + \varepsilon_{ht} \quad (1)$$

where:

- y_{ht} is the value of the response variable for a unit in the population under study. Its value is measured in each group and each t, i.e., before and after the introduction of the treatment. It will correspond either to the i-th or to j-th observation at time t depending on the group (treated or untreated) of the unit;
- β_0 is the intercept of the regression model, common to treated and untreated units;
- $D1$ is the *Time Period Dummy*, that is, a dummy variable that takes the value 0 or 1 depending on whether the h-th observation of the response variable refers to the pre- ($D1 = 0$) or post-treatment period ($D1 = 1$) independently of the group (treated or control) the observation belongs to. It simply indicates whether the treatment existed or not in that period t (independently of which unit was treated);
- $D2$ is the *Treatment Indicator Dummy*, that is, a dummy variable that takes the value 0 or 1 depending on whether the h-th observation refers to an individual in the control group (untreated) or in the treatment group, respectively, independently of the specific time period t. Therefore, $D2 = 0$ when the observation belongs to an untreated unit and $D2 = 1$ when the observation belongs to a treated unit (independently of when the treatment was introduced). Clearly, in the simplified example of this section with only two periods, $D2 = 0$ means that the unit is never treated. More complex settings are discussed in the subsequent sections;
- $D1 \times D2$ is the interaction term between the time dummy and the treatment dummy. It is the most important coefficient to estimate as it measures the average effect of treatment on treated units.

As will be explored below, the above Equation (1) represents the basic but elegant form of a DiD analysis for the homogenous case with no cofactors. In Table 1, we analytically discuss the relevance of each coefficient. Here, we stress that the *elegance* of DiD (Goodman-Bacon, 2021, p. 254) makes it clear which comparisons generate the estimates, what leads to bias, and how to test the design. The expression in terms of sample means connects the regression to potential outcomes and shows that, under a parallel trend assumption, a two-group/two-period (2 $\times$ 2) DiD identifies the average treatment effect on the treated.

Table 1. Combinations of periods and treatment in a 2 × 2 DiD design.

	***D*1 = 0**	***D*1 = 1**
***D*2 = 0**	$y_{ht} = \beta_0 + \varepsilon_t$	$y_{ht} = \beta_0 + \beta_1 + \varepsilon_t$
***D*2 = 1**	$y_{ht} = \beta_0 + \beta_2 + \varepsilon_t$	$y_{ht} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \varepsilon_t$

The estimated coefficients of Equation (1) have definite relations with the critical points of Figure 1. These relations are illustrated in the following 2 × 2 Table 1, which gives a more explicit description of how the states of the world (time and treatment) can combine and how they affect the realisation of y in the above Equation (1).

In what follows, the estimated coefficients obtained from an OLS regression of the model (1) correspond to the expected values presented above. For the fitted model, the corresponding expectations are as follows.

We use the caps (ˆ) above the coefficients to indicate that they are the estimated (fitted) values of the corresponding coefficients. Replacing y_{ht} with the expected value of y_{ht} also allows us to drop the error term ε_t since by hypotheses in a well-behaved OLS regression model, the expected value of the error term is a zero mean and constant variance term. Hence, we can rewrite the content of each cell of the 2 × 2 matrix of Table 1 as follows.

The northwest cell is as follows:

$$E[y_{ht}|D1 = 0, D2 = 0] = \hat{\beta}_0$$

In terms of the hypothetical dataset generating Figure 1, $\hat{\beta}_0$ corresponds to point B and must be interpreted as the average baseline intercept common to the two groups (a constant).

The northeast cell is as follows:

$$E[y_{ht}|D1 = 1, D2 = 0] = \hat{\beta}_0 + \hat{\beta}_1$$

In terms of the hypothetical data generating Figure 1, $\hat{\beta}_0$ still corresponds to point B and, as above, it must be interpreted as the model baseline average (constant). $\hat{\beta}_1$, which corresponds to the slope of segment DB, is the time trend coefficient in the control group (constant before and after the introduction of the treatment).

The southwest cell is as follows:

$$E[y_{ht}|D1 = 0, D2 = 1] = \hat{\beta}_0 + \hat{\beta}_2$$

In terms of the hypothetical data generating Figure 1, the following is observed:

(i) $\hat{\beta}_0$ corresponds to point B, as above, and must be interpreted as the model baseline average (constant);

(ii) $\hat{\beta}_2$, which corresponds to segment AB, is the constant difference between the two groups before the treatment.

The southeast cell is as follows:

$$E[y_{ht}|D1 = 1, D2 = 1] = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$$

In terms of the hypothetical data generating Figure 1, the sum $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$ corresponds to point C.

We now proceed to calculate the difference in the expected value of y between the before- (pre-) and after (post-)-treatment phases of this study. For the treatment group, the difference in expectations works out as follows:

$$E[y_{ht}|D1 = 1, D2 = 1] - E[y_{ht}|D1 = 0, D2 = 1] = (\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3) - (\hat{\beta}_0 + \hat{\beta}_2) = \hat{\beta}_1 + \hat{\beta}_3$$

which is the difference in estimated response between the after-treatment and before-treatment phases of this study recorded within the treatment group. Similarly, for the control group, we have:

$$E[y_{ht}|D1 = 1, D2 = 0] - E[y_{ht}|D1 = 0, D2 = 0] = (\hat{\beta}_0 + \hat{\beta}_1) - (\hat{\beta}_0) = \hat{\beta}_1$$

The above is the difference in estimated response within the control group between the after-treatment and before-treatment phases of this study.

The difference between the two differences measures the average net effect of the treatment on the treated group, as follows:

$$E[DiD\ Effect] = (\hat{\beta}_1 + \hat{\beta}_3) - (\hat{\beta}_1) = \hat{\beta}_3$$

The estimated coefficient $\hat{\beta}_3$ is what we called the *ATET* (*Average Treatment Effect upon Treated*) and it was obtained from the estimates of a linear model in which there are no cofactors (other independent variables affecting y). Its difference with respect to the similar measure of the treatment called *ATE* is discussed later.

For a more analytically grounded derivation of the estimated average effect of the treatment, one may consult Angrist and Pischke (2009, p. 229), who discuss the expected DiD effect and then show the OLS regression that may be used for its estimation. Following the opposite route, Wooldridge (2010, p. 148) first starts with an OLS regression equation augmented with treatment dummies and then expresses and interprets the estimated relevant dummy as an estimate of the expected DiD treatment effect.

The following remarks summarise the conditions of the application of OLS as part of a DiD strategy.

Remark 1. *The basic ingredient: randomness.*

We start with a set of i.i.d. individuals $i = 1, \ldots, n$ and a tuple (Y_i, X_i, D_i), where $Y_i \in \mathbb{R}$ is the response variable, $X_i \in \mathbb{R}$ is a cofactor vector (it may not be present) and $D_i \in \{0, 1\}$ is the treatment assignment. We assume that the *outcome* (i.e., the realisation of the response variable) depends on existence/absence of the treatment, and is defined as follows:

$$Y_i = Y_i(D_i) = \begin{cases} Y_i(0) \equiv \textit{The response we had observed if } D_i = 0 \\ \quad (\textit{potential unobserved outcome}) \\ \\ Y_i(1) \equiv \textit{The response we actually observe with } D_i = 1 \\ \quad (\textit{realized actual outcome}) \end{cases}$$

We define the causal effect of the treatment of the sample of treated units as $Y_i(1) - Y_i(0)$, i.e., as the difference between the potential and actual outcomes of each individual i, so that *on average* (i.e., with respect to the population) we determine that the Average Treatment Effect upon Treated (*ATET*) is $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$.

It is worth stressing again that each realisation of the response can be observed in just one state of the world, i.e., conditional on either $D_i = 0$ or $D_i = 1$, not both. Under the assumption that the treatment assignment is random (there is no systematic association between the potential outcome of an individual and the condition of being treated), OLS methods can help us to overcome this missing data problem.

Remark 2. *The basic ingredients: ATET and OLS estimator.*

The OLS method of estimation of Equation (1) correctly identifies the ATET in a DiD regression under the following conditions:

1. Parallel trend (the response variable for treated and untreated units follows the same time path);
2. No anticipation effects (treated units cannot adjust their behaviour on the basis of information on an incoming treatment);
3. No spillover impacts of the treatment (the treatment does not have an impact outside the treated units).

OLS allows us to define the estimate, which involves unobservable counterfactuals in a form (Equation (1)) that depends only on observed outcomes. This process is known as the "identification process."

Then, ATET is the expected value of the DiD effect between the treatment and control group (i.e., *CH* in Figure 1). After the DiD model is estimated, the estimated coefficient of the interaction term ($D1 \times D2$), i.e., $\hat{\beta}_3$, will give us the estimated difference-in-differences effect of the treatment that we are seeking. The coefficient's *t*-score and corresponding *p*-value will tell us whether the effect is statistically significant, and if so, we can construct the 95% or 99% confidence intervals around the estimated coefficient using the coefficient's standard error reported by the model output.

Finally, recall that we have randomly selected the participants (treated) and the non-participants (untreated). Therefore, carrying out a DiD study, we do not ask ourselves the following question: why did the non-participants not participate? To some extent, this question is outside the realm of DiD analysis. The general DiD assumption is that there is a sort of powerful external force determining, in a random way, the correct random sampling.[3]

In applied research, the issue of no anticipation effect might pass unnoticed, and for that reason we want to stress it a bit more. Suppose that one wants to look at the impact on youth employment of a policy of minimum wage change adopted by local authorities. The concern is that local authorities may increase minimum wages during good times, so that labour demand will cause the trajectory of youth employment to differ between treated and control areas. However, this effect could be autonomous and unrelated to the effect of the minimum wage policy.

2.1. Violations of the Parallel Trend Assumption

As indicated by Remark 1 and Figure 1, the parallel trend assumption constitutes a critical component of DiD. Violations of the parallel trend assumption in DiD analysis implies that pre-treatment response trends are different for the treatment and control groups and would not show a similar potential parallel path in the treatment period in the absence of treatment. This unfortunate but likely condition would lead to biased estimates of the causative treatment effect because the estimated treatment coefficient will correspond to the estimate of the true treatment effect plus a measure of the different pre-existing trends. In the case of a simple 2-period and 2 group homogenous staggered case of Section 2, the above condition can be generated by time-varying unobserved confounding factors that produce selective treatment responses to common shocks for the two groups. At the same times, we may imagine that trends were originally parallel and that an economic shock other than (but concomitant with) the treatment has generated divergent post-treatment trends and has complicated the application of the DiD method.

Pre-trend testing is common but suffers from low power and can introduce pre-test bias. The low power character of pre-trend testing is discussed by a large literature on DiD and many attempts have been carried on elaborating different approaches to the testing

procedures and their results. A range of approaches to this issue discussed in the DiD literature are reviewed below.

Rambachan and Roth (2023, p. 2055–56) propose methods for robust inference and sensitivity analysis in empirical settings where the parallel trends assumption may not hold but the post-treatment violations of parallel trends is not "too different" from the pre-trends violations. They postulate that it is reasonable to conjecture that, when it existed, the pre-existing discrepancy in trends will carry on from the pre-treatment to the post-treatment periods. Yet, this discrepancy can be extrapolated and, instead of requiring that parallel trends condition holds exactly, restrictions can be imposed on how different the post-treatment violations of parallel trends can be from the pre-treatment level. To illustrate, it can be assumed that a specific disparity in trends in the employment rate between treated and untreated regions would persist subsequent to the implementation of, say, a policy wage intervention. Consequently, if the control group (untreated units in untreated regions) exhibits post-treatment higher employment growth, it can be hypothesized that the treated group's employment would have grown at a higher rate "anyhow". A comparison of the actual employment rate to this theoretical counterfactual would allow for the determination of whether the greater growth is attributable to the treatment but, under the above condition, would not provide unbiased estimates of the treatment effect.

Rambachan and Roth (2023) posit that researchers may instead prefer to consider robustness to some degree of deviation from the pre-existing trend. In this case, linear extrapolation would need only be "approximately" correct, instead of being totally accurate. The discrepancy between the observed trend and the preexisting path is permitted to deviate nonlinearly by an amount designated as M. The magnitude of this deviation is directly proportional to the amount of divergence from the preexisting trend. Subsequent to the rejection of the parallel trends assumption, the (pseudo-)parallel trend is examined through the assessment of the constraints on the potential post-treatment variations in trends (the value of the aforementioned M) in relation to the pre-treatment trend estimate. These restrictions are predicated on the premise that pre-trends are indicative of counterfactual post-treatment differences in trends. Subsequently, the researchers' paper demonstrates that, given a specific value of M, it is possible to ascertain a confidence set for the treatment parameter of interest. We deviate from the "pure" parallel trend assumption by implementing a partial identification approach, as outlined in Remark 1. This deviation stems from the necessity of a proper identification procedure. For instance, Rambachan and Roth (2023) examined the impact of a teacher collective bargaining reform on employment, finding that parallel trends emerged for males, but a pre-existing negative trend was evident for females. The DiD estimate for males at M = 0 (linear extrapolation of the pre-existing trend) is obtained, and it is subsequently demonstrated that confidence intervals increase in size as M is increased, i.e., as a greater deviation in trends is permitted. Conversely, for females, the DiD estimator reveals an opposing sign. Rambachan and Roth (2023) posit that applied researchers must motivate the imposed restrictions on parallel trends with economically compelling reasons, report robust confidence intervals, and conduct formal sensitivity analyses to ascertain the assumptions necessary to draw any conclusions.[4]

A second approach is provided by Bilinski and Hatfield (2020). The traditional parallel trend pre-tests are the subject of critique due to the possibility of failing to reject parallel trends, due to either low test power or, conversely, high test power. Nevertheless, the rejection of parallel trends offers limited insight into the magnitude of the violation and its implications for the results. Bilinski and Hatfield (2020) argue that the most popular approach to testing parallel trend (H_0 that pre-intervention trends are parallel was acknowledged of 293 quotations in their Table 1 list) is incorrect and frequently misleading.[5] Hence, they present test reformulations in a non-inferiority framework that, according to their

findings, rule out violations of model assumptions that exceed a threshold. Then, they focus on the parallel trends assumption, for which they propose what they call a "one step up" method: (1) reporting treatment effect estimates from a model with a more complex trend difference than is believed to be the case and (2) testing that the estimated treatment effect falls within a specified distance of the treatment effect from the simpler model. This reduces bias while also considering power, controlling mean-squared errors. Their base model also aligns power to detect treatment effects with power to rule out violations of parallel trends.

A third approach was proposed by Freyaldenhoven et al. (2019) and involves the introduction of instrumental variables (IV) to adjust for the possible violation of parallel trends. These IV can be conceptualised as a covariate influenced by confounders yet not by the treatment. Subsequently, the covariate can be utilised to elucidate the dynamics of the confounding variable and adjust the DiD estimations for it, thereby providing the impact of the policy change.[6] The aforementioned authors provide two examples of their procedure. The initial inquiry concerns the impact of SNAP program participation on household spending. The primary dataset encompasses SNAP participation and the resulting expenditure outcomes. The objective of the present study was to examine the impact of the program in question. A secondary concern that was addressed was the possibility that income trends could influence both program participation and spending. Utilising a secondary dataset encompassing SNAP participation and income, researchers can instrument labour participation with leads of income. This approach necessitates the assumption that households do not curtail their labour supply in anticipation of receiving the program, thereby enabling the application of the aforementioned method.

2.2. The Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978, 1980, 1990a)

DiD identification require that the treatment applied to one (or more) unit does not affect the outcome for other units. Following the definition of Angrist et al. (1996), Rubin (1980, 1990b), and Wooldridge (2010, p. 905), by Stable Unit Treatment Value Assumption (SUTVA) in causal studies we mean that the potential outcome for a generic unit does not depend on the treatment status of the other units nor on the mechanism by which units are assigned to the control and treatment groups. In other words, treated and untreated units are expected not to mutually interfere and do not influence their outcomes (Cox, 1958). The above authors themselves point out that the assumption is critical and does not always match with real situations. For instance, let us consider a generic market in which operators mutually know and interact, thus influencing the reactions to exogeneous or external events, or policies in which "spillover effects" among neighbours can affect the choices of people involved in the experiment (Sobel, 2006). Similarly, one could consider panel data settings in which units interact across temporal (e.g., anticipation effects), cross-sectional (Xu, 2024), and spatial dimensions (Wang, 2021; Xu, 2024). Imbens and Rubin (2015, p. 10) use the example of the fertiliser applied to one plot that affected the yields in contiguous untreated plots. Another example might be that of students assigned to attend a tutoring program to improve their grades (treated units) who might interact with other students in their school who were not assigned to the tutoring program (untreated control units) and influence the grades of the latter. Treated students might "informally" affect the performance of the control students since their interaction can generate spillover effects of the treatment in favour of untreated students. Under these circumstances, to enable causal inference, the analysis might be completed at the school level rather than the individual level. SUTVA would then require no interference across schools, a more plausible assumption than no interference across students.

Hence, SUTVA demands that the potential outcomes for some untreated unit do not vary with the treatments assigned to some other treated units. In other words, a subject's potential outcome is not affected by other subjects' exposure to the treatment. The SUTVA implies that each individual has one and only one potential outcome under each exposure condition, that is, with and without treatment (Schwartz et al., 2012), thus making the causal effect "stable". On the contrary, when the SUTVA is not fulfilled, there could exist multiple potential outcomes for each individual under each exposure condition (i.e., the causal effect is not unique), potentially leading to lead to misleading inferences. In non-economic frameworks, researchers often add a second aspect of stability in causal studies closely related to the original SUTVA, that is, the so called "consistency assumption" (Cole & Frangakis, 2009; VanderWeele, 2009) or "no-multiple-versions-of-treatment assumption", which states that potential outcomes of individuals exposed to the treatment coincide with their observed outcomes. In other words, there are no hidden forms of treatment leading to different potential outcomes (Cerqua et al., 2023, 2024; Bosco et al., 2025).

Laffers and Mellace (2020) introduced a third source of violation of the SUTVA, that is, the presence of measurement errors in either the observed outcome or the treatment indicator. While this new perspective extends the definition of the SUTVA, the authors also propose a way to relax the assumption by means of a sensitivity study. Specifically, they suggest computing the maximum share of units for which SUTVA can be violated without changing the conclusion about the sign of the treatment effect. According to the specificities of the empirical setting of interest, several other attempts to extend and to relax the SUTVA can be found in the recent literature (see, for instance, the papers by Qiu & Tong, 2021; VanderWeele et al., 2015, for a recent review on causal inference in the presence of interference). For instance, considering the case when all units are affected by the treatment, Cerqua et al. (2022) make use of a machine learning counterfactual framework in which the no-interference part of the SUTVA is substituted by a milder definition only requiring that the potential outcomes for treated units are not affected by the individual characteristics of the other treated units. Indeed, in Cerqua et al. (2023), the authors remove entirely the no-interference assumption and rely solely on the no-multiple-versions-of-treatment assumption, as they are aware that in many socio-economic applications, agents are sensibly affected by interference across both space and time. Other strategies attempt to relax the assumption by using clustered or hierarchical data structures (for instance, individuals living in restricted areas such as neighbourhoods) with potential spatial spillovers. VanderWeele (2010), for instance, introduced the definition of individual-and-neighbourhood-level SUTVA and neighbourhood-level SUTVA to deal with empirical settings in which cluster-level interventions are considered. Among others, Huber and Steinmayr (2021) allow for the interaction between individuals and higher-level structures (e.g., regions) and suggest a non-parametric modelling to separate individual-level treatment effects from spillover effects. However, while the SUTVA may be violated on the individual level, it must hold at the aggregate level. The latter can be referred to as the *regional SUTVA*, which admits spillover effects between individuals within regions, but rules out spillovers across regions. Under this new setting, the total treatment effect may be split up into an individual effect and a within-region spillover effect driven by the treatment of other individuals in the region. Eventually, Ogburn et al. (2020, 2024) considered the potential spillover effect produced by a network in which individuals mutually interact and treated individuals may spread the treatment to their social contacts.

Remark 3. *The basic ingredients: the SUTVA assumption.*

The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

Essentially, Remark 3 states that an individual's potential outcome under a given treatment depends neither on the treatment received by other individuals nor on different versions of the treatment itself. In other words, each individual has only one potential outcome for each level of treatment, and this outcome is independent of the treatment received by others. Specifically, we must make sure (i) that an individual's outcome is not influenced by the treatment received by other individuals—for example, if we are evaluating the effect of a drug, SUTVA implies that whether a patient takes the drug does not affect the outcome of another patient who might not take it—and (ii) that for each level of treatment, there is a unique version of the treatment that leads to a given potential outcome. This means that there are no different versions of the treatment that could lead to different outcomes for the same individual.

We may conclude that SUTVA is crucial for the correct interpretation of causal effects because, if violated, it can lead to biased estimates of treatment effects. For example, if interference is present (a student treated with a new textbook shares the improvements in their knowledge with an untreated fellow student), we may not be able to distinguish the effect of the treatment from the effect of the interactions between treated and untreated individuals.

2.3. Exogeneity and Identification DiD and Traditional Econometrics

In OLS regression analysis, we are interested in assessing the effect of a (usually) continuous variable x on a dependent Y under the hypothesis of exogeneity. The "true" causal effect of x on Y can be identified as long as independent changes in x only produce a direct effect on Y by ruling out any potential indirect effect of x on Y occurring via the relation of x with unobservable factors. Without this exogeneity condition, OLS produced biased estimated parameters. Using Cerulli's (2015, p. 6) example, we assume that the regression model is as follows:

$$Y = \beta x + u$$

where β represents the causal effect of x on Y and u is a non-observable factor. By differentiation, we have the following:

$$dy/dx = \beta + du/dx$$

The model is identified as long as $\frac{du}{dx} = 0$. If $\frac{du}{dx} \neq 0$, the autonomous changes in x are not exogenously determined, as x has also an indirect effect on Y through its effect on u, and since u is not observed, we cannot separate the direct effect (β) and the indirect effect ($\frac{du}{dx}$) and the model is no longer identified.

The counterfactual approach of the DiD to causality can be reformulated in terms of OLS model with the x assuming a binary form (say x_0 for the treated and x_1 for the untreated) instead of a continuous x. If we can observe two responses (Y_0 state: treatment and Y_1 state: no treatment), we write the following:

$$Y_1 = \beta x_1 + u_1$$

$$Y_0 = \beta x_0 + u_0$$

Subtracting the second equation from the first, we have the following:

$$Y_1 - Y_0 = \beta(x_1 - x_0) + u_1 - u_0$$

Or

$$\Delta y = \beta \Delta x + \Delta u$$

Then

$$\frac{\Delta y}{\Delta x} = \beta + \frac{\Delta u}{\Delta x}$$

If $\frac{\Delta u}{\Delta x} \neq 0$, a bias like the previous one is generated even when we use binary-form data typically associated with treatment events and counterfactuals. Exogeneity of treatment is a necessary condition.

In DiD analysis, we may go a little further and portrait the identification problem using the assumptions made above. We can re-write the target estimand (which involved unobserved counterfactuals) in a form that depends only on observed outcomes. In DiD, we call this process "identification". To do so, we assume that the change in response from pre- to post-intervention in the control group is a good proxy for the counterfactual change in untreated potential outcomes in the treated group. When, in a 2 periods framework, we observe the treated and control units only once before treatment ($t = 1$) and once after treatment ($t = 2$), we write this as follows:

$$E\Big[y^0(2) - y^0(1)\Big|D = 1\Big] = E\Big[y^0(2) - y^0(1)\Big|D = 0\Big]$$

Notice that it involves unobserved counterfactual outcomes, namely $y^0(2)\big|D = 0$ (the potential realisation of y in case of no treatment; recall from Figure 1 that these data no not exist). This is another way to state the parallel trend assumption or the counterfactual assumption.

We also need to make another, more explicit assumption of Section 2.1. For DiD, the treatment status of a unit can vary over time. However, we only permit two treatment histories: never treated (the control group) and treated in the post-intervention period only (the treated group). Thus, we will use $D = 0$ and $D = 1$ to represent the control and treated groups, with the understanding that the treated group only receives treatment whenever $T > T_0$. Every unit has two potential outcomes, but we only observe one—the one corresponding to their actual treatment status. The consistency assumption links the potential outcomes $y^d(t)$ at time t with treatment d with treatment $d \in D = (0, 1)$ to the observed outcomes $y(t)$:

$$y(t) = (1 - D)y^0(t) + Dy^1(t)$$

Finally, we add the assumption that future treatment does not affect past outcomes. Thus, in the pre-intervention period, the potential outcome with (future) treatment and the potential outcome with no (future) treatment are the same (no anticipation effects).

Using the assumptions made above, we can re-write the target estimand (which involved unobserved counterfactuals) in a form that depends only on observed outcomes. In DiD, this process is specifically called "identification" and should not be confused with the specification problems typical of traditional OLS single equation regressions or with the so-called over or under identification problems emerging from multi-equation OLS systems. The DiD identification relies on the Counterfactual Assumption and the Consistency Assumption discussed above and ends with the familiar DiD estimator where, for reducing notation, we use D instead of $D2$ of Equation (1) and indicate periods as numbers between parenthesis (see Callaway, 2022, p. 8):

$$
\begin{aligned}
ATET &= E\left[y^1(2) - y^0(2) \middle| D = 1\right] \equiv \text{Definition of } ATET \\
&= E\left[y^1(2) \middle| D = 1\right] - E\left[y^0(2) \middle| D = 1\right] \\
&= E\left[y^1(2) \middle| D = 1\right] - \left\{E\left[y^0(2) - y^0(1) \middle| D = 0\right] + E\left[y^0(1) \middle| D = 1\right]\right\} \\
&\quad \text{by counterfactual assumption} \\
&= \left\{E\left[y^1(2) \middle| D = 1\right] - E\left[y^0(1) \middle| D = 1\right]\right\} - \left\{E\left[y^0(2) \middle| D = 0\right] - E\left[y^0(1) \middle| D = 0\right]\right\} \\
&= \{E[y(2)|D = 1] - E[y(1)|D = 1]\} - \{E[y(2)|D = 0] - E[y(1)|D = 0]\} \\
&\quad \text{by consistency assumption}
\end{aligned}
$$

You may compare the above ATET with the result obtained in Section 2.1. To simplify reading and comparison, we summarise the meaning of the above terms as follows:

- $E[y(2)|D = 1]$ is the post-intervention average response of the treated group.
- $E[y(1)|D = 1]$ is the pre-intervention average response of the treated group.
- $E[y(2)|D = 0]$ is the post-intervention average response of the control group.
- $E[y(1)|D = 0]$ is the pre-intervention average response of the control group.

To summarise the above derivation, we observe that DiD identification begins with the ATET, then the Counterfactual Assumption and the Consistency Assumptions are applied to obtain the familiar DiD estimator.

When we observe the treated and control units many times before and after treatment, we must adapt the target estimand and identify assumptions accordingly. Identification problems with multi-period DiD are discussed later; Appendix A provides a guided example with an easy visualisation of the dataset (see Table A1).

3. The OLS Version of the Two-Way Fixed Effects Regression (TWFE)

TWFE is the most common way to implement a DiD identification strategy under the assumption of treatment homogeneity. In this section, we present what Roth et al. (2023, p. 2224) call a "static" TWFE, which regresses the outcome variable on individual and period fixed effects and an indicator for whether the unit h is treated in period t. Recall that in Section 2.1, we defined the following:

$$ATET = E\left[y^1(2) - y^0(2) \middle| D2 = 1\right]$$

$$\equiv \{E[y(2)|D2 = 1] - E[y(1)|D2 = 1]\} - \{E[y(2)|D2 = 0] - E[y(1)|D2 = 0]\}$$

Then, the estimated ATET can be written by replacing population means by their sample analogues (indicated by upper bars) to obtain the following:

$$\hat{ATET} = \{E[\overline{y}(2)|D2 = 1] - E[\overline{y}(1)|D2 = 1]\} - \{E[\overline{y}(2)|D2 = 0] - E[\overline{y}(1)|D2 = 0]\}$$

The above expression is algebraically equivalent to either of the following OLS regression systems:

$$\begin{cases} y_{ht} = \theta_t + \eta_i + \alpha D_{ht} + v_{ht} \\ \\ y_{ht} = \theta_t + \eta_j + \alpha D_{ht} + v_{ht} \end{cases}$$

where i indicates treated units, j indicates untreated units, and t is time, while h in Equation (2) below can be either i or j. The interpretation of the quantities involved in (2) is the following:

- y_{ht} is the response variable;
- θ_t is a time effect;
- η_i or η_j are unit (not group) fixed effects;

- D_{ht} is the dummy (indicator) for whether or not unit h is affected by the treatment in period t (the term $D1 \times D2$ of the last column of Table A1);
- v_{ht} are idiosyncratic, time-varying unobservable factors.

Equivalently, the previous system can be written in a single equation (panel data) version as follows:

$$y_{ht} = \eta_h + \theta_t + \alpha[D1_h \times D2_t] + \varepsilon_{ht} \tag{2}$$

where η_h is the individual's fixed effect, θ_t is the time fixed effect, and $[D1_h \times D2_t]$ is the treatment dummy interaction having the coefficient α. We can estimate Equation (2) and interpret the estimated coefficients according to the result reported in Remark 4.

Remark 4. *Causal interpretation of the TWFE estimator.*

Under parallel trend, treatment homogeneity, SUTVA, no anticipation, and no spillover, $\hat{\alpha}$ in Equation (2) is the TWFE estimate of the causal effect of receiving the treatment.

As the very name suggests, TWFE is the case where there are exactly two time periods, where no units are treated in the first time period, and where some units become treated in the second time period only while other units remain untreated in the second time period. Notice that when we say periods we do not necessarily refer to units of time (years, months, etc.) but to "time intervals": the first (possibly composed by several years, several months, etc.) in which there is no treatment for nobody and the second (possibly composed by several years, several months, etc.) in which some units are treated (uniformly).

To illustrate TWFE formally, we need some notation. Let us define the following quantities:

- t^* and $t^* - 1$ are the two periods of interest that, for simplicity, correspond to two years;
- D_h is the treatment indicator $D1 \times D2$ of Table A1 so that

$$D_h = \begin{cases} 1 \text{ for treated units during treatment periods} \\ 0 \text{ for untreated units} \end{cases}$$

Then, for $t \in \{t^* - 1, t^*\}$ let $y_{ht}(1)$ be the unit's potential treated response in period t and correspondingly $y_{ht}(0)$ be the unit's potential untreated response in period t. Impose that $y_{ht^*-1}(1) = y_{ht^*-1}(0)$ for all units. This is the *no anticipation condition* presented in Section 1. It states that the treatment should not affect the response variable in periods before the treatment takes place. The result from the above assumption and conditions is the following:

$$y_{ht*-1} = y_{ht*-1}(0) \text{ and } y_{ht*} = D_h y_{ht*}(1) + (1 - D_h) y_{ht*}(0)$$

In the first time period, we observe untreated potential outcomes for the response variable for all units, and in the second period, we observe treated potential outcomes of the response variable for treated units and untreated potential outcomes of the response variable for untreated units.

Using the above definitions, we may define the ATET resulting from the DiD identification of the treatment effect as follows:

$$ATET = E[y_{t^*}(1) - y_{t^*}(0) \mid D = 1]$$

which is equivalent to the one given in the previous section.

Using Callaway's (2022, p. 6) definition, the ATET is the mean difference between treated and untreated potential outcomes among the treated group. Perhaps the main reason that the DID literature most often considers identifying the ATET rather than, say, the

average effect of treatment is that, for the treated group, the researcher observes untreated potential outcomes (in pre-treatment time periods) and treated potential outcomes (in post-treatment time periods). The DID identification strategies exploit the above framework. As a result, it is natural to identify causal effect parameters that are local to the treated group. Clearly, the model presented in Equation (2) is the static specification of the TWFE, which yields a sensible *estimand* when there is no heterogeneity in treatment effects across either time or units. Following Roth et al. (2023, p. 2224), we can stress the relevance of these hypotheses more formally.

Define a period (e.g., year) $g > t$ and let $\tau_{h,t}\ (g) = Y_{h,t}(g) - Y_{h,t}(\infty)$. Suppose that for all units h, $\tau_{h,t}(g) = \tau$ whenever $t \geq g$. This implies that (a) all units have the same treatment effect and (b) the treatment has the same effect regardless of how long it has been since treatment started. Then, under a suitable generalisation of the parallel trends assumption and no anticipation assumption, the population regression coefficient α in Equation (2) is equal to τ.

Yet issues arise, however, when there is heterogeneity of treatment effects over time, as shown in Borusyak and Jaravel (2018), de Chaisemartin and D'Haultfoeuille (2020), and Goodman-Bacon (2021), among others. More generally, if treatment effects vary across both time and units, then $\tau_{h,t}(g)$ may obtain a negative weight in the TWFE estimand for some combinations of t and g.

Figure 2 gives the idea of a parallel trend with 3 units (unit 1 drawn in blue colour is untreated) and 10 periods. The plot has been generated using the data of Table A2 in Appendix A. The following plot illustrates the time paths of the response variable. Notice that Unit 1 is never treated, while Unit 2 and Unit 3 start the treatment at time $t = 5$ and are always treated from $t = 5$ to $t = 10$.

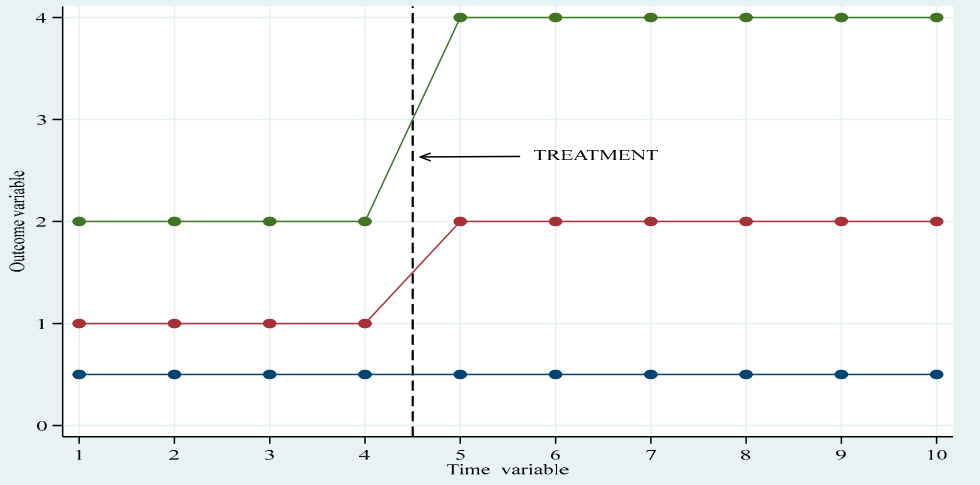


Figure 2. Parallel trends plot with two treated units (homogeneous case) and one control unit. The bottom solid line is the response variable y of the untreated Unit 1. The treatment is introduced from period $t = 5$. The two upper lines correspond to the data of Units 2 and 3. Trends are imposed to be parallel in treated and untreated periods for simplicity. Comparing the treated units with the control group might give us an understanding of whether the treatment is responsible for the change in the line path of the response variables for Unit 2 and Unit 3.

3.1. Testing for the Parallel Trends and Anticipation Effects Assumptions in the TWFE Model

Given the fundamental importance of the parallel trend assumption, a natural question is how to test for parallel trends in a panel data TWFE? In order to find answers, we start from the above model written as follows:

$$y_{ht} = \beta_0 + \sum_{k=T_0}^{T} \beta_k I(t = k \cap D_h = 1) + \alpha_h + \gamma_t + \varepsilon_{ht}$$

where $I(\cdot)$ is the indicator function. The treatment is indexed by D_h, with D_h = 1 indicating that observation h is part of the treated groups/units and D_h = 0 indicating that it is part of the comparison/control group. Let t index time $\{1, \ldots, T\}$ and suppose that an intervention begins at time T_0 for the treated units (same treatment periods for all treated units: the so-called homogenous case. See below). All other symbols have their usual meaning. The treatment effects of interest are β_k, representing differential post-period changes in the treated group relative to comparison at each time point. The average of these coefficients, that is, the following:

$$\beta = \frac{1}{T - T_0 - 1} \sum_{k=T_0}^{T} \beta_k$$

is the average ATET. Then, we may exploit the above definition of the coefficients to derive some tests of the DiD identification hypothesis with respect to parallel trends and anticipation effects.

Parallel trends test: the slope test. In a parallel trends test for DiD identification, we may try and estimate whether there is a difference in slope between treatment and comparison groups prior to the intervention. Call θ the different coefficient. Then, rewrite the above equation in a form that incorporates the pre-treatment coefficient θ:

$$y_{ht} = \beta'_0 + \sum_{k=T_0}^{T} \beta'_k I(t = k \cap D_h = 1) + \theta(D_h \times t) + \alpha_h + \gamma_t + \varepsilon'_{ht}$$

We can now test whether the (pre-treatment) differential slope θ = 0. If the null hypothesis for this test is not rejected (i.e., Prob. > 0.05), researchers may conclude that trends are parallel.

Anticipation effects: Researchers may instead examine the validity of the identification of the parameter by DiD by testing whether there is a significant "treatment" effect prior to the intervention, that is, an effect starting at $T^* < T_0$. In this context, they might use the modified original panel model:

$$y_{ht} = \beta_0 + \sum_{k=T^*}^{T_0-1} \theta_k I(t = k \cap D_h = 1) + \alpha_h + \gamma_t + \varepsilon_{ht}$$

and estimate it by omitting data from after T_0. If the test statistics

$$\theta = \frac{1}{T - T^* - 1} \sum_{k=T^*}^{T_0-1} \theta_k$$

is significant, this again suggests a violation of parallel trends. (Alternatively, a joint F-test can be used to test whether placebo effects at all possible $T^* < T_0$ were insignificant.)

3.2. More on the Parallel Trend Assumption

We have already stressed that DiD does not identify the treatment effect if treatment and control groups were on different trajectories prior to the treatment (common trend or parallel trend assumption).

With respect to Equation (1) as the OLS equation of our DiD model we recall the following:

- Selection bias relates to the fixed characteristics of the units η_h;
- Time trend θ_t is the same for treated and untreated units.

These assumptions guarantee that the common trends assumption is satisfied, but quite unfortunately, they cannot be easily tested directly, and sometimes parallel trends are checked by mere visual plot inspection.

In Figure 1, we illustrated the case of a clear pre-treatment parallel trend. In other cases, the violation of the assumption may be easily evaluated. The time paths of Figure 1 generate a case of non-optically distorted test of parallel trend. One arrives at the same conclusion after seeing an alternative presentation of the plots in which the G2 line start from A (after elimination of the difference AB, which is the idiosyncratic constant element) and detects whether G1 and G2 lines overlap before the introduction of the treatment and diverge in period 2. In another section, we provide an example of the latter way of plotting and interpreting path of the data. An example is reproduced in the next section (Figure 3 below).

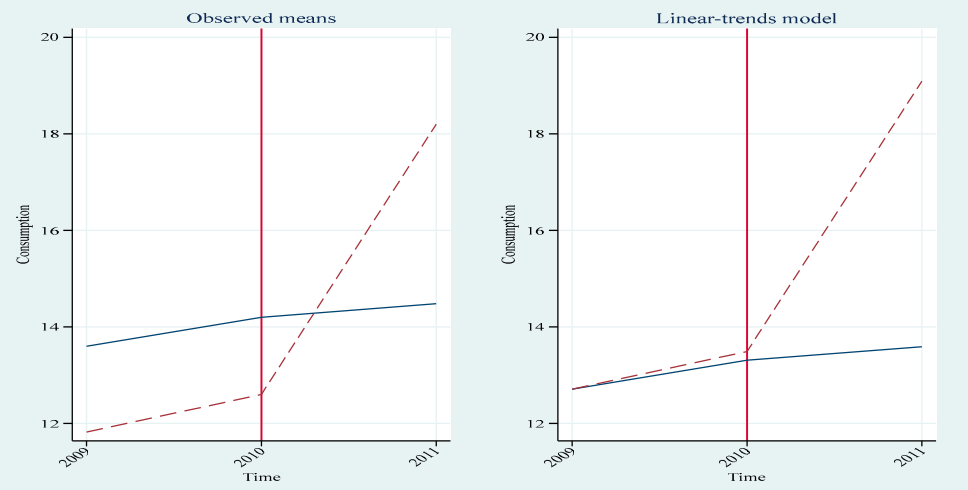


Figure 3. Parallel trends plots using the worked example data of Table A4. The plot depicts the mean values of the response variable before and after treatment (year 2010). The left plot shows the observed means whereas the right plot illustrates the linear trend of both series of means after they are forced to start from the same intercept (the initial difference is supposed to remain constant and is removed). Under parallel trend, the elimination of the initial (supposedly constant) difference should make the pre-treatment path overlap. If the treatment is effective, the post-treatment paths should show appreciable differences between each other (treated data = dashed lines). Please recall that the presence of parallel trends cannot be diminished and debased to a mere matter of good *optical* observation, however sophisticated the plots' definition can be. Plots are generated using Stata 18 package.

Let us go back to the initial 2×2 case and present a discussion of the parallel trend relevance. We had two groups, one treated and one untreated, and we indicated them as follows: $g \in \{0,1\}$, where 0 is untreated (control) and 1 is treated. We also had 2 years, and then we write $t \in \{0,1\}$, where 0 is the before-treatment period and 1 is the treatment period. To guarantee of a consistent estimate of the ATET, we need to make the following parallel trend assumption:

$$E(y_{i01}|D_{gt}=1) - E(y_{i00}|D_{gt}=1) = E(y_{i01}|D_{gt}=0) - E(y_{i00}|D_{gt}=0)$$

If the treated units had not received the treatment, the response variable for groups defined by $D_{gt}=1$ and $D_{gt}=0$ should show the same paths as in Figures 1 and 2. The group effects must be time-invariant, and the time effect must be group-invariant.

Within a two-period framework, the possible "test" of this assumption is only graphical, but for more than two periods, the same testing procedures based on the Wald test are available. Many econometric software packages offer such statistical tests. We will present them alongside applications at the end of this review.

In the linear case, Wooldridge (2025) has shown that tests of the Parallel Trend assumption are easily carried out in the context of pooled OLS estimation. In other words, in linear DiD models within a staggered treatment framework, the parallel trends assumption can be tested using pooled OLS estimation. This approach leverages the inclusion of cohort and time period dummies, along with cohort-by-time treatment indicators, in a linear regression model. The key idea is that under the parallel trend assumption, the coefficients on these interaction terms, when estimated via pooled OLS, are consistent for the estimation of ATET. Moreover, the tests are the same whether based only on the $D_{it} = 0$ observations (imputation regression) or on pooled OLS using all observations—provided full flexibility is allowed in the treatment indicators. In other words, tests obtained pooling over the entire sample are equivalent to the commonly used 'pre-trend' tests (i.e., common tests used to examine the parallel trends assumption) that use only the untreated observations. As discussed by Wooldridge (2025), this means the tests using post-treatment data are not 'contaminated' by using treated observations—if the treatment effects are allowed to be flexible.

The algebraic equivalence of the pooled tests and pre-trends tests carries over to the nonlinear case, provided the canonical link function is used in the Linear Exponential Function (LEF). In a DiD analysis with panel data, when using a linear exponential family (LEF) with the canonical link function, the pooled tests (using all data) and pre-trends tests (using only untreated observations) are algebraically equivalent. This means that the same underlying statistical properties and results can be obtained regardless of whether you pool all the data or focus only on pre-treatment observations to check for parallel trends. Technically, if one uses a different mean function or different objective function, the test should be carried out using only the $D_{it} = 0$ observations (although it seems unlikely that the difference would be important in practice). Wooldridge (2023) recently discussed the non-linear case.

In general, one should consider that the implications for applied work revolve around the (often-implausible) parallel trend assumption needed for the identification (using non-treated post treatment observations as counterfactuals) of a DiD model. Yet, rather than just asserting that parallel trends hold or abandoning projects where a pre-test rejects parallel trends (not to speak of the so-called *optical* test based the trend plots!), new approaches focus on thinking carefully about what sort of violations of parallel trends are plausible and examining robustness to these. Importantly, these methods should be used when there is reason to be sceptical of parallel trends ex ante, regardless of the outcome of a test of whether parallel trends hold pre-intervention. This type of descriptive analysis will allow one to get bounds on likely treatment effects. For instance, a recent application comes from Manski and Pepper (2018), who look at how right-to-carry laws affect crime rates, obtaining bounds on the treatment effect under different assumptions about how much the change in crime rates in Virginia would have differed from those in Maryland in the absence of this policy change in Virginia.

In summary, the default DiD estimation equation should allow for a linear trend difference. This is a key recommendation of Bilinski and Hatfield (2020).

The specific approach used to examine robustness depends on how many pre-periods you have. For instance, with only a small number of pre-intervention periods, the Rambachan and Roth approach of bounding seems most applicable for sensitivity analysis;

when you have more periods, you can consider fitting different pre-trends as in Bilinski and Hatfield (2020). Some issues are discussed in sections below.

3.3. OLS and the Efficiency of the Estimation of the Treatment

Generalising our discussion beyond the two-group and two-period case, we may note here that the response variables under investigation may vary at the group and time levels. This implies that it may be necessary to correct the estimations to take serial correlation into account and avoid inefficient coefficient estimates as well as unreliable standard errors. When serial correlation is present and the number of groups is not too small, Bertrand et al. (2004) suggest the use of robust standard errors defined at the group level. Moreover, Bester et al. (2011) further show that using cluster-robust standard errors and critical values of a t distribution with degrees of freedom, equal to the total number of groups minus one, is asymptotically valid for a fixed number of groups. In other words, they show that consistency does not require the number of groups to be randomly large and that we could still obtain reliable standard errors even when the number of groups is not too large. Unfortunately, cluster-robust standard errors may still have poor correcting effects when the number of groups is too small or when the number of treated groups is small relative to the control groups.

For cases where the number of groups is small, one may use the wild cluster bootstrap that imposes the null hypothesis that the *ATET* is zero. Cameron et al. (2008) and MacKinnon and Webb (2018) show that the wild cluster bootstrap provides better inference than using cluster-robust standard errors with critical values determined according to time and group number. On the other hand, Imbens and Kolesar (2016) show that with a small number of groups, one may use bias-corrected standard errors with the degrees of freedom adjustment proposed by Bell and McCaffrey (2002). Another alternative is proposed by Donald and Lang (2007), who propose a specific aggregation strategy and show that their method works well when the number of groups is small but the number of individuals in each group is large.

The reader is referred to all the papers quoted above for a discussion of technical details that cannot be conducted in the context of this review. Yet a practician should bear in mind that when there is a large disparity between treatment and control groups, or there is only one treated group, or when the group sizes vary significantly, cluster-robust standard errors and the other methods mentioned above may fail to provide efficient solutions. Then, the bias-corrected and cluster–bootstrap methods may provide an improvement over the cluster-robust standard errors. For a more discussion on this issue, see Cameron and Miller (2015), MacKinnon (2019), and MacKinnon et al. (2023) as well as the references therein.

4. Simple Worked Examples

We offer two simple numerical examples of ATET estimation with OLS and of the interpretation of the estimated coefficients. The second example relates to the interpretation of the response variable paths (before and after the treatment) as a tool for the evaluation of the presence of parallel trends.

4.1. Example n.1: Equivalence Between the OLS Estimation and the Calculation Based on Mean Differences

Assume we have a total of 10 consumers in 2 equal-sized groups (a group of 5 treated consumers and a group of 5 untreated consumers); 2 periods corresponding to two years, namely 2010 and 2011; a treatment occurring at the end of 2010 (e.g., a reduction in the consumption tax for the treated group only). We name treated consumers Mrs. 1–5 and untreated consumers Mrs. 6–10. Data and dummies are presented in Table A3 in the Appendix A.

Using the above dataset, we show by direct calculation of mean values how to recover the ATET induced by the treatment. We need the following quantities:

- The mean Consumption in the Control group before the treatment is as follows:

$$E[y|D1 = 0 \wedge D2 = 0] = 14.2$$

- The mean Consumption in the Treated group before treatment is as follows:

$$E[y|D1 = 0 \wedge D2 = 1] = 12.6$$

- The mean Consumption in the Control group after the treatment is as follows:

$$E[y|D1 = 1 \wedge D2 = 0] = 14.48$$

- The mean Consumption in the Treated group after the treatment is as follows:

$$E[y|D1 = 1 \wedge D2 = 1] = 18.2$$

Results can be synthesised in the following 2×2 format reported in Table 2:

Table 2. Numerical values of the differences used for the Average Treatment Effect.

	Control	**Treated**
Pre-Treatment	14.2	12.6
Post-Treatment	14.48	18.2

Therefore, we obtain the ATET as the difference in the two differences:

$$DiD = 3.72 - (-1.6) = 5.32$$

Clearly, the above calculation does not tell us how "good" the computed ATET is from an inferential point of view. In other words, 5.32 has no CI around it or p-values. That is why we must re-obtain the result following a route that allows us to introduce inferential elements. Now we estimate ATET with the OLS after the creation using the above-defined $D1$, $D2$, and TRET = $D1 \times D2$ (reported in small letters in Table 3). We run the OLS (with the option of robust SE) regression of Equation (1) and obtain the results reported below.

Table 3. Result of OLS regression.

Linear regression				Number of obs	-	20
				F(3, 16)	-	4.63
				Prob > F	-	0.0163
				R-squared	-	0.5951
				Root MSE	-	1.8926
consumption	coefficient	Roust std. err.	t	$p > \|t\|$	[95% conf.	interval]
d1	0.28	0.6822023	0.41	0.687	−1.166204	1.726204
d2	−1.6	1.183216	−1.35	0.195	−4.108306	0.9083058
TRET	5.32	1.692749	3.14	0.006	1.731532	8.908468
_cons	14.2	0.583095	24.35	0.000	12.96389	15.43611

Then, the estimated ATET is 5.32, which, according to the t-test reported in the table, is statistically significant at any level. Recall that TRET is the $D1 \times D2$ dummy variable. We can interpret the above estimated coefficients as it follows:

- The estimated Constant = 14.2 (with a p-value smaller than 0.05) is the mean value of the Consumption in the control group in 2010 (i.e., before the treatment). We can compare it with the result obtained from the numerical calculation reported above. The two figures coincide.
- If we sum the coefficient Constant and the d2 coefficient, i.e., if we calculate 14.2 + (– 1.6), we obtain 12.6. This is the expected Consumption of the control group in 2011, i.e., during the year of treatment.
- If we sum up the coefficient Constant and the d1 coefficient, i.e., if we calculate 14.2 + 0.28 = 14.48, we obtain the mean value of the Consumption in the treatment group in 2010, i.e., before the treatment.
- The estimated TRET = 5.32 is the (statistically significant) treatment effect. Treated units increase their average consumption by EUR 5.32 with respect to untreated individuals.

In formula, we may write, after indicating differences with the symbol Δ, the calculation of ATET as follows:

$$E\Delta[Consump|TaxTreatment = 1] - E\Delta[Consump|TaxTreatment = 0] = 3.72 - (-1.6) = 5.32$$

The above results are obtained as OLS under the robust SE option, which allows for adjusting the model-based standard errors using the empirical variability in the model residuals, i.e., the difference between observed and predicted outcomes, as suggested by Bertrand et al. (2004).

The example shows that the DiD strategy relies on two differences. The first is a difference across time periods. Separately for the treatment group and the control group, we compute the difference in the outcome means before and after the treatment. As discussed in Section 2, this across-time difference eliminates time-invariant unobserved group characteristics that confound the effect of the treatment on the treated group. This is important in the above example because the absence of cofactors might obscure the impact of the characteristics of the two units. Yet there may be time-varying unobserved confounders with an effect on the outcome mean, even after we control for time-invariant unobserved group characteristics. Therefore, we have used the second difference—i.e., the difference between the treatment group and the control group. Our DiD eliminates time-varying confounders by comparing the treatment group with a control group that is subject to the same time-varying confounders as the treatment group.

In a nutshell, *ATET* is then consistently estimated in the example as a one parameter (TRET) in a linear OLS equation by differencing the mean outcome for the treatment and control groups over time to eliminate time-invariant unobserved characteristics and also differencing the mean outcome of these groups to eliminate time-varying unobserved effects common to both groups.

4.2. Example n.2: Illustration of Parallel Trends

Using the data provided in Table A4 in Appendix A (data stuck in panel data form, i.e., in the version that is always recommended is defined by individual identity and time indicator) we generated a working dataset. Since the treatment is introduced at the end of 2010, we show graphically in Figure 3 below that the parallel trend exists. One can write the Fixed Effect panel data version of Equation (1) with time and individual effects and estimate both *ATET* and the time effect.

As one can see, in both groups, there was an increase in the mean of the outcome response variable after 2010. Therefore, the increase in the mean value of the response in the treatment group cannot be attributable entirely to the treatment (see Section 3). Yet the deviation from a common trend was more sizable from the treatment group, and the difference may visually indicate the effect of the treatment.

5. ATET vs. ATE

In the previous sections, we have used the acronym ATET to indicate the estimation of the average causal effect on treated units when our dataset includes both pre-treatment and post-treatment observations. It should not be confused with the Average Treatment Effect (ATE), which measures the effect of a treatment on a group of units estimated when we have observations recorded only for the after-treatment period (we do not have pre-treatment observations). Yet we would like to know if the treatment has an effect on the response variable y of the treated vs. untreated units. In an ideal world, we would observe y when a subject is treated (which we denote in what follows as y_1) and we would observe y when the same subject is not treated (which we denote as y_0). If the only difference in the data generation process of treated and untreated responses is the presence or absence of the treatment, we could average the difference between y_1 and y_0 across all the subjects in our dataset to obtain a measure of the average impact of the treatment. However, this ideal experiment setting is almost never available because we cannot observe a specific subject having received the treatment and having not received the treatment. When, for instance, the response is the level of consumption and the treatment is the presence or the absence of a consumption tax for a specific group of consumers, it is impossible to observe the consumers' expenditure under both treatment (the presence of the tax) and absence of the treatment (no taxes). As a result, we cannot estimate individual-level treatment effects because of a missing-data problem. Econometricians have developed potential-outcome models to overcome this problem. Potential-outcome models bypass this missing-data problem and allow us to estimate the distribution of individual-level treatment effects. A potential-outcome model specifies the potential outcomes that each individual would obtain under each treatment level, the treatment assignment process, and the dependence of the potential outcomes on the treatment assignment process. These models are beyond the purpose of this DiD review.

To illustrate the difference between ATE and ATET estimates, we follow Cameron and Trivedi (2005, p. 866). Define $\Delta = y_1 - y_0$, the above difference between the response variable in the treated and untreated states. Returning to Figure 1, one immediately realises that Δ cannot be observed (Group 2 after the treatment). Then, we define the following:

$$ATE = \mathbb{E}[\Delta] = \textit{Population average Treatment Effect}$$

whereas

$$ATET = \mathbb{E}[\Delta[D = 1] = \textit{Population average Treatment Effect upon Treated}$$

with the sample analogues (using the hypothesis of Section 1):

$$\hat{ATE} = M^{-1}\sum_{j=1}^{M}[\Delta_j]$$

$$\hat{ATET} = \left(\sum_{j=1}^{M}(D1 \times D2)_j\right)^{-1\sum_{j=1}^{M}(D1\times D2)_j}\sum_{j=1}[\Delta_j|D_j = 1]$$

The $\hat{ATE}$ version may be useful when the treatment has "universal applicability" (Cameron & Trivedi, 2005, p. 866) and we may consider the effect of the treatment for a randomly selected member of the population.

On the contrary, the $\hat{ATET}$ version is the measure of the average effect on treated units. It may be useful when the treatment has universal applicability and one wants to estimate its effect on a randomly selected subset of the population. Yet the estimation of the ATE is not straightforward, because as was mentioned above, we cannot simultaneously observe average outcomes of participants who are at the same time not participants, and a control group does not exist. An indication on how to specify treated and "untreated" observation to estimate the ATE is found in Cameron and Trivedi (2005, p. 867). Techniques are available to estimate various versions of $\hat{ATE}$. Wooldridge (2010, Ch. 21) discusses the assumptions and identification of ATE and presents the results (p. 929) of different estimation approaches.

6. The Confounding Factors

At the beginning of Section 1, we stated that confounding factors should be controlled for in DiD analysis. A confounder in DiD is a variable with a time-varying effect on the response outcome or a time-varying difference between groups. For example, if we run a DiD study on heart disease and therapy effects, we know that some coffee drinkers are smokers whilst some others are not. So, smoking is a confounding variable in the study of the association between coffee drinking and heart disease. The increase in heart disease may be due to smoking and not to coffee and can interact with the treatment administered to some units of patients. Hence, in DiD, we may adopt as a starting concept the colloquial definition of a confounder in cross-sectional settings: a variable associated with both treatment and outcome. As in the example of coffee drinkers, we may then think that the confounding elements in a DiD analysis arise because some covariates evolve over time differently in the treated and control groups or because the effects of covariates on outcomes vary over time. Then, confounders that vary over time and/or have time-varying effects on the outcome can cause violations of the parallel trends assumption. This concern has led scholars to develop methods to estimate the ATET coefficients under the assumption that parallel trends hold conditionally on covariates (see Roth et al., 2023 for a recent review). Methods that make a conditional parallel trends assumption prevailingly assume that control for pre-treatment covariates suffices. Researchers are often explicitly cautioned against controlling for post-treatment variables to avoid potential "post-treatment bias" (Rosenbaum, 1984; Myint, 2024).

To see why confounding factors can affect adversely our DiD estimations, we should recall that in DiD, our target estimand is the average effect of treatment on the treated (ATET):

$$\hat{ATET}(t^*) = E\left[y^1(t^*) - y^0(t^*)\middle| D = 1\right]$$

for some time $t^* \geq T_0$ where T_0 is the time the intervention is introduced to the treatment group.

Yet, in most settings, a confounder is a factor associated with both treatment D and the response variable. This is why randomised trials are not subject to bias through confounders—no factor is associated with the randomly assigned treatment. In other words, the potential outcomes and treatment are independent. Otherwise, we must make the following orthogonality assumptions:

- Assumption of unconditional independence between response and treatment:

$$y^d \perp D$$

or

- Assumption of conditional (on covariates X) independence between response and treatment: $y^d \perp D|X$

In both versions, the treatment D is independent of the potential outcomes y^d, either unconditionally or conditional on X.

As for practical applications, notice that these relations are only satisfied in randomised trials; otherwise, there is no guarantee that X is sufficient to make D and y^d conditionally independent. Even if we continue collecting covariates, it is likely that some unmeasured new covariates are still a common cause of D and y^d. Paradoxically, the fewer covariates we have, the smaller the probability of running into confounding factors trouble.

In summary, in DiD studies, the presence of confounding factors violates the counterfactual assumption when the following apply:

(1) The covariate is associated with treatment;
(2) There is a time-varying relationship between the covariate and outcomes;
(3) There is differential time evolution in covariate distributions between the treatment and control populations (the covariate must have an effect on the outcome).

As a conclusion, we may state that confounders are covariates that change differently over time in the treated and comparison group or have a time-varying effect on the outcome. When the confounder is appropriately included in a DiD regression model, unbiased estimates of ATET can be obtained with optimal SE. However, when a time-varying confounder is affected by the treatment, DiD may not be generate unbiased estimates of the causal effect. For more in-depth discussions of confounding for DiD, we recommend Wang et al. (2024) and Zeldow and Hatfield (2021).

7. More than Two Periods with Homogeneity

Assume we have some groups of units observed for more than two time periods (e.g., years > 2). After some year, a treatment is introduced and imposed to only a randomly selected subset of groups. If the treatment is administered to that subset of groups at the same moment and is maintained till the end of the time period and the rest of the groups are never affected, we have a case of panel data homogenous DiD. This is the case of no differential treatment time across groups. The opposite case is given by the administration of the same treatment at different moments to different groups (group 1 receives the same treatment at $g > t_0 >$ initial year, some other group at g_{+1}, some other at g_{+2}), where t_0 is the year of the first administration of the treatment to some group. This is the heterogeneous case. In this review, we always assume, for both the homogenous and the heterogeneous case, that once the treatment is administered, it stays in operation until the end of the sample period under study. Callaway (2022, p. 10) calls this Staggered Treatment Assumption, also called staggered case.

We start with the staggered homogeneous case. We specify this assumption as a Remark 5.

Remark 5. *Staggered Homogeneous Treatment Assumption* (*from Callaway, 2022, p. 10*).

For any unit and all $t = (1, \ldots, T)$ we assume that $D2_{it-1} = 1 \rightarrow D2_{it} = 1$. In other words, we assume that the treatment, once introduced, is active for **all treated units** until the end of the sample period. Then, in a sense, the treatment is irreversible.

The following Table 4 shows the distribution of the treatment across units and time that defines homogeneity.

Table 4. Homogeneous staggered treatment.

YEARS	2000	2001	2002	2003	2004	2005	2006	2007
UNITS	Period 1 (no treatment Dit = 0)			Period 2 (treatment Dit = 0 ∧ Djt = 1)				
1 2	Never treated							
3	Not yet treated			Treated since 2003 until 2007				
4	Not yet treated			Treated since 2003 until 2007				
5	Not yet treated			Treated since 2003 until 2007				
6	Not yet treated			Treated since 2003 until 2007				

Table 4 contains an example of a six-unit and eight-year panel with homogeneity of treatment and irreversibility. Here, we account for a treatment design where (i) the design is *staggered*, meaning that groups' treatments do not change over time and can change at most once; (ii) the treatment is binary; and (iii) there is no variation in treatment timing, that is, all treated groups start receiving the treatment at the same date and all the treated units terminate the treatment at the end of the sample period.

The TWFE can be employed to estimate a *DiD* model when data are generated according to the above framework. As a result, one may calculate the *ATET* at any $t > T_0$ of the post treatment period starting in T_0 as follows:

$$ATET(t > T_0) = E\left[y_h^{treated}(t) - y_h^{untreated}(t)\middle| D = 1\right] \forall t > T_0$$

or the average ATET as follows:

$$\overline{ATET}(t > T_0) = E\left[\bar{y}_h^{treated}(t > T_0) - \bar{y}_h^{untreted}(t > T_0)\middle| D = 1\right] \forall t > T_0$$

Notice that with the panel TWFE model, we increase the statistical power of DiD (under parallel average outcomes in pre-to post intervention periods), but the possible presence of serial correlation in treatment and outcome variables may be a problem (see Section 9.2).

8. More Than Two Periods with Heterogeneity

Assume now that we have some groups of units observed for more than two years and that the treatment is administered to a subset of groups in different moments during the sample period. If the treatment, once introduced, is maintained until the end of the period, we have a case of panel data DiD with staggered treatment (see Remark 5 above). We will always refer to Remark 5 as a case of "treatment irreversibility". The following Table 5 provides an illustration of the staggered treatment design for the case of six units and eight time periods, with two never-treated units and four units that started to be treated at different moments.

Table 5. Heterogeneous staggered treatment.

YEARS UNITS	2000	2001	2002	2003	2004	2005	2006	2007
1 2	Never treated							
3	Not yet treated		Treated since 2002					
4	Not yet treated			Treated since 2003				
5	Not yet treated				Treated since 2004			
6	Not yet treated					Treated since 2005		

Table 5 Example of a six-unit and eight-year panel with heterogeneity of treatment (different treatment windows). Here, we account for a treatment design where (i) the design *is staggered*, meaning that groups' treatments change over time and can change at most once; (ii) the treatment is binary; and (iii) there is more than one variation in treatment timing, that is, treated groups start receiving the treatment at different dates.

When the treatment is introduced in different periods of time, its impact changes (within treated units) over time, and we face a situation when average treatment effects vary over time and over cohorts (i.e., each group of units whose treatment started in the same moment and lasted for the same time). Note that in Table 5, each unit from 3 to 6 is a specific cohort. In general, a cohort can be formed by a plurality of groups/units. If we had an extra unit (say, number 7) with a treatment starting in 2005 (beginning) and ending in 2007 (end), that unit would form a cohort with unit 6. We plot below the possible time paths before and after treatments in Figure 4.

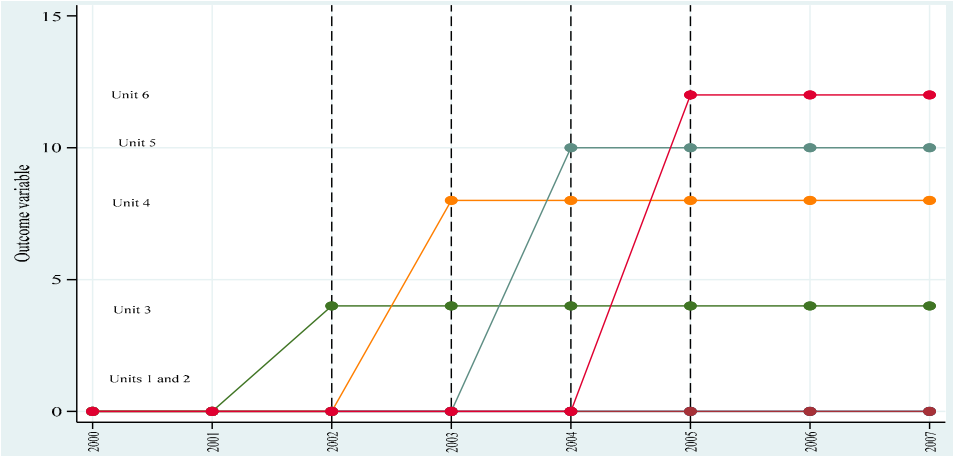


Figure 4. Time paths recorded for six units with two untreated units and four units subjected to different time treatment. In Figure 4, we plot the simulated response variable of each of the six units illustrated in Table 5 before and after each treatment (2002, 2003, 2004, and 2005). Treatments are represented by vertical lines. Data are simulated; so, we put $y = 0$ for units 1 and 2 from 2000 to 2007 (always untreated) as well as for the other units before their treatments, i.e., at least from 2000 to 2002 (excluded). We assume, for the sake of illustration, that treatment causes a positive shift, that is, $\Delta y > 0$.

With heterogeneity of treatment, ATET cannot be estimated by mere application of the TWFE method since the DiD estimate of the treatment effect depend on the choice of the evaluation window. In other words, when groups are treated at different points in time, the assumption about a constant ATET may be violated because the standard DiD estimator estimates an ATET that is common to all groups across time. When groups are treated at different points in time, the assumption about a constant ATET may be violated. Callaway (2022, p. 3) discusses this issue and what are the possible effects of the "bad comparisons" resulting from using for comparison groups that were treated in previous periods.

Different estimators can be employed to overcome the above difficulties. We concentrate on four commonly employed estimators: the extended two-way fixed effects (TWFE), regression adjustment (RA), inverse-probability weighting (IPW), and augmented inverse-probability weighting (AIPW). Some of them fit a model for the response/outcome variable of interest; others fit a model for the treatment or not response and treatment.

Surveys and discussion are, among others, found in Callaway (2022),[7] Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfoeuille (2023), Roth (2022), and Roth et al. (2023).

Before formally introducing these staggered DiD estimators, some general identification assumptions are necessary to ensure the validity of the estimation strategy:

1. Irreversibility of the treatment or staggered treatment: This assumption posits that once units receive treatment, they remain treated throughout the observation period.
2. Parallel Trends Assumption with respect to Never-Treated Units: When we examine groups and periods where treatment is not applied (C = 1), we assume the average potential outcomes for the group initially treated at time g. The group that never received treatment would have followed similar trends in all post-treatment periods $t \geq g$. Then, we have T = (1, ..., S) and g = (2, ..., S) with t $\geq$ g. However, this assumption relies on two important conditions:
   a. There must be a sufficiently large group of units that have never received treatment in our data.
   b. These never-treated units must be similar enough to the units that eventually receive treatment that we can validly compare their outcomes.

   In situations where these conditions are not met, we can use an alternative parallel trends assumption that involves the not-yet-treated units as valid comparison groups.
3. Parallel Trends Assumption with respect to Not-Yet-Treated Units: When we are studying groups treated first at time g, we assume that we can use the units that are not yet treated by time s (where s $\geq$ t) as valid comparison groups for the group initially treated at time g.

It is noteworthy to mention that a common problem to any estimation strategy is the choice of the control units. When there is heterogeneity, the control group can be defined in either way: (a) one can use the units that are never treated; (b) one can use the units not in cohort g and not yet treated by time t, where g is the year of the beginning of the treatment of the cohort. In the following subsections, we will consider a panel of G groups observed at T periods, respectively indexed by the placeholders g and t, which can refer to any group or time period.

For practical understanding, Table A5 reported in Appendix B provides a worked dataset with 13 units and 6 years. Treatment is staggered and irreversible until 2014. Data must be interpreted according to the summary information provided after the table, which contains the dataset and the estimated results obtained by employing the above-mentioned techniques. Notice that, in the worked example, g can be 2011, 2012, or 2013 for the three cohorts.

8.1. The Extended TWFE Method (Wooldridge, 2025)

According to Wooldridge (2025), "there is nothing inherently wrong with using TWFE in situations such as staggered interventions". He proposed an extended TWFE estimator in DiD research design to account for block and staggered treatments based on his finding that the traditional TWFE estimator and a two-way Mundlak (TWM) estimator are equivalent. To show the equivalence, Wooldridge (2025) defines the two-way Mundlak regression as a regression of Y_{it} on a constant term, X_{it} (independent variable of interest), $T^{-1}\sum_{\forall t} X_{it}$ (the unit-specific average over time), and $N^{-1}\sum_{i=1}^{N} X_{it}$ (the cross-sectional average). By the Frisch–Waugh–Lovell theorem and some algebraic calculations, we can see the coefficient of X_{it} is the same as the one in the traditional TWFE regression discussed for the homogenous case in Section 2. Moreover, adding time-invariant variables (Z_{it}) and unit-invariant variables (M_t) does not change the coefficient of X_{it}.

Based on the findings above, Wooldridge (2025) finds that an unbiased, consistent, and asymptotic efficient estimator for heterogeneous ATETs in DiD can be obtained by running a TWFE regression with an inclusion of interactions between treatment-time cohorts and time or, equivalently, by running a pooled OLS regression with an inclusion of panel-level

averages of covariates. This estimator allows for heterogenous effects over time, over covariates, or over both.

As an illustration, we rewrite the traditional TWFE DiD regression of Section 2:

$$y_{ht} = \theta_t + \eta_h + \alpha D_{ht} + \upsilon_{ht}$$

in the extended Wooldridge (2025)'s proposed model:

$$y_{ht} = \eta + \sum_{g=q}^{T} \alpha_g G_{ht} + \sum_{s=q}^{T} \gamma_s F_s + \sum_{g=q}^{T} \sum_{s=q}^{T} \beta_{gs} D_{ht} G_{hg} F_s + \upsilon_{ht}$$

where q denotes the first period in which the treatment occurs, G_{hg} is a group dummy, and F_s is a dummy indicating the post-treatment period ($F_s = 1$ if $t = s$, where $s \in [q, T]$).

In the post-estimation results obtained with Extended TEFE, only the ATT estimates (for each cohort) at the treatment time and for the periods thereafter are shown; this is because Wooldridge (2025) proves that including time dummies and their related interactions for periods prior to the earliest treatment period does not affect the coefficient estimates of interest.

The extended TWFE estimator uses the never-treated group as a control group and has a big advantage: it can be obtained from a very basic regression (pooled OLS) so that most researchers can understand it easily. However, it also has a computational disadvantage (there are many interactions, and therefore the computation of a great number of coefficient estimates is necessary).

8.2. The Regression Adjusted Method (Callaway & Sant'Anna, 2021)

To estimate the ATET for each cohort at each time, the RA, IPW, and AIPW estimators transform the estimation into a classical two groups and two periods difference-in-differences setup. Thus, these techniques restrict the data to an estimation sample with only two groups and only two periods based on the values of g and t. As for the two groups, one group includes all observations in cohort g; the other group includes untreated observations not in cohort g, (control group). For the two periods, one period is the data in time t; the other period is a period when cohort g is not treated (base-line time).

The estimation procedures differ in the way control groups are identified. A possibility is to use the units that are never treated as the control group. An alternative is to use as the control group the units not in cohort g and not yet treated at time t.

RA uses the data of the never-treated control group to estimate the information about the effect of the treatment on the outcome/response variable of the treated groups. Therefore, we have as many benchmark (pre-treatment) years (i.e., g–1 periods) as there are years with a new treatment and as many benchmark/control groups as there are never-treated groups. Ra computes ATET for each cohort and time starting from each t before the treatment (g–1) of each treated cohort.

8.3. The Inverse Probability Weighting Method, IPW, (Callaway & Sant'Anna, 2021) and the Augmented IPW (Callaway & Sant'Anna, 2021)

The IPW estimates the probability that the observation in the benchmark group belongs to the treated group to estimate the untreated differences. IPW computes ATET for each cohort and time starting from each t before the treatment (g–1) of each treated cohort. Yet IPW first builds a logistic regression model to estimate the probability of the exposure observed to the treatment for a particular unit/group and uses the predicted probability as a weight in the subsequent analyses. An extended discussion of the Inverse Probability Method is beyond the scope of this review. For an introduction, the reader is

referred to Chesnaye et al. (2022). Eventually, the AIPW estimator combines the RA and IPW estimators.

9. DiD with Complex Data Structure: Clustering and Spatial-Temporal Dependence

Inference and estimation are closely linked. Once we estimate the causal estimand, we want to know how uncertain our estimate is and test hypotheses about it. In this section, we highlight some common challenges and proposed solutions for inference in Did.

Whether the data arise from repeated measures or from repeated cross-sections, data used in diff-in-diff studies are usually not independently and identically distributed (i.i.d.). For example, we often have hierarchical data, in which individual observations are nested within larger units (e.g., individuals in a US state) or longitudinal data, in which repeated measures are obtained for units. In both of these cases, it is assumed that i.i.d. data will result in standard errors that are too small. Also, as previously discussed in Section 2.2, when the assumption of reciprocal independence among the individuals under study is violated, the SUTVA assumption is dramatically violated as well, leading to identifiability issues with the actual treatment effect (S. Sun & Delgado, 2024).

Recall that in Equation (1), there were no cofactors and assume now that we have two subperiods (pre and post treatment). Since treatment is homogeneous (there is no staggered treatment), we may think that we face the panel data version of the 2×2 TWFE model analysed in Sections 1 and 2. However, things may not be so, and two new issues may emerge:

a. Data showing a grouping or clustering structure;
b. Data exhibiting complex dependence generated by spatial and temporal relationships.

9.1. Clustering

When the dataset has a group structure, data are unlikely to be independent across observations. For example, if our data are the individual test scores of students belonging to different classes of different schools, students' tests of pupils belonging to the same class tend to be correlated across each other simply because students are exposed to the same factors: same teachers, same textbooks, same school equipment, etc.). Likewise, individual consumption or work data within a regional zone in a country can be correlated because the consumers/workers in each zone share the same cultural tradition and work/consumption habits. If we call g the group (cluster) of the observations and assume that the treatment is administered to some groups in the homogenous form (see Section 6), the above equation rewrites as follows:

$$y_{hgt} = \beta_0 + \beta_1 \times D1_t + \beta_2 \times D2_{gt} + \beta_3 \times \Big[D1 \times D2\Big]_{gt} + \varepsilon_{hgt}$$

where h is the individual observation, g is the group (cluster) to which the observation belongs, and t is the time indicator. As one can see in the above equation, we have maintained the common intercept. To emphasis the presence of group correlation, the equation ca be rewritten in terms of a random effect model as follows:

$$y_{hgt} = \gamma + \beta_1 \times D1_t + \beta_2 \times D2_{gt} + \beta_3 \times [D1 \times D2]_{gt} + \delta_{hgt}$$

where y_{hgt} is the h-th observation in the g-th group, while γ is an unobserved overall mean (common intercept). The term $\delta_{hgt} = \alpha_g + \varepsilon_{hgt}$ is a random effect term given by the sum of an unobserved random effect shared by all individuals in group g but varying across groups $\left(\alpha_{g)}\right.$ and an unobserved and unstructured noise term uncorrelated in time and across both groups and individuals (ε_{hgt}). For the model to be identified, α_g and ε_{hgt} are

assumed to have expected values of zero and to be uncorrelated among themselves and over time.

If we postulate that the above-mentioned group correlation across individual data exists, the covariance in the error term of two observations drawn from observation in the same group (cluster) in each t is not zero. Following Angrist and Pischke (2009, p. 309), we may write with respect to the original model that the covariance is as follows:

$$E\left[\varepsilon_{hg}\varepsilon_{jg}\right] = \rho_{\varepsilon}\sigma_{\varepsilon}^{2} > 0 \; \forall \; h \neq j \; in \; each \; g \; and \; \forall \; t$$

where ρ_{ε} is the intraclass correlation coefficient of the original error term (ICC).

Then, the question is how to define ICC. In the light of the random effect version of Equation (1) the ICC writes as follows:

$$\rho_{e} = \frac{\sigma_{\alpha}^{2}}{\sigma_{\alpha}^{2} + \sigma_{\varepsilon}^{2}} > 0$$

This ICC is always non-negative, allowing it to be interpreted as the proportion of total variance that exists "between groups." This ICC can be generalised to allow for covariate effects, in which case the ICC is interpreted as capturing the within-class similarity of the covariate-adjusted data values. Recall that this expression can never be negative (unlike Fisher's original formula), and therefore, in samples from a population which has an ICC of 0, the ICCs in the samples will be higher than the ICC of the population.[8]

9.2. Serial Correlation

In the 2-year framework of DiD typical of the 2 × 2 model of Section 2.1, serial correlation (i.e., the tendency of a variable and a lagged version of itself, for instance, a variable at times t and at $t-1$, to be correlated with one another over periods of time) is not a real problem. Yet DiD analysis is often performed using data which have a time dimension greater than two. Although the sample can still be divided into two "treatment periods" (D = 0 and D = 1), each period can be composed by more time observations (annual, quarterly, etc.) of the response variable and cofactors, if present. Therefore, the serial correlation problem cannot be ignored. Moreover, if we have a panel data structure, we also have individual effects to consider alongside time effects.

Rewrite the above basic OLS equation in panel data form and use h for individuals, g for groups and t for time (say year). Recalling that the treatment is imposed at the group level, we have (with no cofactors) the following:

$$y_{hgt} = \beta_{g} + \lambda_{t} + \delta \times D_{gt} + \varepsilon_{hgt}$$

where:

- y_{hgt} is the status of the response variable of individual h in group g in time t;
- β_{g} is the time-invariant group effect;
- λ_{t} is the group-invariant time effect;
- $D_{gt} = [D1 \times D2]_{gt}$ is the interaction dummy representing the treatment state in post-treatment period;
- ε_{hgt} reflects the idiosyncratic variation in the response variable across individuals, groups and time.

If we assume that some of the components of ε_{hgt} are common to individuals in the same group and time (a tax imposed in some regions for some years; a regional business cycle prevailing in some areas; a pandemic affecting only some specific regional areas and lasting for some years; etc.), we may think of ε_{hgt} as the sum of two components. One is a

group-year shock (v_{gt}) and the other is an idiosyncratic individual component (η_{hgt}) such that the above estimand rewrites as follows:

$$y_{hgt} = \beta_g + \lambda_t + \delta \times D_{gt} + v_{gt} + \eta_{hgt}$$

Following Angrist and Pischke (2009, p. 317), we assume the following:

$$E\left[v_{gt}\right] = E\left[\eta_{hgt}|g,t\right] = 0$$

Group- and time-specific random effects generate a clustering problem that affects statistical inference. In a 2×2 framework (two years and two groups), we have no way to distinguish the difference-in-differences generated by a policy change from the difference-in-differences due to the fact that the response variable in a group (the treated) is merely subject to some cyclical path when the other (control) is not. The common pre-treatment parallel trend assumption may fail.

The solution suggested by Angrist and Pischke (2009, p. 317) is to increase the time and group dimension of the sample (more years and more groups). Actually, increasing the time dimension of the sample is a solution only if we are prepared to assume that v_{gt} is not plagued by serial correlation, which is hard to maintain, particularly with economic data: unemployment in one region (group) is most likely related to previous unemployment in that region (group). A suggested correction can be the clustering of standard errors by groups only, and not by groups and time (passing the clustering buck one level higher).[9] Whether or not this solves the problem is still controversial because clustered standard errors are not robust to any sort of heteroskedasticity or serial correlation (personal view).

Yet the great advantage of having many time periods (say, many years) is that the presence (and the order) of serial correlation for the response variable can be tested by employing a test for serial correlation with panel data. Indeed, the independent variable of interest in DiD estimation (e.g., the passage of a law in the very well-known Bertrand et al., 2004 example) may itself be very serially correlated, which will exacerbate the bias in standard errors. I consider it advisable to run various tests for serial correlation in fixed-effects panel data regression models, particularly when there are a small number of time periods relative to groups/clusters.

9.3. Spatial Dependence

When the data are georeferenced (i.e., each individual is uniquely identified by a pair of coordinates) or are organised according to a geographical/spatial/lattice/areal structure (e.g., individuals belonging to administrative regions), the independence assumption may be violated due to the potential spillover (or contagion) effect given by the spatial proximity (Elhorst, 2010). Spatial econometric models can easily deal with spatial interactions and spillover effects among units by extending the classical regression models to include spatial lagged terms. Spatial lags can be determined either by the neighbourhood or by the physical distance and can be applied to either dependent variables, covariates or random effects. Under this spatial econometric setting, Qiu and Tong (2021) combines difference-in-difference estimator and spatial regression models into a two-period spatial DiD hedonic framework. The causal regression model is then specified as follows:

$$y_{ht} = \rho W y_{ht} + \beta_1 D1_{it} + \beta_2 D2_{ht} + \beta_3 [D1 \times D2]_{ht} + u_{ht}$$

$$u_{ht} = \lambda W u_{ht} + \varepsilon_{ht}$$

where W is a row-standardised $n \times n$ spatial weighting matrix containing information on the spatial relationship between observations, ρ and λ are the spatial parameters which

measure the strength of the spatial dependence in the dependent variable and error term, respectively. y, β_1, β_2, β_3 and ε_{it} are the usual regression terms previously introduced. While the interpretation of the marginal effects for continuous variable is the same as in the classical cross-sectional DiD model, the interpretation of marginal effects for treatment effects including the spillover treatment effects are different (see Section 2.3 of Qiu & Tong, 2021 for an analytical discussion on the new interpretations of the causal effects). In fact, individuals in the control group can also be affected by the treatment through treated unit houses due to spatial and/or social interactions. Alternative specifications of the above spatial regression model can be found in the works of, among others, Delgado and Florax (2015) and S. Sun and Delgado (2024). In particular, in the work of Delgado and Florax (2015), the authors consider a local spatial DID model able to explicitly capture the effect on an individual that comes from the treatment of their neighbours, while in S. Sun and Delgado (2024), the authors expand the dynamic treatment DID estimator by Callaway and Sant'Anna (2021) to a spatial setting with spillovers among units.

10. The Most Relevant Issues Discussed in This Review and Some Further Research Directions

DiD is at the core of a recent revolution in empirical economics because it aims at "discovering" if a time contingent causal-effect relationship (a post hoc, ergo propter hoc relationship) between a response variable and a treatment/event is statistically consistent with the data. Angrist and Pischke (2010) convincingly describe DiD as "probably the most widely applicable design-based estimator."

In this review, we have presented the DiD method as a suitable approach for estimating causal effect relationships. Nevertheless, the efficacy of DiD is contingent upon a comprehensive array of assumptions concerning the data generation process and its underlying statistical characteristics. Practitioners must be aware that, possibly more than with other estimation procedures, in numerous practical economic applications DiD may not constitute a design-based credible estimation method. This may be given to the likely absence of a genuinely randomized experimental design, to the absence of robust testing procedure for the parallel trend assumption, to the violation of the SUTVA conditions, and to the many other pitfalls that characterizes numerous instances of actual DiD applications. Clearly, with the exception of the parallel trend problem the majority of the methodological drawbacks that are addressed in this review pertain to aspects that are not unique to DiD, and as such, they may not garner the attention of scholars. In general, they are derived from standard regression analysis, particularly when the structure of the dataset is panel data with more than two periods (i.e., one pre-treatment time and one post-treatment time) and two units (i.e., one treated and one untreated unit). Furthermore, given DiD's regression representation, it frequently lacks the capacity to inherently offer more compelling evidence of a causal effect than that provided by regression analysis itself (Kahn-Lang & Lang, 2020, p. 613). Consequently, it is imperative to acknowledge that the same regression issues that plague more conventional regression analysis can resurface in DiD applied studies. In this review, we have placed particular emphasis on issues pertaining to the model's proper specification and the orthogonality of the response variable of interest to the error term, conditional on the controls. We have also emphasized in various sections that specific problems characterize the application of numerous variants of DiD. In what follows, as a way of informal general warning, we single out some of the specific problems that might affect applied DiD studies.

- As in many causal inference procedures, DiD relies on strong assumptions that may be difficult to test. The key assumption (parallel trends) is that the outcomes of the treated and comparison groups would have evolved similarly in the absence of treatment

(under a *vis inertiae* as alluded in the title). Yet, even in simple 2 units and 2 periods case statistical tests have low power, and the issue becomes more complicated in the multi-unit and multi-period cases. The search for the existence of parallel trends might became a search for the Arabian Phoenix since it requires elaborated statistical tests. The simple graphical appearance of a commune time path of mean realizations in the pre-treatment period might be a misleading suggestion of the perpetuation of a similar potential parallel trend path in the post treatment periods (when counterfactuals cannot be observed);

- Therefore, without a true randomized experiment, tools like DiD do not broaden the range of "natural experiments" we can use to identify causal effects.
- Even in the case of true randomization, SUTVA problems (i.e., the so-called spill-over effects across treated and untreated unites) might plague estimations and make it difficult to identify a DiD model that consistently estimate ATET (which requires unique potential outcome for each individual under each exposure condition).
- Often the interpretation of the role of covariates in DiD estimates is difficult and, sometimes, even what a covariate is might be controversial. In fact, DiD does not require the treated and comparison groups to be balanced on covariates, unlike in cross-sectional OLS studies. Thus, a covariate that differs by treatment group (i.e. that is group specific) and is associated with the response is not necessarily a confounder in DiD. Only covariates that differ by treatment group and are associated with outcome trends are confounders in DiD as these can be the ones that violate the identification assumptions because they are correlated with both the response and the treatment. Since confounders can bias the estimated treatment effect by violating the parallel trends assumption, practitioners should pay extra care in evaluating whether a factor can be considered as a confounder.
- In this review we have discussed linear basic OLS version of DiD application, without discussing any issue related to the functional form. Yet, there may be cases where it can matter whether the "correct" model is a linear probability model, probit or logit, since they may assume different counterfactuals. Determining that two groups would have experienced parallel trends requires, first of all, a justification of the chosen functional forms for the adopted model.

Although not explicitly addressed in the review, it is noteworthy to mention some further issues raised by recent literature concerning (a) applications with small-sized treatment groups; (b) the construction/definition of the counterfactual group; and (c) the potential high dimensionality and non-linearity of some empirical relationships.

First, let us consider the issue on few treated units, also denoted as the rare events data issue, which poses the challenge of getting reliable inference in settings where standard inference based on asymptotic theory are unrealistic even with large total sample sizes (Alvarez et al., 2025). For instance, to address this problem, Pernagallo and Vitali (2025) provides estimates based on a logistic regression for rare events in the context of real estate investment decisions for survey data with a strong unbalance between purchasing and non-purchasing respondents. To have a general perspective, the readers can refer to the papers by Alvarez and Ferman (2020) and (Alvarez et al., 2025), which review heuristics modifications to improve the finite-sample performance of some existing methods and address the role of spatially correlated errors in determining the asymptotic properties of the DiD estimator for few treated situations.

Regarding the second and third issues, modern statistical machine learning (ML) algorithms can represent effective tools that researchers interested in causal inference should include in their toolkit (Ahrens et al., 2021). As for the counterfactual matter, readers can refer to the Machine Learning Control Method proposed by Cerqua et al. (2022, 2023)

to forecast the counterfactual values with machine learning (hence, to estimate individual, average, and conditional average treatment effects) when a credible control group is not available (Cerqua et al., 2024). Eventually, when considering high dimensional or non-linear settings, of particular relevance is the Double-Debiased ML for obtaining reliable inferential results on the parameters of interest in highly-dimensional models (Ahrens et al., 2025; Chernozhukov et al., 2016, 2017, 2022) and for the estimation of causal treatment effects (Chernozhukov et al., 2016, 2017, 2018) also in combination with other techniques like causal mediation analysis (Farbmacher et al., 2022). The importance of double/debiased ML in casual inference ground on its ability to provide flexible da-ta-driven methods to estimate the average treatment effects under several settings (e.g., binary, multiple and continuous treatments) and to allow for a flexible estimation of heterogeneous effects and of treatment assignment rules (Knaus, 2022). For further in-sight on the use of ML algorithms for causal inference, such as tree-based algorithms (Souto & Neto, 2025) or LASSO-based models (Shortreed & Ertefaie, 2017), we refer the reader to the recent comprehensive review by Yao et al. (2021).

Another line of investigation that should receive greater attention is represented by the possible mixed or hierarchical structure of the data set whose implication for the study of the treatment-effect relationships are not (to our knowledge) extensively investigated. Mixed models may improve DiD analysis by incorporating random effects for panel and clustered data structure (e.g. students within schools, patients within hospitals, taxpayers in different regions) that are commonly used in policy analysis where the *simultaneous estimation of the multiple sources of response variation may prove as important as the very ATET estimation*. Moreover, mixed models may provide more accurate standard errors and more reliable estimates of the treatment effect coefficients.

It is finally essential to emphasize that it is not in the nature of DiD to enable the investigation of the underlying causes of the observed differences in response levels between the treated and control groups in the pre-treatment period (a question that may concern economists and policy makers). For example, DiD will not help practitioners to understand the reasons for the higher unemployment rate in region A compared to region B, both before and in the presence of treatment, *regardless* the modification the treatment has caused to the unemployment in treated groups. Neither can we understand why the experimental design may have been unsuccessful (absence of significant ATET on unemployment in treated groups) and how to evaluate whether the absence of statistically significant differences in the post-treatment periods among the treated and untreated units is responsible for the failure.

Everything notwithstanding, it is important to conclude that even basic DiD can contribute to overcoming some of the identification difficulties associated with more traditional OLS-based methods, such as the exogeneity issue. Even this sole advantage would represent a significant "productive asset" possessed by DiD and a great advantage over other empirical procedures. This alone would justify the enormous body of research conducted by means of DiD methods.

11. Some Examples of DiD Applications

In this section, we present a selection of DiD applications in which the authors study the behavioural responses of various outcome variables to events such as new taxation, energy prices and regulation reforms, as the latter are introduced in various markets/sectors. The selection does not simply reflect the preferences of the authors of the present review. It is also motivated by the methodological content of the quoted papers, particularly when the authors of the papers employ some variants of the basic DiD techniques reviewed here. Therefore, the reading of the original papers is strongly recommended because it represents

a necessary integration of this manuscript, and the readers should bear in mind that the following sections do not substitute a sound studying and understanding of the original papers (*Dixit et salvavi animam meam*).

11.1. The Elasticity of Taxable Income (Feldstein, 1995)

A long-standing problem of applied public economics/finance is as follows: How do we estimate the total welfare loss associated with taxes, in particular with income taxes? The modern literature on taxes and labour supply discusses two main alternatives:

1. The structural approach (closer to the "old" theoretical analysis of labour responses to income taxation), which separately accounts for each of the potential responses to taxation (intensive and extensive) and then aggregate.
2. The DiD approach first proposed by Feldstein (1995), which aims estimate the elasticity of taxable income with respect to the net-of-tax rate and claims that this elasticity is a sufficient statistic for calculating the possible deadweight loss of income taxation.

Feldstein's (1995) paper is the starting point of a whole new literature on this topic.[10] He argued in favour of the idea that focusing on labour supply misses margins at which individuals might also respond to taxation. The latter may be (a) the intensity of work (effort), training, occupation and career choices; (b) the form and timing of compensation; or (c) tax avoidance and tax evasion. Then, Feldstein diverted the research's attention from pure labour supply response to income taxation to the analysis of the effects of income taxation on the entire level of income as a tax base (the taxable income). Moreover, he argued that the elasticity of taxable income is a sufficient statistic for the empirical study of the effects of income taxes.

To correctly identify the above elasticity, he employs a DiD method and used a Treasury Department panel of more than 4000 taxpayers to estimate the sensitivity of taxable income to changes in tax rates on the basis of a comparison of the tax returns of the same individual taxpayers before and after Reagan's 1986 tax reform. Therefore, in Feldstein's paper, one will find neither the equivalent of Equation (1) in Section 2, nor the test statistics recommended for parallel trend, anticipation effects, etc.

To describe the results of the paper, we follow Feldstein and define the following:

- TI = taxable income (defined as an aggregate measure of income from various sources);
- τ = proportional income tax rate.

Then, TI depends on the tax rate τ and the net-of-tax income, NTI, is as follows:

$$NTI = (1 - \tau)TI$$

When the tax rate changes, the taxable income may change as a result of some behavioural reaction of the taxpayer. A measure of the reaction is the elasticity of the taxable income. We can calculate the elasticity of taxable income with respect to the net-of-tax rate $(1 - \tau)$ by totally differentiating the TI:

$$dTI = \frac{\partial TI}{\partial(1-\tau)} d\tau$$

The above rewrites as follows:

$$dTI = \underbrace{\left[\frac{\partial TI}{\partial(1-\tau)} \frac{(1-\tau)}{TI}\right]}_{\eta_{TI,(1-\tau)}} TI \frac{d\tau}{(1-\tau)}$$

Then, the problem is how to identify the elasticity of taxable income $\eta_{TI,(1-\tau)}$ since (the conventional view apparently is) that the tax rate is endogenous to choice of income (reverse causality), whereas for empirical purposes, we need exogenous variation in tax rates to identify the elasticity. This is where DiD somehow enters the analysis.

Feldstein's goal was to estimate causal effect of (net-of-tax rate on taxable income. To identify the elasticity of taxable income, he used the variation in marginal tax rates (MTRs) generated in the USA by the TRA1986 reform of President Reagan. Then, since changes in MTRs differ between taxpayers according to tax brackets to account for initial differences in taxable income, the author compares the change in taxable income in one income group (say A) to the change in taxable income in another income group (say B).

The empirical DiD specification used to estimate the ATET generated by the tax reform is the following:

$$y = \beta_0 + \beta_1 Post + \beta_2 Treatment + \delta\,[Post \times Treatment] + \varepsilon$$

where:

- *Post* is the dummy variable for the reform period (1 if after-reform and 0 pre-reform);
- *Treatment* is the dummy identifying the treated income group (Treatment = 1) and the untreated group (Treatment = 0);
- Post × Treatment is the DiD variable given by the interaction between the above two;
- δ is the coefficient of interest which measures ATET;
- and ε is the classical error term.

Yet, y is not a measure of labour supply but a percentage change in the tax return, i.e., the Adjusted Gross Income (AGI) before and after the reform for various subsets of taxpayers. According to Feldstein (1995, p. 555):

> *The use of tax return data rather than of a household survey permits analysing the response of taxable income as a whole and not just of labour force participation and working hours. A panel, in which each individual is observed both before and after the change in tax rates, permits a "differences-in-differences" form of estimator that identifies the tax effect in a way that is not available with a single year's cross section.*

Indeed,

$$\hat{\delta} = ATET_{\text{Tax Reform}} = \left(TI_{Post1986,A} - TI_{Before1986,A}\right) - \left(TI_{Post1986,B} - TI_{Before1986,B}\right)$$

In the above equation, the first difference controls for time-invariant differences in the earning potential of high-income and low-income groups, assumed to be A and B. The second difference controls for time effects that affect the two groups identically. The difference with respect to DiD of Section 2 is that there is no untreated control group in the model, but treatment and control groups differ in the intensity of treatment (poor taxpayers are a control for rich taxpayers, and vice versa).[11]

To satisfy the DiD identifying assumptions discussed in Section 1, Feldstein had to assume the following:

- The income growth rate is the same for all income earners (medium, high and highest tax brackets) in the absence of the treatment ("parallel trend assumption").
- The taxpayers cannot adjust their income in 1985 (last year before reform) to "choose" their change in tax rate through TRA1986 ("no selection into treatment" and no anticipation effect).
- The comparison of taxpayers that vary in the intensity of treatment (instead of comparing taxed to untaxed taxpayers) is legitimate. Implicitly, he needs to assume that

the elasticity of taxable income is constant in income, i.e., the same across all income groups. This last assumption will reappear in other papers.

The main target of Feldstein' paper was not the pure estimation of the ATET of the model but the use of the estimated coefficient to estimate the causal effect of (net-of) tax rate on taxable income. His general result is that the larger the increase in the net-of-tax rate (i.e., the decrease in marginal tax rate), the larger the increase in income declared for tax purposes. He reported the following elasticities (Feldstein, 1995, p. 565):

The results show the following:

- Estimates of the elasticities are high estimates, ranging from 1 to 3.
- The so-called Laffer rate, i.e., the rate that maximises the tax revenue, changes with the elasticity and corresponds to $1/(1+\epsilon)$.
- The USA are on the wrong side of the Laffer curve (excessive levels of income tax rates).

We now consider how he employed DiD to compute the above elasticities. The difference in adjusted taxable income (ATI in column 2) is divided by the difference in net of tax rate (13.04) = 1.10, and so on and so forth. However, more recent estimates at the layers state that these estimates are excessively high.

Feldstein's analysis raises some questions.

- No proper untreated control group is present in the study. Treatment and control groups differ in *the intensity* of treatment.
- An equal elasticity of taxable income across income distribution is assumed. The elasticity of taxable income is likely higher for high-income taxpayers (with more adjustment opportunities).
- Small and unstratified sample: very few high-income taxpayers are included.
- The presence of increasing earning inequality in the US determined for non-tax reasons should be considered.
- The results may be affected by a regression-to-the-mean bias due to classification of treatment groups by pre-treatment income: rich people in year t may tend to revert to the mean in year $t+1$.
- Panel analysis introduces a downward bias in the estimated elasticity if marginal tax rate for rich people decreases.
- It is unclear whether the common trend assumption really holds. Not even the simplest tests are conducted (parallel trends, anticipation effects, etc.).
- Estimated elasticity overestimates welfare loss if the behavioural response involves transfers between individuals.
- The study provides some unclear indications about the effects of changes in MTR on the aggregate income tax yield, but it is silent about taxpayers' behavioural reactions to income taxation in spite of the claim that "The Tax Reform Act of 1986 is a particularly useful natural experiment for studying the responsiveness of taxpayers to changes in marginal tax rates" (Feldstein, 1995, p. 552). The potential role that confounders (likely affected by the treatment) may play in this estimation is completely ignored.

11.2. Top Income Taxation and the Migration Decisions of Rich Taxpayers (Kleven et al., 2013)

The paper reviewed in this section uses DiD to analyse possible income tax-induced migration across countries and tries to estimate the causal relationship between tax rates and migration. It uses a combination of graphical evidence and systematic multinominal regression (DiD with cofactors) and employs synthetic control.[12]

Specifically, Kleven et al. (2013) analyse the effects of top tax rates on international migration of football players in 14 European countries since 1985. They also conduct country case studies and multinomial regressions and find evidence of strong mobility

responses to tax rates, with an elasticity of the number of foreign (domestic) players to the net-of-tax rate around one (around 0.15). The paper shows evidence of sorting effects (low taxes attract high-ability players who displace low-ability players) and displacement effects (low taxes on foreigners displace domestic players).

The research question is as follows: How do tax rates impact "labour" mobility of professional football players in Europe once the three-player limitations was abolished by the Bosman Ruling of 1995? Kleven et al. (2013) claim that to conduct their study, the average tax rate (ATR) is the appropriate tax rate for location decision and that taxpayer considers overall tax burden of location decision (an extensive margin decision).[13]

The paper aims at estimating two key elasticities:

$$\varepsilon_{nf} = \frac{dp_{nf}}{d\left(1-\tau_{nf}\right)}\frac{1-\tau_{nf}}{p_{nf}} \varepsilon_{nd} = \frac{dp_{nd}}{d(1-\tau_{nd})}\frac{1-\tau_{nd}}{p_{nd}}$$

where:

- p_{nd} = total of domestic players in country n;
- p_{nf} = total of foreign players in country n.

The two elasticities represent the percentage variation in the number of foreign (domestic) players in country n with respect to the variation in the net-of-tax rate on foreign (domestic) players in country n.

Kleven et al. (2013, p. 1904) provide graphical cross-country evidence on the relationship between the top earning tax rate and in-migration of foreign players, out-migration of domestic players, and club performance. Each panel consists of two graphs, with the pre-Bosman era (1985–1995) on the left and the post-Bosman era (1996–2008) on the right. In each panel, the authors depict the best linear fit using a univariate regression (with no country weights).

Then, they estimate corresponding elasticities by regressing the log y-axis outcome on the log of the net-of-tax rate (again with no country weights) of those weights. For a country-specific tax reform case study, Kleven et al. (2013, p. 1907) present elasticity estimates for a DiD comparison of the treatment country and the synthetic control country before and after the reform. During the pre-Bosman period, the fraction of foreigners is generally very low and there is no correlation between the fraction of foreigners and tax rates. After the Bosman ruling, the fraction of foreigners is much higher in every country (between 5 percent and 25 percent), and there is a significant negative correlation with the top earning tax rate. The implied elasticity of the fraction of foreigners with respect to the net-of-tax rate is zero pre-Bosman era, but very large at 1.22 (0.45) in the post-Bosman era. Panel B of Figure 1 plots the average fraction of players of a given nationality playing in their home league against the average top earnings tax rate on domestic residents. In the pre-Bosman era, the fraction of players playing at home is very high in all countries (between 90 percent and 100 percent across the entire sample). After the Bosman ruling, the fraction playing at home drops in almost all countries, and the negative correlation with tax rates becomes much stronger. The implied elasticity of the fraction playing at home with respect to the net-of-tax rate was modest pre-Bosman, at 0.09 (0.04), and much higher post-Bosman, at 0.29.

The elasticities are always for foreign players and are obtained from a 2SLS regression of (see the Notes to Table 1 at page 1906 of the original paper)

$$\log(P_{ct}) = e \times \log(1-\tau_{ct}) + \beta \times I(c=T) + \gamma I(t \geq t_0) + \varepsilon$$

instrumented with $I(c = T) \times I(t > t_0)$, where c is country (the treatment country is T, i.e., a synthetic control) and p is the number of foreign players, τ is the top marginal tax rate, t is the year, and t_0 is the year of the reform.

The possible limitations of the analysis are the following

- In the graphical analysis, the elasticities of the average tax rate are not presented for the pre-Bosman period and the Danish case studies because of a lack of individual earnings data before 1996. Similarly, the average tax rate elasticity for Spain is based on the 1996–2003 versus 2004–2008 comparison. It is therefore difficult to conduct a complete comparison study (not even graphical).
- The sample used is limited to a very special category of privileged *migrants* (the well-paid football players whose behaviour is affected by several treatment-related confounding factors). Out-of-sample projections seem problematic.
- The Bosman ruling could have had differential impacts on low-tax and high-tax countries for non-tax reasons. Tax rates may correlate with country size and thus league quality. Better leagues may have benefited more from Bosman ruling.
- Football player contracts are generally signed in advance with respect to the year of the actual transfer and then anticipation effects of the Borman ruling might be present.
- Other factors could have changed from the pre-Bosman to the post-Bosman era that impacted low-tax and high-tax countries differentially.

11.3. Toxic Emissions and the Environment (Zhou et al., 2019; Dong et al., 2022)

Emission trading (buying and selling permissions to pollute the environment by releasing CO2 particles. . .) is supposed to be a market-driven mechanism able to reduce carbon intensity production processes. It has been widely used in western countries, and it has produced debatable results in terms of reduction in TONs of carbon emissions and emission price determination. In 2013, the Chinese government established pilot carbon emission trading programs in seven provinces. The papers discussed in this section conduct an empirical analysis, using a decomposition and DiD approach of the effects of the 2013 environmental policy. The main conclusions are as follows: (1) Overall, China's emission trading pilots have driven a significant decline in the carbon intensity, resulting in an average annual decline of approximately 0.026 tons/10,000 CNY in the pilot provinces. (2) In the sample period, emission trading pilots had a sustained and stable effect on carbon intensity with no time lag. (3) Emission trading pilots reduce the carbon intensity by adjusting the industrial structure. In contrast, energy structure and energy intensity channels have not yet been realised.

Zhou et al. (2019, p. 516) use a Propensity Score Matching (PSM) approach before the implementation of DiD to enhance the selection of the appropriate control group from the untreated provinces. According to the authors, this helped solve possible endogenous problems and ensure that the DiD estimation results were unbiased.

After establishing the control group, the DiD approach was used by Zhou et al. (2019, p. 517) to evaluate the overall effect of emission trading pilots on carbon intensity:

$$Dif\_CI_{it} = \alpha_0 + \alpha_1 Y + \alpha_2 R + \alpha_3 (Y \times R) + \gamma_i + \gamma_t + e_{it}$$

where i denotes provinces and t denotes years. *Dif_CI* denotes first-order differences in the carbon intensity (dependent variable); α is the coefficient of the independent variable with α_3 corresponding to the ATET; γ_i and γ_t represent province-fixed and time-fixed effects, respectively; and e is the random error. Y corresponds to years with the new regulation and R to the regulated (pilot) provinces/units.

Below, we report the results from a table taken from the original paper (their Table 6). Column (1) reports the results of the fixed-effect estimation based on matched data. The coefficient of the variable Y × R was significantly negative. This indicates that implementing the emission trading pilots resulted in an average annual decrease in the carbon intensity of 0.026 tons/10,000 CNY. In addition, column (2) reports the results based on panel data; the results are consistent with column (1), indicating robust estimation results. Columns (3) and (4) show the DID estimation results using non-matched data.

Table 6. Overall effect of emission trading pilots.

Variables	Matched (1)	Matched (2)	Non-Matched (3)	Non-Matched (4)
Y	−0.061 *** (−2.807)		−0.029 ** (−2.073)	
R	−0.036 *** (−2.869)		−0.036 *** (−2.649)	
Y × *R*	−0.026 *** (−2.986)	−0.026 *** (−3.156)	0.038 (0.704)	−0.026 (−1.186)
_cons	−0.062 *** (−2.850)	−0.096 *** (−4.734)	−0.072 *** (−3.088)	−0.111 *** (−5.370)
Province/Year fix effects	Yes	Yes	Yes	Yes
A-R^2	0.196	0.008	0.170	0.001

Notes: *t* values are shown in brackets; ***, ** indicates statistical significance at 1% and 5% levels, respectively. Source: Zhou et al. (2019, p. 516).

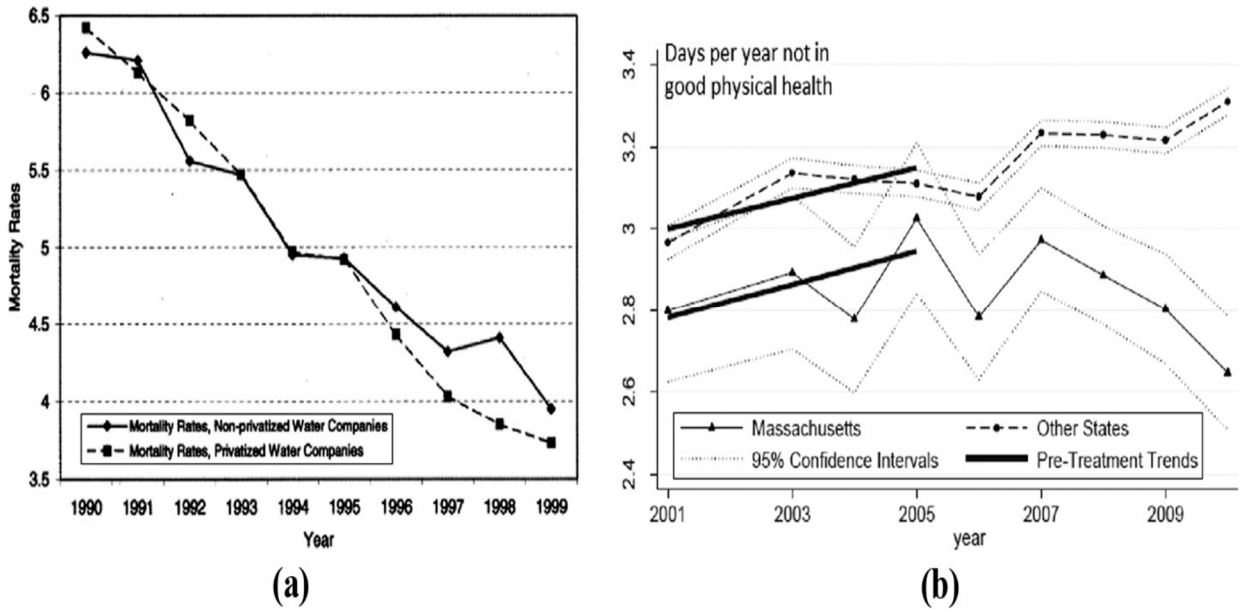
The pilots have no significant effect on the downward trend in carbon intensity. This may be because the control group, before matching, included provinces in the western regions. The western regions have experienced a rapid drop in carbon intensity, weakening the significance of pilot effects on reducing carbon intensity. Using the PSM approach to remove the unsuitable provinces from the control group can ensure the DID approach generates unbiased estimation results. The authors interpret their DiD results as an indication that the adoption of the emission trading reform has effectively reduced China's carbon intensity.

Similar results are provided by Dong et al. (2022, p. 12), who also estimate the effects of the infrastructure transformation and greenhouse gas emission performance improvement. Their DiD results show that information infrastructure exerts significant emission reduction compression in cities with large size, advanced digital economy, and leading economic status, while its impact on greenhouse gas emission performance drops in other cities.

11.4. Regulation, Privatisation, Management (Galiani et al., 2005)

Galiani et al. (2005) and Gertler et al. (2016) study the impact of privatising water services on child mortality in Argentina. Using a decade of mortality data and comparing areas with privatised (treatment) and non-privatised water companies (control), they observe similar pre-reform (pre-1995) trends that support the parallel trend assumption of their DiD work (plot a of the figure reproduced below).

The authors find a statistically significant reduction in child mortality in areas with privatised water services (panel a). Panel (b) provides another example, with data on a health variable measuring days in which patients are not in good health conditions. This variable is recorded both before and after the 2006 Massachusetts reform, as illustrated by Courtemanche and Zapata (2014). A more formal approach to provide support for the parallel trend assumption was followed by conducting a placebo regression, which applies the DiD method to the pre-reform data itself. There should then be no significant "treatment effect". When running such placebo regressions, one option is to exclude all post-treatment observations and analyse the pre-reform periods only (if there are enough data available).



A line of investigation similar to the one quoted above is provided by Schnabl (2012). He studies the effects of the 1998 Russian financial crisis on bank lending and uses two years of pre-crisis data for a placebo test, whereas an alternative is to use all data and add to the regression specification interaction terms between each pre-treatment period and the treatment group indicator(s). The latter method is used by Courtemanche and Zapata (2014), studying the above Massachusetts health reform. A further robustness test of the DiD method is to add specific time trend terms for the treatment and control groups, respectively, in equations like our Equation (1) of Section 2.1, and then check that the difference in trends is not significant (see Wing et al., 2018, p. 459). A general review of the above papers is presented by Fredriksson and de Oliveira (2019).

Author Contributions: Conceptualisation, B.P.B. and P.M.; methodology, B.P.B. and P.M.; software, P.M.; investigation, B.P.B.; resources, B.P.B. and P.M.; writing—original draft preparation, B.P.B. and P.M.; writing—review and editing, B.P.B.; supervision, P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All the data used in the examples are included in Appendices A and B of this paper.

Acknowledgments: This research work was partially supported by the University of Milan-Bicocca. The authors want to acknowledge the contribution of many departments' colleagues to a previous working paper version of this review and thank three anonymous referees for their help, encouragement and indications. The authors are the only responsible for any remaining errors and mistakes.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper. This research was conducted independently, and no financial, personal, or professional relationships could be construed as influencing the findings or interpretations presented in this study. Additionally, there are no competing interests related to the funding, data collection, or analysis that could affect the integrity of this research.

Appendix A. An Example with Easy Visualisation of the Dataset

Assume there are three randomly selected groups of consumers, A, B, and C, whose consumption is recorded from 2000 until 2006. For simplicity, each group is composed of five people. At the beginning of 2003, a treatment (e.g., a commodity tax reduction) is

introduced by the local governments where A and B live, and it is maintained until, and including, 2006. Therefore, we are dealing with a two-period model: the first period/phase (3 years) without any treatment and the second period/phase (4 years) with the treatment affecting some unites. Let us assume that there are barriers that do not permit consumers C to move to a locality with lower taxes.

Table A1 is a basic example of data stuck in panel data form. The first column shows years; the second shows the response variable (the first pedis refers to the individual; the second to her/his group; the third to the year); the rest of the columns show the two dummies and their product. Groups A and B received the treatment (tax reduction) all in the same year and group C was never treated.

The case of Table A1 (same treatment periods for each treated units) can be called a staggered treatment effect with homogeneity. Recall that the alternative is the case of a staggered treatment effect with heterogeneity (cohorts treated from different initial moments until the end of the sample's recorded period).

Table A1. Example of panel dataset structure for DiD with more than two years and units.

Year	***y*** **Consumption Expenditure of an Individual Belonging to a Group Recorded in a Year**	***D*1 = Time Period Treatment Dummy** **0 If It Is a Year with No Treatment** **1 If It Is a Year When Treatment Existed**	***D*2 = Treatment Application** **0 If the Individual Is Never Treated** **1 If the Individual Is Treated (Sooner or Later)**	**TREATMENT = *D*1 × *D*2** **0 Will Indicate the Individual Is Not Affected by the Tax Policy** **1 Will Indicate That in a Certain Year the Individual Is Affected by the Tax Policy**
2000	y_{1A2000}	0	1	0
2001	.	0	1	0
2002	.	0	1	0
2003	y_{1A2003}	1	1	1
2004	.	1	1	1
2005	.	1	1	1
2006	y_{1A2006}	1	1	1
2000	y_{2A2000}	0	1	0
2001	.	0	1	0
2002	.	0	1	0
2003	y_{2A2003}	1	1	1
2004	.	1	1	1
2005	.	1	1	1
2006	y_{2A2006}	1	1	1
		…….		
2000	y_{1C2000}	0	0	0
2001	.	0	0	0
2002	.	0	0	0
2003	y_{1C2003}	1	0	0
2004	.	1	0	0
2005	.	1	0	0
2006	y_{1C2006}	1	0	0
		…….		
2000	y_{5C2000}	0	0	0
2001	.	0	0	0
2002	.	0	0	0
2003	y_{5C2003}	1	0	0
2004	.	1	0	0
2005	.	1	0	0
2006	y_{5C2006}	1	0	0

Table A2. Example of a DiD dataset.

Units ID	TIME	TRET = ($D1 \times D2$)	Response Variable
1	1	0	0.5
1	2	0	0.5
1	3	0	0.5
1	4	0	0.5
1	5	0	0.5
1	6	0	0.5
1	7	0	0.5
1	8	0	0.5
1	9	0	0.5
1	10	0	0.5
2	1	0	1
2	2	0	1
2	3	0	1
2	4	0	1
2	5	1	2
2	6	1	2
2	7	1	2
2	8	1	2
2	9	1	2
2	10	1	2
3	1	0	2
3	2	0	2
3	3	0	2
3	4	0	2
3	5	1	4
3	6	1	4
3	7	1	4
3	8	1	4
3	9	1	4
3	10	1	4

Table A3. Dataset used for the worked example #1 Section 4.

Consumers' Id	Time	Consumption EUR	*D*1	*D*2
1	2010	12	0	1
2	2010	9	0	1
3	2010	13	0	1
4	2010	14	0	1
5	2010	15	0	1
6	2010	13	0	0
7	2010	14	0	0
8	2010	13	0	0
9	2010	16	0	0
10	2010	15	0	0
1	2011	15	1	1
2	2011	17	1	1
3	2011	19	1	1
4	2011	18	1	1
5	2011	22	1	1
6	2011	13.5	1	0
7	2011	14	1	0
8	2011	15	1	0
9	2011	15.5	1	0
10	2011	14.4	1	0

Table A4. Dataset used for the worked example #2 in Section 4 (Parallel trend).

Consumers' Id	Time	Consumption EUR	*D*1	*D*2
1	2009	11	0	1
1	2010	12	0	1
1	2011	15	1	1
2	2009	8.6	0	1
2	2010	9	0	1
2	2011	17	1	1
3	2009	12.5	0	1
3	2010	13	0	1
3	2011	19	1	1
4	2009	13	0	1
4	2010	14	0	1
4	2011	18	1	1
5	2009	14	0	1
5	2010	15	0	1
5	2011	22	1	1
6	2009	12	0	1
6	2010	13	0	0
6	2011	13.5	1	0
7	2009	13.7	0	0
7	2010	14	0	0
7	2011	14	1	0
8	2009	12.7	0	0
8	2010	13	0	0
8	2011	15	1	0
9	2009	14.9	0	0
9	2010	16	0	0
9	2011	15.5	1	0
10	2009	14.7	0	0
10	2010	15	0	0
10	2011	14.4	1	0

Appendix B. Dataset for Non-Homogeneous DiD Estimation

This appendix contains data and estimations of heterogenous DiD estimation discussed in Section 8.

Table A5. Example of dataset for staggered heterogeneous DiD treatment.

ID	Year	Consumption	*D*1	*D*2	TRET	First Year of Treatment
1	2009	11	1	0	0	2011
1	2010	12	1	0	0	2011
1	2011	15	1	1	1	2011
1	2012	14.8	1	1	1	2011
1	2013	15.8	1	1	1	2011
1	2014	17	1	1	1	2011

Table A5. *Cont.*

ID	Year	Consumption	*D*1	*D*2	TRET	First Year of Treatment
2	2009	8.6	1	0	0	2011
2	2010	9	1	0	0	2011
2	2011	17	1	1	1	2011
2	2012	18	1	1	1	2011
2	2013	18.8	1	1	1	2011
2	2014	19.1	1	1	1	2011
3	2009	12.5	1	0	0	2011
3	2010	13	1	0	0	2011
3	2011	19	1	1	1	2011
3	2012	19.8	1	1	1	2011
3	2013	21	1	1	1	2011
3	2014	22	1	1	1	2011
4	2009	13	1	0	0	2011
4	2010	14	1	0	0	2011
4	2011	18	1	1	1	2011
4	2012	19.1	1	1	1	2011
4	2013	22	1	1	1	2011
4	2014	21.8	1	1	1	2011
5	2009	14	1	0	0	2011
5	2010	15	1	0	0	2011
5	2011	22	1	1	1	2011
5	2012	21.9	1	1	1	2011
5	2013	22.2	1	1	1	2011
5	2014	22	1	1	1	2011
6	2009	12	0	0	0	Never treated
6	2010	13	0	0	0	Never treated
6	2011	13.5	0	1	0	Never treated
6	2012	13.9	0	1	0	Never treated
6	2013	14.2	0	1	0	Never treated
6	2014	15.1	0	1	0	Never treated
7	2009	13.7	0	0	0	Never treated
7	2010	14	0	0	0	Never treated
7	2011	14	0	1	0	Never treated
7	2012	14.9	0	1	0	Never treated
7	2013	15.1	0	1	0	Never treated
7	2014	14.9	0	1	0	Never treated
8	2009	12.7	0	0	0	Never treated
8	2010	13	0	0	0	Never treated
8	2011	15	0	1	0	Never treated
8	2012	15.5	0	1	0	Never treated
8	2013	16.1	0	1	0	Never treated
8	2014	17.2	0	1	0	Never treated
9	2009	14.9	0	0	0	Never treated
9	2010	16	0	0	0	Never treated
9	2011	15.5	0	1	0	Never treated
9	2012	16	0	1	0	Never treated
9	2013	16.7	0	1	0	Never treated
9	2014	17	0	1	0	Never treated
10	2009	14.7	0	0	0	Never treated
10	2010	15	0	0	0	Never treated
10	2011	14.4	0	1	0	Never treated
10	2012	15	0	1	0	Never treated
10	2013	15.7	0	1	0	Never treated
10	2014	16.1	0	1	0	Never treated

Table A5. *Cont.*

ID	Year	Consumption	*D*1	*D*2	TRET	First Year of Treatment
11	2009	13.1	1	0	0	2012
11	2010	14	1	0	0	2012
11	2011	14.8	1	0	0	2012
11	2012	16	1	1	1	2012
11	2013	16.2	1	1	1	2012
11	2014	15.5	1	1	1	2012
12	2009	12.9	1	0	0	2012
12	2010	13.3	1	0	0	2012
12	2011	14.7	1	0	0	2012
12	2012	16.1	1	1	1	2012
12	2013	16.7	1	1	1	2012
12	2014	18	1	1	1	2012
13	2009	12	1	0	0	2013
13	2010	12.8	1	0	0	2013
13	2011	13	1	0	0	2013
13	2012	13.9	1	0	0	2013
13	2013	15.4	1	1	1	2013
13	2014	16	1	1	1	2013

Additional information for using data of Table A5 to estimate alternative versions of the staggered DiD models are provided below.

Description of the variables of Table A5

- **DEPENDENT VARIABLE: CONSUMPTION**
- COACTOR: INCOME
- HETEROGENOUS TREATMENT: A Consumption Credit (for instance a policy measure that supports consumption (for instance a consumption local credit card with public warrant.

DiD DUMMIES

*D*1 = 0 if the consumer was never treated
*D*1 = 1 if the consumer was treated, sooner or later
*D*2 = 0 if the treatment did not exist in that year for that consumer
*D*2 = 1 if the treatment exists in that year for that consumer

ID: CONSUMERS

1 to 5 are Treated from 2011
6 to 10 are Never Treated
11 to 12 are Treated from 2012
13 is Treated from 2013 to 2014

TREATMENT TIMING

From 2009 to 2010, no treatment existed.
From 2011 to 2012, there was a treatment on individuals 1, 2, 3, 4, and 5.
In 2012, a treatment was extended to individuals 11 and 12.
In 2013, a treatment further extended to individuals of unit 13.

SUMMARY OF UNITS AND COHORTS			
Cohorts	Units and Observations		
	Never Treated Units	5 units	30 Observations
First Cohort	Units Treated from 2011	5 units	30 Observations
Second Cohort	Units Treated from 2012	2 units	12 Observations
Third Cohort	Units Treated from 2013	1 unit	6 Observations

The units treated since 2011 form the first cohort, and so on. Once the treatment is introduced, each unit in the treated cohort remains treated until the end of the sample period. TWFE, RA and IPW estimations of ATET can be obtained by applying the methods presented in Section 8 and following.

More information on the behaviour of the response variable is provided by the following plots of the mean values shown in Figure A1.

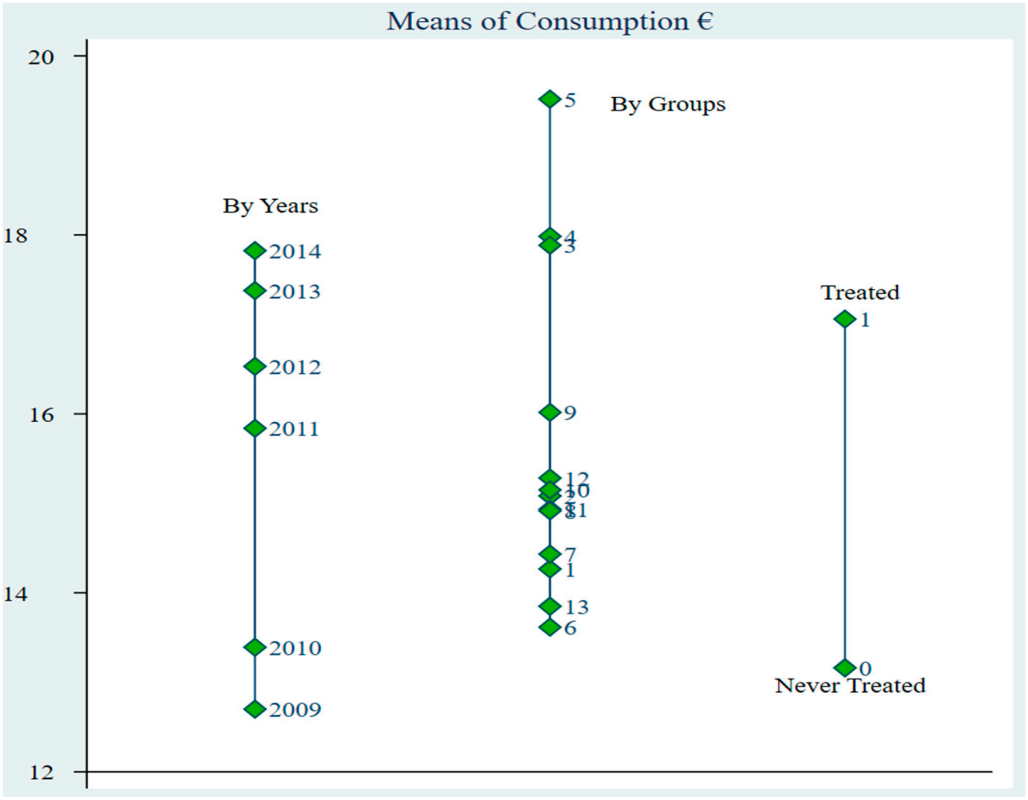


Figure A1. Consumption means clustered by year, group, and treatment assignment. *Notes*. Mean values are computed using the dataset provided in Table A5.

Table A6. DiD estimates (without cofactors) obtained using the data reported in Table A5.

		ATET (SE in Parenthesis)			
Cohorts	**YEARS**	**TWFE**	**RA**	**IPW**	**AIPW**
2011	2010	/ /	0.18 (0.20)	0.18 (0.20)	0.18 (0.2)
	2011	5.20 *** (0.99)	5.32 *** (0.93)	5.32 *** (0.93)	5.32 *** (0.93)
	2012	5.22 *** (1.1)	5.26 *** (1.01)	5.26 *** (1.01)	5.26 *** (1.01)
	2013	6.0 *** (1.06)	6 *** (0.97)	6 *** (0.97)	6 *** (0.97)
	2014	5.92 *** (1.03)	5.92 *** (0.96)	5.92 *** (0.96)	5.92 *** (0.96)

Table A6. *Cont.*

		ATET (SE in Parenthesis)			
Cohorts	**YEARS**	**TWFE**	**RA**	**IPW**	**AIPW**
2012	2010	//	0.05 (0.24)	0.05 (0.24)	0.05 (0.24)
	2011	//	0.82 (0.47)	0.82 (0.47)	0.82 (0.47)
	2012	1.23 ** (0.31)	0.72 *** (0.10)	0.72 *** (0.10)	0.72 *** (0.10)
	2013	1.17 ** (0.44)	0.62 * (0.23)	0.62 * (0.23)	0.62 * (0.23)
	2014	0.97 (1.27)	0.42 (0.94)	0.42 (0.94)	0.42 (0.94)
2013	2010	//	0.2 (0.16)	0.2 (0.16)	0.2 (0.16)
	2011	//	−0.08 (0.42)	−0.08 (0.42)	−0.08 (0.42)
	2012	//	0.32 *** (0.07)	0.32 ** (0.08)	0.32 ** (0.08)
	2013	1.5 *** (0.22)	1 *** (0.09)	1 *** (0.09)	1 *** (0.09)
	2014	1.35 ** (0.43)	1.1 ** (0.24)	1.1 (0.25)	1.1 (0.25)
Overall ATET		4.32 ** (1.11)	4.22 *** (1.003)	4.22 *** (1.00)	4.22 *** (1.00)
Average ATET by years					
2011		5.2 *** (0.99)	5.32 *** (0.93)	5.32 *** (0.93)	5.32 *** (0.93)
2012		4.08 ** (1.16)	3.96 ** (1.05)	3.96 ** (1.05)	3.96 ** (1.05)
2013		4.2 ** (1.21)	4.03 ** (1.08)	4.03 ** (1.08)	4.03 ** (1.08)
2014		4.11 ** (1.28)	3.94 ** (1.13)	3.94 ** (1.13)	3.94 ** (1.13)

Notes. Notice that the absence of covariates RA, IPW, and AIPW generate the same ATET estimations for each cohort as well as the same average ATET for 2011. The reader may replicate the exercise and include the cofactor income (included in the dataset) and obtain different ATET/CATET estimates. The reader may replicate the above exercise to evaluate whether the methods give the a common average ATET by cohorts. Asterisks correspond to usual significance. // means that no estimation is conducted by the model.

Notes

[1] We lastly recall that software packages useful to implement basic and more advanced DiD methods can be found in the following websites (listed in alphabetic order): R®: https://asjadnaqvi.github.io/DiD/docs/02_R/; Stata®: https://asjadnaqvi.github.io/DiD/docs/01_stata/. We did not access the websites during the editing of this paper.

[2] The present review does not address the domain of healthcare, as readers can access numerous DiD studies that have been utilised to evaluate novel policies and healthcare programs. For instance, in the United States, numerous studies have estimated the effects of expanded Medicaid eligibility through the Affordable Care Act (ACA). In the aftermath of the Supreme Court's decision regarding the ACA, each state within the United States is empowered to determine its own Medicaid eligibility criteria, with the option to expand its threshold. This methodological framework enabled the establishment of groups comprising treated states and comparison (untreated) states, thereby facilitating the implementation of DiD. These studies have contributed to ongoing policy debates in the US regarding the future of the ACA, and readers are encouraged to consult the relevant literature (see Zeldow & Hatfield, 2021 for an introduction).

[3] Consequences of non-random assignment are discussed, among others, by Cerulli (2015, p. 17).

[4] Researchers can then also find the breakdown point—how much of a deviation from the pre-existing difference in trends is needed before we can no longer reject the null of no parallel trend.

[5] The authors acknowledge that previous researchers have stressed that p-values are not meant to express the strength of evidence in favour of the null.

[6] According to the authors, this procedure does not simply control for this covariate but rather allows for its use in a 2SLS or GMM estimator.

[7] Callaway (2022) discusses an ampler set of estimation strategies. According to Callaway (2022, p. 4), all of them explicitly make, in a first step, the same good comparisons that show up in the TWFE regression (i.e., the comparisons that use units that become treated relative to units that are not-yet-treated) while explicitly avoiding the "bad comparisons" that show up in the TWFE regression (i.e., the comparisons that use already-treated units as the comparison group). Then, in a second step, they combine these underlying treatment effect parameters into target parameters of interest such as an overall average treatment effect on the treated. See Section Alternative Approaches in Callaway (2022, p. 20).

[8] Several different ICC statistics have been proposed, not all of which estimate the same population parameter. There has been considerable debate about which ICC statistics are appropriate for a given use, since they may produce markedly different results for the same data.

[9] A list of bias correction procedures is provided by Angrist and Pischke (2009, pp. 320–322).

[10] In a later paper (Feldstein, 1999) he also argues that traditional analyses of the income tax greatly underestimate deadweight losses by ignoring its effect on forms of compensation and patterns of consumption. He calculated the full deadweight loss using the compensated elasticity of taxable income to changes in tax rates because leisure, excludable income, and deductible consumption are assumed (by Feldstein) to be a Hicksian composite good. According to his estimations a deadweight loss of as much as 30% of revenue or more than ten times Harbergers classic 1964 estimate. The relative deadweight loss caused by increasing existing tax rates is substantially greater and, according to Feldstein's results, may exceed USD 2 per USD 1 of revenue. Some enormous measure, one should say!

[11] The treatment incorporated in the Feldstein's analysis was the 1986 US tax reform that lowered marginal tax rates, and simultaneously broadened tax bases. The two elements were designed to net out. Approximately no revenue and distributional effects absent behavioural responses means that approximately there are no income effects. Important as the aim is to estimate the compensated elasticity of taxable income.

[12] This *review* does not discuss synthetic controls. One should see Abadie et al. (2015). A synthetic control can be constructed as a weighted average of several units combined to recreate the trajectory that the response variable of a treated unit would have followed in the absence of the treatment. Recent advances in this field can be found in Y. Sun et al. (2025).

[13] This is in contrast with the view that the appropriate tax rate for decisions on the intensive margin is the marginal tax rate (MTR = tax rate on the last euro earned). In the paper ATR is not exact but approximated (for a subsample of football players). Since these taxpayers earn very high salaries, authors approximate the ATR by the top marginal tax rate (MTR). An alternative, and possible more reliable procedure is followed by Moretti and Wilson (2017). By focusing on the locational outcomes of star scientists, defined as scientists with patent counts in the top 5 percent of the distribution, their paper quantifies how sensitive is migration by these stars to changes in personal and business tax differentials across states in the USA. The study uncovers large, stable, and precisely estimated effects of personal and corporate taxes on star scientists, migration patterns. The long-run elasticity of mobility relative to taxes is 1.8 for personal income taxes, 1.9 for state corporate income tax, and -1.7 for the investment tax credit.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510. [CrossRef]

Ahrens, A., Aitken, C., & Schaffer, M. E. (2021). Using machine learning methods to support causal inference in econometrics. In S. Sriboonchitta, V. Kreinovich, & W. Yamaka (Eds.), *Behavioral predictive modeling in economics* (pp. 23–52). Springer International Publishing. [CrossRef]

Ahrens, A., Chernozhukov, V., Hansen, C., Kozbur, D., Schaffer, M., & Wiemann, T. (2025). An introduction to double/debiased machine learning. *arXiv*, arXiv:2504.08324. [CrossRef]

Alvarez, L., & Ferman, B. (2020). Inference in difference-in-differences with few treated units and spatial correlation. *arXiv*, arXiv:2006.16997.

Alvarez, L., Ferman, B., & Wüthrich, K. (2025). Inference with few treated units. *arXiv*, arXiv:2504.19841. [CrossRef]

Angrist, J. D., GImbens, W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. [CrossRef]

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

Angrist, J. D., & Pischke, J.-S. (2010). The Credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. [CrossRef]

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, *67*(4), 648–660. [CrossRef]

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., & Sant'Anna, P. (2025). Difference-in-differences designs: A practitioner's guide. *arXiv*, arXiv:2503.13323. [CrossRef]

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–181.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust difference-in-differences estimates? *Quarterly Journal of Economics*, *119*, 249–275. [CrossRef]

Bester, C. A., Conley, T. G., & Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, *165*, 137–151. [CrossRef]

Bilinski, A., & Hatfield, L. (2020, August 5). *Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions*. JSM 2020 Virtual Conference, in Virtual. Available online: https://ww2.amstat.org/meetings/jsm/2020/onlineprogram/AbstractDetails.cfm?abstractid=312323 (accessed on 10 August 2025).

Borusyak, K., & Jaravel, X. (2018). *Revisiting event study designs*. Social science research network. SSRN Scholarly Paper ID 2826228. Available online: https://scholar.harvard.edu/files/borusyak/files/borusyak_jaravel_event_studies.pdf (accessed on 10 August 2025).

Bosco, B. P., Bosco, C. F., & Maranzano, P. (2025). Labour responsiveness to income tax changes: Empirical evidence from a DID analysis of an income tax treatment in Italy. *Empirical Economics*, *69*(2), 787–828. [CrossRef]

Callaway, B. (2022). Difference-in-differences for policy evaluation. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population* (pp. 1–61). Springer. [CrossRef]

Callaway, B., & Sant'Anna, P. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*, 200–230. [CrossRef]

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414–427. [CrossRef]

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. [CrossRef]

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Card, D., & Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and Pennsylvania. *American Economic Review*, *84*(4), 772–793.

Cerqua, A., Letta, M., & Menchetti, F. (2022). *Losing control (group)? The machine learning control method for counterfactual forecasting*. SSRN.

Cerqua, A., Letta, M., & Menchetti, F. (2023). The machine learning control method for counterfactual forecasting. *arXiv*, arXiv:2312.05858. [CrossRef]

Cerqua, A., Letta, M., & Menchetti, F. (2024). Causal inference and policy evaluation without a control group. *arXiv*, arXiv:2312.05858.

Cerulli, G. (2015). Econometric evaluation of socio-economic programs. In *Theory and applications*. Springer-Verlag GmbH.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–265. [CrossRef]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2016). Double/debiased machine learning for treatment and causal parameters. *arXiv*, arXiv:1608.00060.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. [CrossRef]

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*, *25*(3), 576–601. [CrossRef]

Chesnaye, N., Stel, S., Tripepi, G., Dekker, F. W., Fu, E. L., Zoccali, G., & Jager, K. J. (2022). An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, *15*(1), 14–20. [CrossRef]

Cole, S. R., & Frangakis, C. E. (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, *20*(1), 3–5. [CrossRef]

Courtemanche, C. J., & Zapata, D. (2014). Does universal coverage improve health? The Massachusetts experience. *Journal of Policy Analysis and Management*, *33*, 36–69. [CrossRef]

Cox, D. R. (1958). *Planning of experiments, Wiley series in probability and statistics—Applied probability and statistics section*. John Wiley & Sons Inc.

de Chaisemartin, C., & D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996. [CrossRef]

de Chaisemartin, C., & D'Haultfoeuille, X. (2023). Two-way fixed effects estimators with heterogeneous treatment effects: A survey. *Econometrics Journal*, *26*, C1–C30. [CrossRef]

Delgado, M. S., & Florax, R. J. G. M. (2015). Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction. *Economics Letters*, *137*, 123–126. [CrossRef]

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, *89*(2), 221–233. [CrossRef]

Dong, F., Li, Y., Qin, C., Zhang, X., Chen, Y., Zhao, X., & Wang, C. (2022). Information infrastructure and greenhouse gas emission performance in urban China: A difference-in-differences analysis. *Journal of Environmental Management*, *316*, 115252. [CrossRef]

Elhorst, J. P. (2010). Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis*, *5*(1), 9–28. [CrossRef]

Farbmacher, H., Huber, M., Laffers, L., Langen, H., & Spindler, M. (2022). Causal mediation analysis with double machine learning. *The Econometrics Journal*, *25*(2), 277–300. [CrossRef]

Feldstein, M. (1995). The effect of marginal tax rates on taxable income: A panel study of the 1986 tax reform act. *Journal of Political Economy*, *103*, 551–572. [CrossRef]

Feldstein, M. (1999). Tax avoidance and the deadweight loss of the income tax. *The Review of Economics and Statistics*, *81*(4), 674–680. [CrossRef]

Fredriksson, A., & de Oliveira, G. M. (2019). Impact evaluation using difference-in-differences. *RAUSP Management Journal*, *54*(4), 519–532. [CrossRef]

Freyaldenhoven, S., Hansen, C., & Shapiro, J. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, *109*, 3307–3338. [CrossRef]

Galiani, S., Gertler, P., & Schargrodsky, E. (2005). Water for life: The impact of the privatization of water services on child mortality. *Journal of Political Economy*, *113*, 83–120. [CrossRef]

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. The World Bank.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. [CrossRef]

Huber, M., & Steinmayr, A. (2021). A framework for separating individual-level treatment effects from spillover effects. *Journal of Business and Economic Statistics*, *39*(2), 422–436. [CrossRef]

Imbens, G. W., & Kolesar, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, *98*(4), 701–712. [CrossRef]

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Kahn-Lang, A., & Lang, K. (2020). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business and Economic Statistics*, *38*, 613–620. [CrossRef]

Kleven, K. J., Landais, C., & Saez, E. (2013). Taxation and international migration of superstars: Evidence from the European football market. *American Economic Review*, *103*(5), 1892–1924. [CrossRef]

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627. [CrossRef]

Laffers, L., & Mellace, G. (2020). *Identification of the average treatment effect when SUTVA is violated* (Vol. 3). Discussion Papers on Business and Economics, University of Southern Denmark.

MacKinnon. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, *52*(3), 851–881. [CrossRef]

MacKinnon, J. G., Ørregaard Nielsen, M., & Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, *232*(2), 272–299. [CrossRef]

MacKinnon, J. G., & Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal*, *21*, 114–135. [CrossRef]

Manski, F. C., & Pepper, J. V. (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics*, *100*(2), 232–244. [CrossRef]

Moretti, E., & Wilson, D. J. (2017). The effect of state taxes on the geographical location of top earners: Evidence from star scientists. *American Economic Review*, *107*(7), 1858–1903. [CrossRef]

Myint, L. (2024). Controlling time-varying confounding in difference-in-differences studies using the time-varying treatments framework. *Health Services and Outcomes Research Methodology*, *24*, 95–111. [CrossRef]

Ogburn, E. L., Shpitser, I., & Lee, Y. (2020). Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(4), 1659–1676. [CrossRef] [PubMed]

Ogburn, E. L., Sofrygin, O., Díaz, I., & van der Laan, M. J. (2024). Causal inference for social network data. *Journal of the American Statistical Association*, *119*(545), 597–611. [CrossRef]

Pernagallo, G., & Vitali, G. (2025). Housing market investment in peripheral areas: Evidence from Italy. *Papers in Regional Science*, *104*(5), 100113. [CrossRef]

Qiu, F., & Tong, Q. (2021). A spatial difference-in-differences approach to evaluate the impact of light rail transit on property values. *Economic Modelling*, *99*, 105496. [CrossRef]

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, *90*, 2555–2591. [CrossRef]

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, *147*(5), 656–666. [CrossRef]

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, *4*(3), !305–322.

Roth, J., Sant'Anna, P., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*, 2218–2244. [CrossRef]

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58. [CrossRef]

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593. [CrossRef]

Rubin, D. B. (1990a). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*, 279–292. [CrossRef]

Rubin, D. B. (1990b). On the application of probability theory to agricultural experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*(4), 472–480. [CrossRef]

Schnabl, P. (2012). The international transmission of bank liquidity shocks: Evidence from an emerging market. *The Journal of Finance*, *67*, 897–932. [CrossRef]

Schwartz, S., Gatto, N. M., & Campbell, U. B. (2012). Extending the sufficient component cause model to describe the Stable Unit Treatment Value Assumption (SUTVA). *Epidemiologic Perspectives and Innovations*, *9*(1), 3. [CrossRef]

Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*(4), 1111–1122. [CrossRef]

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, *101*(476), 1398–1407. [CrossRef]

Souto, H. G., & Neto, F. L. (2025). Forests for differences: Robust causal inference beyond parametric DiD. *arXiv*, arXiv:2505.09706. [CrossRef]

Sun, S., & Delgado, M. S. (2024). Local spatial difference-in-differences models: Treatment correlations, response interactions, and expanded local models. *Empirical Economics*, *67*(5), 2077–2107. [CrossRef]

Sun, Y., Xie, H., & Zhang, Y. (2025). Difference-in-differences meets synthetic control: Doubly robust identification and estimation. *arXiv*, arXiv:2503.11375.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883. [CrossRef]

VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods and Research*, *38*(4), 515–544. [CrossRef] [PubMed]

VanderWeele, T. J., Tchetgen, E. J. T., & Halloran, M. E. (2015). Interference and sensitivity analysis. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *29*(4), 687. [CrossRef]

Wang, Y. (2021). Causal inference with panel data under temporal and spatial interference. *arXiv*, arXiv:2106.15074.

Wang, Y., Samii, C., Chang, H., & Aronow, P. (2024). Design-based inference for spatial experiments under unknown interference. *arXiv*, arXiv:2010.13599. [CrossRef]

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. *Annual Review of Public Health*, *39*, 453–469. [CrossRef]

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). The MIT Press.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *Econometrics Journal*, *26*, C31–C66. [CrossRef]

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Empir Econ*. [CrossRef]

Xu, Y. (2024). Causal inference with time-series cross-sectional data: A reflection. In J. M. Box-Steffensmeier, D. P. Christenson, & V. Sinclair-Chapman (Eds.), *Oxford handbook of engaged methodological pluralism in political science*. Oxford University Press. [CrossRef]

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, *15*(5), 74. [CrossRef]

Zeldow, B., & Hatfield, L. A. (2021). Confounding and regression adjustment in difference-in-differences studies. *Health Services Research*, *56*, 932–941. [CrossRef]

Zhou, B., Zhang, C., Song, H., & Wang, Q. (2019). How does emission trading reduce China's carbon intensity? An exploration using a decomposition and difference-indifferences approach. *Science of the Total Environment*, *676*, 514–523. [CrossRef]