IMPACT EVALUATION USING DIFFERENCE-IN-DIFFERENCES

July, 2019

ANDERS FREDRIKSSON

CORS - Center for Organization Studies, FEA-USP, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 908, São Paulo CEP 05508-900, SP, Brazil.
E-mail: anders.fredriksson@usp.br, Phone: +551130910616 (corresponding author)

GUSTAVO MAGALHÃES DE OLIVEIRA

CORS - Center for Organization Studies, FEA-USP, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 908, São Paulo CEP 05508-900, SP, Brazil.
E-mail: gustavomoliv@gmail.com, Phone: +551130910616

Purpose

This paper aims to present the Difference-in-Differences (DiD) method in an accessible language to a broad research audience from a variety of management-related fields.

Design/methodology/approach

The paper describes the DiD method, starting with an intuitive explanation, goes through the main assumptions and the regression specification, and covers the use of several robustness methods. Recurrent examples from the literature are used to illustrate the different concepts.

Findings

By providing an overview of the method, we cover the main issues involved when conducting DiD studies, including the fundamentals as well as some recent developments.

Originality/value

The article can hopefully be of value to a broad range of management scholars interested in applying impact evaluation methods.

1. INTRODUCTION

Difference-in-Differences (DiD) is one of the most frequently used methods in impact evaluation studies. Based on a combination of before-after and treatment-control group comparisons, the method has an intuitive appeal and has been widely used in economics, public policy, health research, management and other fields. After the introductory section, this paper outlines the method, discusses its main assumptions, then provides further details and discusses potential pitfalls. Examples of typical DiD evaluations are referred to throughout the text, and a separate section discusses a few papers from the broader management literature. Conclusions are also presented.

Differently from the case of randomized experiments that allow for a simple comparison of treatment and control groups, DiD is an evaluation method used in non-experimental settings. Other members of this "family" are matching, synthetic control and regression discontinuity. The goal of these methods is to estimate causal effects of a program when treatment assignment is non-random; hence there is no obvious control group.[1] Although random assignment of treatment is prevalent in medical studies and has become more common also in the social sciences, through e.g. pilot studies of policy interventions, most real-life situations involve non-random assignment. Examples include the introduction of new laws, government policies and regulation.[2] When discussing different aspects of the DiD method, a much researched 2006 healthcare reform in Massachusetts, that aimed to give nearly all residents healthcare coverage, will be used as an example of a typical DiD study object. In order to estimate the causal impact of this and other policies, a key challenge is to find a proper control group.

In the Massachusetts example, one could use as control a state that did not implement the reform. A DiD estimate of reform impact can then be constructed, which in its simplest form is equivalent to calculating the after-before difference in outcomes in the treatment group, and subtracting from this difference the after-before difference in the control group. This double difference can be calculated whenever treatment and control group data on the outcomes of interest exist before and after the policy intervention. Having such data is thus a prerequisite to apply Difference-in-Differences. As will be detailed below, however, fulfilling

[1] The reader is assumed to have basic knowledge about regression analysis (e.g. Wooldridge, 2012) and also about the core concepts in impact evaluation, e.g. identification strategy, causal inference, counterfactuals, randomization and treatment effects (e.g. Gertler, Martinez, Premand, Rawlings, and Vermeersch, 2016, chapters 3-4; White and Raitzer, 2017, chapters 3-4).

[2] In this text, the terms policy, program, reform, law, regulation, intervention, shock or treatment are used interchangeably, when referring to the object being evaluated, i.e. the treatment.

this criterion does not imply that the method is always appropriate or that it will give an unbiased estimate of the causal effect.

Labor economists were among the first to apply DiD methods.[3] Ashenfelter (1978) studied the effect of training programs on earnings and Card (1990) studied labor market effects in Miami after a (non-anticipated) influx of Cuban migrants. As a control group, Card used other US cities, similar to Miami along some characteristics, but without the migration influx. Card and Krueger (1994) studied the impact of a New Jersey rise in the minimum wage on employment in fast-food restaurants. Neighboring Pennsylvania maintained its minimum wage and was used as control. Many other studies followed.

Although the basic method has not changed, a number of issues have been brought forward in the literature, and academic studies have evolved along with these developments. Two non-technical references covering DiD are Gertler et al. (2016) and White and Raitzer (2017), whereas Angrist and Pischke (2009, chapter 5) and Wooldridge (2012, chapter 13) are textbook references. In chronological order, Angrist and Krueger (1999), Bertrand, Duflo, and Mullainathan (2004), Blundell and Costa Dias (2000, 2009), Imbens and Wooldridge (2009), Lechner (2011), Athey and Imbens (2017), Abadie and Cattaneo (2018) and Wing, Simon, and Bello-Gomez (2018) also review the method, including more technical content. The main issues brought forward in these and other references are discussed below.

2. THE DIFFERENCE-IN-DIFFERENCES METHOD

The Difference-in-Differences method combines insights from cross-sectional treatment-control comparisons and before-after studies, for a more robust identification. First consider an evaluation that seeks to estimate the effect of a (non-randomly implemented) policy ("treatment") by comparing outcomes in the treatment group to a control group, with data from after the policy implementation. Assume there is a difference in outcomes. In the Massachusetts health reform example, perhaps health is better in the treatment group. This difference may be due to the policy, but also because there are key characteristics that differ between the groups and that are determinants of the outcomes studied, e.g. income in the health reform example: Massachusetts is relatively rich, and wealthier people on average have better health. A remedy for this situation is to evaluate the impact of the policy after controlling for the factors that differ between the two groups. This is only possible for

[3] Lechner (2011) provides a historical account, including Snow's study of cholera in London in the 1850s.

observable characteristics, however. Perhaps important socioeconomic and other characteristics that determine outcomes are not in the dataset, or even fundamentally unobservable. And even if it would be possible to collect additional data for certain important characteristics, the knowledge about which are all the relevant variables is imperfect. Controlling for all treatment-control group differences is thus difficult.

Consider instead a before-after study, with data from the treatment group. The policy under study is implemented between the before and after periods. Assume a change over time is observed in the outcome variables of interest, such as better health. In this case the change may have been caused by the policy, but may also be due to other changes that occurred at the same time as the policy was implemented. Perhaps there were other relevant government programs during the time of the study, or the general health status is changing over time. With treatment group data only, the change in the outcome variables may be incorrectly attributed to the intervention under study.

Now consider combining the after-before approach and the treatment-control group comparison. If the after-before difference in the control group is deducted from the same difference in the treatment group, two things are achieved. First, if other changes that occur over time are also present in the control group, then these factors are controlled for when the control group after-before difference is netted out from the impact estimate. Second, if there are important characteristics that are determinants of outcomes and that differ between the treatment and control groups, then, as long as these treatment-control group differences are constant over time, their influence is eliminated by studying changes over time. Importantly, this latter point applies also to treatment-control group differences in time-invariant unobservable characteristics (as they are netted out). One can thus get around the problem, present in cross-sectional studies, that we cannot control for unobservable factors (further discussed below).

To formalize some of what has been said above, the basic Difference-in-Differences study has data from two groups and two time periods, and the data is typically at the individual level, that is, at a lower level than the treatment intervention itself. The data can be repeated cross-sectional samples of the population concerned (ideally random draws) or a panel. Wooldridge (2012, chapter 13) gives examples of DiD studies using the two types of data structures and discusses the potential advantages of having a panel rather than repeated cross sections (also refer to Angrist and Pischke, 2009, chapter 5; and Lechner, 2011).

With two groups and two periods, and with a sample of data from the population of interest, the DiD estimate of policy impact can be written as follows:

$$DiD =$$

$$\left(\bar{y}_{s=Treatment,t=After} - \bar{y}_{s=Treatment,t=Before}\right) - \left(\bar{y}_{s=Control,t=After} - \bar{y}_{s=Control,t=Before}\right) \quad (1)$$

where y is the outcome variable, the bar represents the average value (averaged over individuals, typically indexed by i), the group is indexed by s (because in many studies, policies are implemented at the state level) and t is time. With before and after data for treatment and control, the data is thus divided into the four groups and the above double difference is calculated. The information is typically presented in a 2×2 table, then a third row and a third column are added in order to calculate the after-before and treatment-control differences and the DiD impact measure. Figure 1 illustrates how the DiD estimate is constructed.

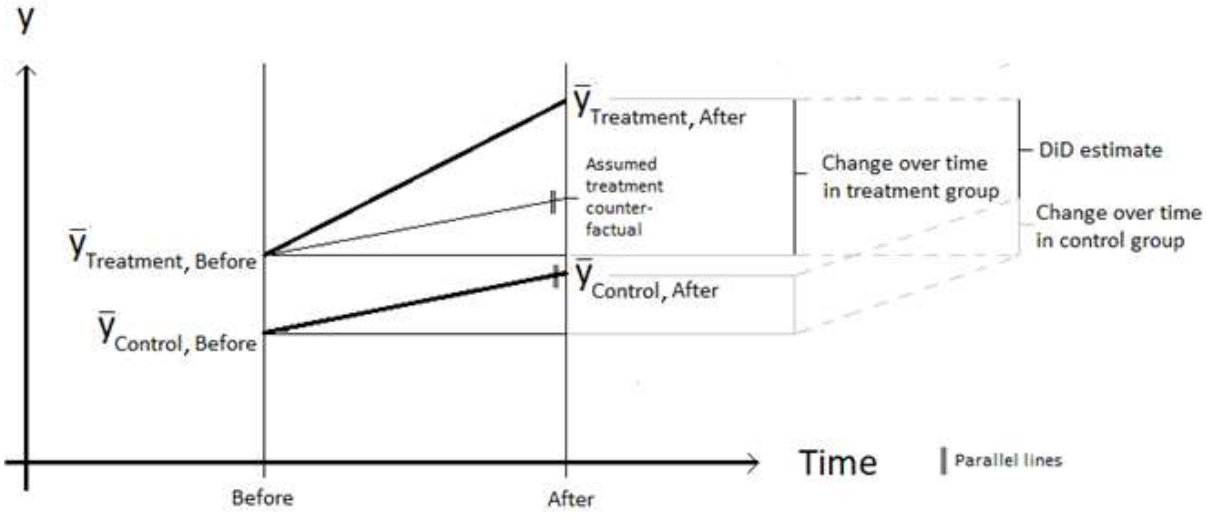

Figure 1. Illustration of the two-group two-period Difference-in-Differences estimate. The assumed treatment group counterfactual equals the treatment group pre-reform value plus the after-before difference from the control group.

The above calculation and illustration say nothing about the significance level of the DiD estimate, hence regression analysis is used. In an OLS framework, the DiD estimate is obtained as the β-coefficient in the following regression, in which A_s are treatment/control

group fixed effects, B_t before/after fixed effects, I_{st} is a dummy equaling 1 for treatment observations in the after period (otherwise it is zero) and ε_{ist} the error term[4]:

$$y_{ist} = A_s + B_t + \beta I_{st} + \varepsilon_{ist} \quad (2)$$

In order to verify that the estimate of β will recover the DiD estimate in (1), use (2) to get

$$E(y_{ist}|s = Control, t = Before) = A_{Control} + B_{Before}$$

$$E(y_{ist}|s = Control, t = After) = A_{Control} + B_{After}$$

$$E(y_{ist}|s = Treatment, t = Before) = A_{Treatment} + B_{Before}$$

$$E(y_{ist}|s = Treatment, t = After) = A_{Treatment} + B_{After} + \beta$$

In these expressions, $E(y_{ist}|s,t)$ is the expected value of y_{ist} in population subgroup (s,t), which is estimated by the sample average $\bar{y}_{s,t}$. Estimating (2) and plugging in the sample counterpart of the above expressions into (1), with the hat notation representing coefficient estimates, gives $DiD = \hat{\beta}$.[5]

The DiD model is not limited to the 2×2 case, and expression 2 is written in a more general form than what was needed so far. For models with several treatment- and/or control groups, A_s stands for fixed effects for each of the different groups. Similarly, with several before- and/or after periods, each period has its own fixed effect, represented by B_t. If the reform is implemented in all treatment groups/states at the same time, I_{st} switches from zero to one in all such locations at the same time. In the general case, however, the reform is staggered and hence implemented in different treatment groups/states s at different times t. I_{st} then switches from 0 to 1 accordingly. All these cases are covered by expression 2.[6]

Individual-level control variables X_{ist} can also be added to the regression, which becomes

$$y_{ist} = A_s + B_t + cX_{ist} + \beta I_{st} + \varepsilon_{ist}. \quad (3A)$$

[4] The variable denominations are similar to those in Bertrand et al. (2004). An alternative way to specify regression 2, in the 2×2 case, is to use an intercept, treatment- and after dummies and a dummy equaling the interaction between the treatment and after dummies (e.g. Wooldridge, 2012, chapter 13). The regression results are identical.

[5] Angrist and Pischke (2009), Blundell and Costa Dias (2009), Lechner (2011) and Wing et al. (2018) are examples of references that provide additional details on the correspondence between the "potential outcomes framework", the informal/intuitive/graphical derivation of the DiD measure and the regression specification, as well as a discussion of population vs. sample properties.

[6] Note that the interpretation of β changes somewhat if the reform is staggered (Goodman-Bacon, 2018). An even more general case, not covered in this text, is when I_{st} switches on and off. A particular group/state can then go back and forth between being treated and untreated (e.g. Bertrand et al., 2004). Again different is the case where I_{st} is continuous (e.g. Aragón and Rud, 2013).

An important aspect of DiD estimation concerns the data used. Although it cannot be done with a 2×2 specification (as there are four observations only), models with many time periods and treatment/control groups can also be analyzed with state-level (rather than individual-level) data (e.g. US or Brazilian data, with 50 and 27 states, respectively). There would then be no i-index in regression 3A. Perhaps the relevant data is at the state level (e.g. unemployment rates from statistical institutes). Individual-level observations can also be aggregated. An advantage of the latter approach is that one avoids the problem (discussed in section 4) that the within group-period (e.g. state-year) error terms tend to be correlated across individuals, hence standard errors should be corrected. With either type of data, also state-level control variables, Z_{st}, may be included in expression 3A.[7] A more general form of the regression specification, with individual-level data, therefore becomes:

$$y_{ist} = A_s + B_t + cX_{ist} + dZ_{st} + \beta I_{st} + \varepsilon_{ist}. \quad (3B)$$

3. PARALLEL TRENDS AND OTHER ASSUMPTIONS

Estimation of DiD models hinge upon several assumptions, which are discussed in detail by Lechner (2011). The following paragraphs are mainly dedicated to the "parallel trends" assumption, the discussion of which is a requirement for any DiD paper ("no pre-treatment effects" and "common support" are also discussed below). Another important assumption is the Stable Unit Treatment Value Assumption, which implies that there should be no spillover effects between the treatment and control groups, as the treatment effect would then not be identified (e.g. Duflo, Glennerster, and Kremer, 2008). Furthermore, the control variables X_{ist} and Z_{st} should be exogenous, unaffected by the treatment. Otherwise, $\hat{\beta}$ will be biased. A typical approach is to use covariates that predate the intervention itself, although this does not fully rule out endogeneity concerns, as there may be anticipation effects. In some DiD studies and data sets, the controls may be available for each time period (as suggested by the t-index on X_{ist} and Z_{st}), which is fine as long as they are not affected by the treatment. Implied by the assumptions is that there should be no compositional changes over time. An example would be if individuals with poor health move to Massachusetts (from a control state to the treatment state). The health reform impact would then likely be underestimated.

[7] Note that X_{ist} and Z_{st} are both vectors of variables. The X-variables could be e.g. gender, age and income, i.e. three variables, each with individual level observations. Z_{st} can be e.g. state unemployment, variables representing racial composition, number of hospital beds, etc., depending on the study. The regression coefficients c and d are (row) vectors.

Identification based on Difference-in-Differences relies on the parallel trends assumption, which states that the treatment group, absent the reform, would have followed the same time trend as the control group (in the outcome variable of interest). Observable and unobservable factors may cause the level of the outcome variable to differ between treatment and control, but this difference (absent the reform in the treatment group) must be constant over time. Because the treatment group is only observed as treated, the assumption is fundamentally untestable. One can lend support to the assumption, however, through the use of several periods of pre-reform data, showing that the treatment and control groups exhibit a similar pattern in pre-reform periods. If such is the case, the conclusion that the impact estimated comes from the treatment itself, and not from a combination of other sources (including those causing the different pre-trends), becomes more credible. Pre-trends cannot be checked in a dataset with one before-period only, however (Figure 1). In general, such studies are therefore less robust. A certain number of pre-reform periods is highly desirable and certainly a recommended "best practice" in DiD studies.

The papers on the New Jersey minimum wage increase by Card and Krueger (1994, 2000) (the first referred to in Section 1) illustrate this contention and its relevance. The 1994 paper uses a two-period dataset, February 1992 (before) and November 1992 (after). By using DiD, the paper implicitly assumes parallel trends. The authors conclude that the minimum wage increase had no negative effect on fast-food restaurant employment. In the 2000 paper, the authors have access to additional data, from 1991 to 1997. In a graph of employment over time, there is little visual support for the parallel trends assumption. The extended dataset suggests that employment variation may be due to other time-varying factors than the minimum wage policy itself (for further discussion, refer to Angrist and Pischke, 2009, chapter 5).

Figure 2A exemplifies, from Galiani, Gertler, and Schargrodsky (2005) and Gertler et al. (2016), how visual support for the parallel trends assumption is typically verified in empirical work. The authors study the impact of privatizing water services on child mortality in Argentina. Using a decade of mortality data and comparing areas with privatized- (treatment) and non-privatized water companies (control), similar pre-reform (pre-1995) trends are observed. In this case also the levels are almost identical, but this is not a requirement. The authors go on to find a statistically significant reduction in child mortality in areas with privatized water services. Figure 2B provides another example, with data on a

health variable before (and after) the 2006 Massachusetts reform, as illustrated by Courtemanche and Zapata, 2014.

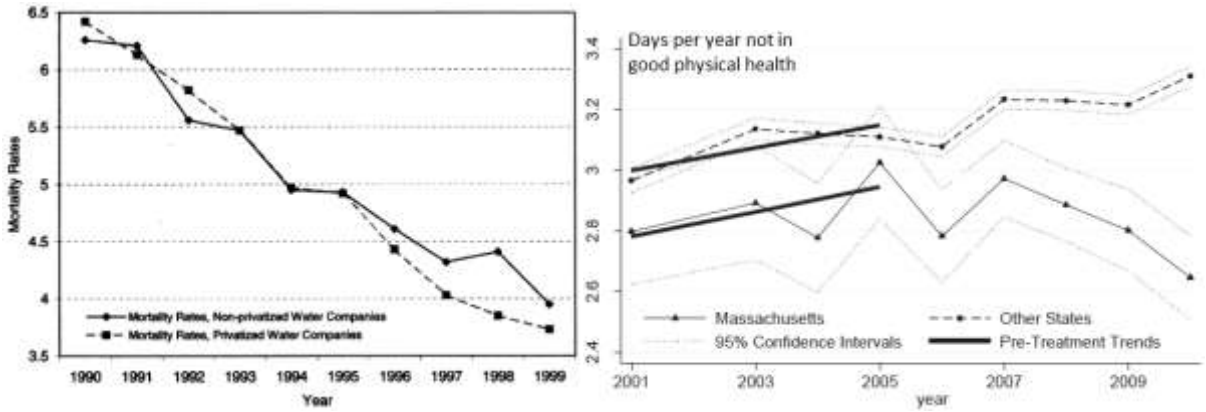

Figure 2. Graphs used to visually check the parallel trends assumption. **2A (left).** Child mortality rates, different areas of Buenos Aires, Argentina, 1990-1999 (reproduced from Galiani et al., 2005). **2B (right).** Days per year not in good physical health, 2001-2009, Massachusetts and control states (from Courtemanche and Zapata, 2014).

A more formal approach to provide support for the parallel trends assumption is to conduct placebo regressions, which apply 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 analyze the pre-reform periods only (if there is enough data available). In line with this approach, Schnabl (2012), who studies the effects of the 1998 Russian financial crisis on bank lending, uses two years of pre-crisis data for a placebo test. 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 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 expression 3B, and then check that the difference in trends is not significant (Wing et al, 2018, p. 459). [8]

The above discussion concerns the "raw" outcome variable itself. Lechner (2011) formulates the parallel trends assumption conditional on control variables (which should be exogenous). One study using a conditional parallel trends assumption is the paper on mining and local economic activity in Peru by Aragón and Rud (2013), especially their Figure 3. Another issue, which can be inspected in graphs such as Figure 2, is that there should be no effect

[8] See also Wing et al. (2018, pp. 460-461) for a discussion of the related concept of event studies. Their set-up can also be used to study short- and long term reform effects. A slightly different type of placebo test is to use control states only, to study if there is an effect where there should be none (Bertrand et al., 2004).

from the reform prior to its implementation. Finally, "common support" is needed. If the treatment group includes only high values of a control variable and the control group only low values, one is in fact comparing incomparable entities. There must instead be overlap in the distribution of the control variables between the different groups and time periods.

It should be noted that the parallel trends assumption is scale dependent, which is an undesirable feature of the DiD method. Unless the outcome variable is constant during the pre-reform periods, in both treatment and control, it matters if the variable is used "as is" or if it is transformed (e.g. wages vs. log wages). One approach to this issue is to use the data in the form corresponding to the parameter one wants to estimate (Lechner, 2011), rather than adapting the data to a format that happens to fit the parallel trends assumption.

A closing remark in this section is that it is worth spending time when planning the empirical project, before the actual analysis, carefully considering all possible data sources, if first hand data needs to be collected, etc. Perhaps data limitations are such that a robust DiD study – including a parallel trend check – is not feasible. On the other hand, in the process of learning about the institutional details of the intervention studied, new data sources may appear.

4. FURTHER DETAILS AND CONSIDERATIONS FOR THE USE OF DIFFERENCE-IN-DIFFERENCES

4.1 Using control variables for a more robust identification

With a non-random assignment to treatment, there is always the concern that the treatment states would have followed a different trend than the control states, even absent the reform. If, however, one can control for the factors that differ between the groups and that would lead to differences in time trends (and if these factors are exogenous), then the true effect from the treatment can be estimated.[9] In the above regression framework (expression 3B) one should thus control for the variables that differ between treatment and control and that would cause time trends in outcomes to differ. With treatment assignment at the state level, this is primarily a concern for state-level control variables (Z_{st}). The main reason for including also individual-level controls (X_{ist}) is instead to decrease the variance of the regression coefficient

[9] In relation to this discussion, note that the Difference-in-Differences method estimates the Average Treatment Effect *on the Treated*, not on the population (e.g. Blundell and Costa Dias, 2009; Lechner, 2011; White and Raitzer, 2017, chapter 5).

estimates (e.g. Angrist and Pischke, 2009, chapters 2 and 5; Wooldridge, 2012, chapters 6 and 13).

Matching is another way to use control variables to make DiD more robust. As suggested by the name, treatment and control group observations are matched, which should reduce bias. First think of a cross-sectional study with one dichotomous state level variable that is relevant for treatment assignment and for outcomes (e.g. Democrat/Republican state). Also assume that, even if states of one category/type are more likely to be treated, there are still treatment and control states of both types ("common support"). In this case, separate treatment effects would first be estimated for each category. The average treatment effect is then obtained by weighting with the number of treated states in each category. When the number of control variables grows and/or take on many different values (or are continuous), such exact matching is typically not possible. One alternative is to instead use the multidimensional space of covariates Z_s and calculate the distance between observations in this space. Each treatment observation is matched to one or several control observations (through e.g. "Mahalanobis matching", "n-nearest neighbor matching"), then an averaging is done over the treatment observations. Coarsening is another option. The multidimensional Z_s-space is divided into different bins, observations are matched within bins and the average treatment effect is obtained by weighting over bins. Yet an option is the propensity score. This one-dimensional measure represents the probability, given Z_s, that a state belongs to the treatment group. In practice, $P(Z_s)$ is the predicted probability from a logit/probit model of the treatment indicator regressed on Z_s. The method thus matches observations based on the propensity score, again using n-nearest neighbor matching, etc.[10]

When implementing matching in DiD studies, treatment and control observations are matched with methods similar to the above, e.g. coarsening or propensity score. In the case of a 2×2 study, a double difference similar to (1) is calculated, but the control group observations are weighted according to the results of the matching procedure.[11] An example of a DiD+matching study of the Massachusetts reform is Sommers, Long, and Baicker

[10] Matching (also referred to as "selection on observables") hinges upon the Conditional Independence Assumption (CIA) (or "unconfoundedness"), which says that, conditional on the control variables, treatment and control would have the same expected outcome, in either treatment state (treated/untreated). Hence the treatment group, if untreated, would have the same expected outcome as the control group, and the selection bias disappears (e.g. Angrist and Pischke, 2009, chapter 3). Rosenbaum and Rubin (1983) showed that if the CIA holds for a set of variables Z_s, then it also holds for the propensity score $P(Z_s)$.

[11] Such a method is used for panel data. When the data is repeated cross sections, each of the three groups treatment-before, control-before and control-after needs to be matched to the treatment-after observations (e.g. Blundell and Costa Dias, 2000; Smith and Todd, 2005).

(2014). With county level data, the authors use the propensity score to find a comparison group to Massachusetts counties.

A third approach using control variables is the synthetic control method. Similar to DiD, it aims at balancing pre-intervention trends in the outcome variables. In the original reference, Abadie and Gardeazabal (2003) construct a counterfactual Basque Country by using data from other Spanish regions. Inspired by matching, the method minimizes the (multidimensional) distance between the values of the covariates in the treatment and control groups, by choosing different weights for the different control regions. The distance measure also depends, however, on a weight factor for each individual covariate. This second set of weights is chosen such that the pre-intervention trend in the control group, for the outcome of interest, is as close as possible to the pre-intervention trend for the treatment group. As described by Abadie and Cattaneo (2018), the synthetic control method aims at providing a "data-driven" control group selection (and is typically implemented in econometrics software packages).

The Massachusetts health study of Courtemanche and Zapata (2014) illustrates a practice for how a DiD study may go about in selecting a control group. In the main specification, the authors use the rest of the United States as control (except a few states), and pre-reform trends are checked (including placebo tests). The control group is thereafter restricted, respectively, to the ten states with the most similar pre-reform health outcomes, to the ten states with the most similar pre-reform health trends and to other New England states only. Synthetic controls are also used. The DiD estimate is similar across specifications.

Related to the discussion of control variables is the threat to identification from compositional changes, briefly mentioned in section 3. Assume a certain state implements a health reform. Compare with a neighboring state. If the policy induces control group individuals with poor health to move to the treatment state, the treatment outcome will then be composed also of these movers. In this case, the ideal is to have data on (and control for) individuals' "migration status". In practice such data may not be available and controls X_{ist} and Z_{st} are instead used. This is potentially not enough, however, as there may be changes also in unobserved factors and/or spillovers and complementarities related to the changes in e.g. socioeconomic variables. One practice used to lend credibility to a DiD analysis is to search for treatment-induced compositional changes by using each covariate as a dependent

variable in an expression 2-style regression. Any significant effect (the β-coefficient) would indicate a potentially troublesome compositional change (e.g. Aragón and Rud, 2013).

4.2 Difference-in-Difference-in-Differences

Difference-in-Difference-in-Differences (DiDiD) is an extension of the DiD concept (e.g. Angrist and Pischke, 2009), briefly mentioned through an example. Long, Yemane, and Stockley (2010) study the effects of the special provisions for young people in the Massachusetts health reform. The authors use data on both young adults and on slightly older adults. Through the DiDiD method, they compare the change over time in health outcomes for young adults in Massachusetts to young adults in a comparison state *and* to slightly older adults in Massachusetts and construct *a triple* difference, with the aim of also controlling for other changes that occur in the treatment state.

4.3 Standard errors[12]

In the basic OLS framework, observations are assumed to be independent and standard errors homoscedastic. The standard errors of the regression coefficients then take a particularly simple form. Such errors are typically "corrected", however, to allow for heteroscedasticity (Ecker-Huber-White heteroscedasticity-robust standard errors). The second "standard" correction is to allow for clustering. Think of individual-level data from different regions, where some regions are treated, others are not. Within a region ("cluster"), the individuals are likely to share many characteristics: perhaps they go to the same schools, work at the same firms, have access to the same media outlets, are exposed to similar weather, etc. Factors such as these make observations within clusters correlated. In effect, there is less variation than if the data had been independent random draws from the population at large. Standard errors need to be corrected accordingly, typically implying that the significance levels of the regression coefficients are reduced.[13]

For correct inference with DiD, a third adjustment needs to be done. With many time periods, the data can exhibit serial correlation. This holds for many typical dependent variables in DiD studies, such as health outcomes, and, in particular, the treatment variable itself. The observations within each of the treatment and control groups can thus be correlated

[12] For a general discussion, refer to Angrist and Pischke (2009) and Wooldridge (2012). Abadie, Athey, Imbens, and Wooldridge (2017), Bertrand et al. (2004) and Cameron and Miller (2015) provide more details.
[13] When there are group effects, it is important to have a large enough number of group-period cells, in order to apply DiD, an issue further discussed in Bertrand et al. (2004).

over time. Failing to correct for this fact can largely overstate significance levels, which was the topic of the much influential paper by Bertrand et al. (2004).

One way of handling the within-group clustering issue is to collapse the individual data to state level averages. Similarly, the serial correlation problem can be handled by collapsing all pre-treatment periods to one before-period, and all post-treatment periods to one after-period. Having checked the parallel trends assumption, one thus works with two periods of data, at the state level (which requires many treatment and control states). A drawback, however, is that the sample size is greatly reduced. The option to instead continue with the individual-level data and calculate standard errors that are robust to heteroscedasticity, within-group effects and serial correlation, are provided by many econometric software packages.

5. EXAMPLES OF DIFFERENCE-IN-DIFFERENCES STUDIES IN THE BROADER MANAGEMENT LITERATURE

The DiD method is increasingly applied in management studies. A growing number of scholars use the method in areas such as innovation (Aggarwal and Hsu, 2014; Flammer and Kacperczyk, 2016; Sing and Agrawal, 2011), board of directors composition (Berger, Kick and Schaeck, 2014), lean production (Distelhorst, Hainmueller, and Locke, 2016), organizational goals management (Holm, 2018), CEO remuneration (Conyon, Hass, Peck, Sadler, and Zhang, 2019), regulatory certification (Bruno, Cornaggia and Cornaggia, 2016), social media (Kumar, Bezawada, Rishika, Janakiraman, and Kannan (2016), employee monitoring (Pierce, Snow, and McAfee, 2015) and environmental policy (He and Zhang, 2018).

Different sources of exogenous variation have been used for econometric identification in DiD papers in the management literature. A few examples are given here. Chen, Grossland and Huang (2014) study the effects of female board representation on mergers and acquisitions. In a robustness test to their main analysis, further addressing the issue that board composition may be endogenous, the authors exploit the fact that female board representation increases exogenously if a male board director dies. A small sample of 24 such firms are identified and matched to 24 control firms, and a basic two-group two-period DiD regression is run on this sample.

Younge, Tong and Fleming (2014) instead use DiD as the main method, and study how constraints on employee mobility affect the acquisition likelihood. The authors use as a source of identification a 1985 change in the Michigan antitrust law that had as an effect that employers could prohibit workers to leave for a competitor. Ten US states, where no changes allegedly occurred around 1985, are used as the control group. The authors also use ("coarsened exact") matching on firm characteristics to select the control group firms most similar to the Michigan firms. In addition, graphs of pre-treatment trends are presented.

Hosken, Olsen and Smith (2018) study the effect of mergers on competition. The authors do not have an exogenous source of variation, which is discussed at length. They compare grocery retail prices in geographical areas where horizontal mergers have taken place (treatment), to areas without such mergers. Several different control groups are constructed and a test with pre-treatment price data only is conducted, to assure there is no difference in price trends. Synthetic controls are also used.

Another study is Flammer (2014), who investigates whether product market competition affects investments in Corporate Social Responsibility (CSR). Flammer (2014) uses import tariff reductions as the source of variation in the competitive environment and compares affected sectors (treatment) to non-affected sectors (control) over time. A matching procedure is used to increase comparability between the groups, and a robustness check restricts the sample to treatment sectors where the tariff reductions are likely to be de facto exogenous. The author also uses control variables in the DiD regression, but as pointed out in the paper, these variables have already been used in the matching procedure and their inclusion does not alter the results.

Lemmon and Roberts (2010) study regulatory changes in the insurance industry as an exogenous contraction in the supply of below-investment-grade credit. Using Compustat data, they undertake a DiD analysis complemented by propensity score matching and explicitly analyze the parallel trends assumption. Iyer, Peydró, da-Rocha-Lopes, and Schoar (2013) examine how banks react in terms of lending when facing a negative liquidity shock. Based on Portuguese corporate loan-level data, they undertake a DiD analysis, with an identification strategy that exploits the unexpected shock to the interbank markets in August 2007. Other papers that have used DiD to study the effect of shocks to credit supply are Schnabl (2012), referenced above, and Khwaja and Mian (2008).

In addition to these topics, several DiD papers published in management journals relate to public policy and health, an area reviewed by Wing et al. (2018). The above referenced Aragón and Rud (2013) and Courtemanche and Zapata (2014) are two of many papers that apply several parts of the DiD toolbox.

6. DISCUSSION AND CONCLUSION

The paper presents an overview of the Difference-in-Differences method, summarized here in terms of some practical recommendations. Researchers wishing to apply the method should carefully plan their research design and think about what the source of (preferably exogenous) variation is, and how it can identify causal effects. The control group should be comparable to the treatment group and have the same data availability. Matching and other methods can refine the control group selection. Enough time periods should be available to credibly motivate the parallel trends assumption and, in case not fulfilled, it is likely that DiD is not an appropriate method. The robustness of the analysis can be enhanced by using exogenous control variables, either directly in the regression and/or through a matching procedure. Standard errors should be robust and clustered in order to account for heteroscedasticity, within-group correlation and serial correlation. Details may differ, however, including what the relevant cluster is, which depends on the study at hand, and researchers are encouraged to delve further into this topic (e.g. Bertrand et al., 2004; Cameron and Miller, 2015). Yet other methods, such as Difference-in-Difference-in-Differences and synthetic controls were discussed, while a discussion of e.g. time-varying treatment effects and another quasi-experimental technique, regression discontinuity, were left out. Several methodological DiD papers were cited above, the reading of which is encouraged, perhaps together with texts covering also other non-experimental methods.

The choice of research method will vary according to many circumstances. Difference-in-Differences has the potential to be a feasible design in many areas of management studies and scholars interested in the topic hopefully found this text of interest. The wide range of surveys and databases - Economatica, Capital IQ and Compustat are a few examples - enables the application of DiD in distinct contexts and to different research questions. Beyond data, the above cited studies also demonstrate innovative ways of getting an exogenous source of variation for a credible identification strategy.

Reference

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*, 465-503.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, *93*(1), 113-132.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (No. Working Paper 24003). National Bureau of Economic Research (NBER).

Aggarwal, V.A., & Hsu, D. H. (2014). Entrepreneurial exits and innovation. *Management Science*, 60(4), 867-887.

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In *Handbook of labor economics* (Vol. 3, pp. 1277-1366). Elsevier.

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

Aragón, F. M., & Rud, J. P. (2013). Natural resources and local communities: evidence from a Peruvian gold mine. *American Economic Journal: Economic Policy*, *5*(2), 1-25.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 47-57.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3-32.

Berger, A. N., Kick, T., & Schaeck, K. (2014). Executive board composition and bank risk taking. *Journal of Corporate Finance*, *28*, 48-65.

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

Blundell, R., & Costa Dias, M. (2000). Evaluation methods for non-experimental data. *Fiscal Studies*, *21*(4), 427-468.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, *44*(3), 565-640.

Bruno, V., Cornaggia, J., & Cornaggia, J. K. (2016). Does regulatory certification affect the information content of credit ratings? *Management Science*, 62(6), 1578-1597.

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

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *ILR Review*, *43*(2), 245-257.

Card, D., & Krueger, A.B. (1994). Wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84,772-793.

Card, D., & Krueger, A. B. (2000). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply. *American Economic Review*, 90(5), 1397-1420.

Chen, G., Crossland, C., & Huang, S. (2014). Female board representation and corporate acquisition intensity. *Strategic Management Journal*, 37(2), 303-313.

Conyon, M. J., Hass, L. H., Peck, S. I., Sadler, G. V., & Zhang, Z. (2019). Do compensation consultants drive up CEO pay? Evidence from UK public firms. *British Journal of Management*, *30*(1), 10-29.

Courtemanche, C. J., & Zapata, D. (2014). Does universal coverage improve health? The Massachusetts experience. *Journal of Policy Analysis and Management*, *33*(1), 36-69.

Distelhorst, G., Hainmueller, J., Locke, R. M. (2016). Does Lean Improve Labor Standards? Management and Social Performance in the Nike Supply Chain. *Management Science*, 63(3), 707-728.

Duflo, E., Glennerster, R. & Kremer, M. (2008). Using randomization in development economics research: A toolkit. In P. Schultz & J. Strauss (Eds.), *Handbook of Development Economics*, Volume 4. Amsterdam, The Netherlands and Oxford, UK: Elsevier, North-Holland.

Flammer, C. (2015). Does product market competition foster corporate social responsibility? *Strategic Management Journal*, 38(2), 163-183.

Flammer, C., Kacperczyk, A. (2016). The impact of stakeholder orientation on innovation: Evidence from a natural experiment, *Management Science*, 62(7), 1982-2001.

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*(1), 83-120.

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

Goodman-Bacon, A. (2018). Difference-in-Differences with variation in treatment timing. NBER Working Paper No. 25018.

Holm, J. M. (2018). Successful problem solvers? Managerial performance information use to improve low organizational performance. *Journal of Public Administration Research and Theory*, *28*(3), 303-320.

Hosken, D. S., Olson, L. M., & Smith, L. K. (2018). Do retail mergers affect competition? Evidence from grocery retailing. *Journal of Economics & Management Strategy*, *27*(1), 3-22.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5-86.

Iyer, R., Peydró, J. L., da-Rocha-Lopes, S., & Schoar, A. (2013). Interbank liquidity crunch and the firm credit crunch: Evidence from the 2007–2009 crisis. *The Review of Financial Studies*, *27*(1), 347-372.

Khwaja, A. I., & Mian, A. (2008). Tracing the impact of bank liquidity shocks: Evidence from an emerging market. *American Economic Review*, *98*(4), 1413-42.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P. K. (2016). From social to sale: The effects of firm-generated content in social media on customer behavior. *Journal of Marketing*, *80*(1), 7-25.

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

Lemmon, M., & Roberts, M. R. (2010). The response of corporate financing and investment to changes in the supply of credit. *Journal of Financial and Quantitative Analysis*, *45*(3), 555-587.

Long, S. K., Yemane, A., & Stockley, K. (2010). Disentangling the effects of health reform in Massachusetts: how important are the special provisions for young adults? *American Economic Review*, *100*(2), 297-302.

Pierce, L., Snow, D. C., McAfee, A. (2015). Cleaning house: The impact of information technology monitoring on employee theft and productivity. Management Science, 61(10), 2299-2319.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Schnabl, P. (2012). The international transmission of bank liquidity shocks: Evidence from an emerging market. *The Journal of Finance*, *67*(3), 897-932.

Singh, J., & Agrawal, A. (2011). Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires. *Management Science*, 57(1): 129-150.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?. *Journal of Econometrics*, *125*(1-2), 305-353.

Sommers, B. D., Long, S. K., & Baicker, K. (2014). Changes in mortality after Massachusetts health care reform. A quasi-experimental study. *Annals of Internal Medicine*, *160*(9), 585-594.

White, H., & Raitzer, D. A. (2017). *Impact evaluation of development interventions: A practical guide*. Asian Development Bank.

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.

Wooldridge, J. M. (2012). *Introductory econometrics:* A modern approach (5th ed.). Mason, OH: South-Western College Publisher.

Younge, K.A., Tong, T.W., & Fleming, L. (2014). How anticipated employee mobility affects acquisition likelihood: Evidence from a natural experiment. *Strategic Management Journal*, 36(5), 686-708.