Methods to Estimate Causal Effects

An Overview on IV, DiD, and RDD and a Guide on How to Apply them in Practice

Matthias Collischon¹

¹ Institut für Arbeitsmarkt- und Berufsforschung (IAB), Research Department PASS

Regensburger Straße 100, 90478 Nürnberg

February 1, 2022

Abstract

The identification of causal effects has gained increasing attention in social sciences over the last years and this trend also has found its way into sociology, albeit on a relatively small scale. This article provides an overview of three methods to identify causal effects that are rarely used in sociology: instrumental variable (IV) regression, difference-in-differences (DiD), and regression discontinuity design (RDD). I provide intuitive introductions to these methods, discuss identifying assumptions, limitations of the methods, promising extension, and present an exemplary study for each estimation method that can serve as a benchmark when applying these estimation techniques. Furthermore, the online appendix to this article contains Stata and R syntax that shows with simulated data how to apply these techniques in practice.

Keywords: causal effects, regression discontinuity design, difference-in-differences, instrumental variables

Supplementary material is available at: <https://osf.io/c7pwk/>

Acknowledgements

The author would like to thank Andreas Eberl, Markus Nagler, Malte Reichelt and Irakli Sauer for their helpful comments and suggestions.

Introduction

Over the last years, social sciences have experienced a shift towards the identification of causal effects. This paradigm shift is most prominent in economics where some refer to it as the credibility revolution (Angrist and Pischke, 2010). It gained so much traction that three of its main protagonists – David Card, Joshua Angrist and Guido Imbens – even won the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 for their empirical and methodological contributions to this field. In quantitative sociology, however, causal analyses are still relatively rare, despite relatively recent contributions that provide comprehensive overviews on causal methods (Gangl, 2010; Bollen, 2012; Legewie, 2012; Morgan and Winship, 2015). While the identification of causal effects is certainly not the only goal of sociology, as there is a huge amount of interesting and relevant qualitative, theoretical or descriptive empirical work, causal analyses are still relevant to test the existence of certain channels that are causally implied by theory or to evaluate policies.

This article provides an overview of three methods to identify causal effects - instrumental variable (IV) estimations, regression discontinuity designs (RDD), and difference-in-differences (DiD) - that can be used to analyze a wide variety of topics that are of interest for sociologists. For example, these methods have been used to investigate whether women's quotas affect gender stereotypes (Paola, Scoppa and Lombardo, 2010), whether class size affects students' achievements (Angrist and Lavy, 1999), whether and how income affects crime (Watson, Guettabi and Reimer, 2019) and whether MTV's 16 and pregnant can reduce teenage birth rates (Kearney and Levine, 2015)¹. Thus, these methods can be used in a wide range of applications concerning topics that lie at the heart of sociology like education, gender differences, or social inequalities in general and that these methods are currently underutilized,

¹ I would also like to point out that their finding is up for debate: Jaeger, Joyce and Kaestner (2018).

a point that is also made by Bollen (2012). Currently, most economists cover causal analyses of these phenomena, even if there is room for these methods in classical, sociological topics.

Reviewing the full online archive of the American Sociological Review (ASR) as of August 2019 shows that only 2 articles employ a regression discontinuity design, 27 articles use an instrumental variable estimation and 7 articles use a difference-in-differences approach², three methods that are commonly used to identify causal effects. Compared to 204 articles that contain the word combination “fixed effects” and 65 articles that contain “propensity score matching”, this number is seemingly small. Thus, in contrast to methods that rely on holding certain characteristics constant to retrieve causal effects (like matching or fixed effects approaches), methods that rely on exogenous variation to obtain causal effects are far less used.

But the rare application in quantitative studies is not the only issue in sociology when it comes to these causal estimation methods. Researchers often abstain from good practices in applying these methods in papers. For example, in the case of IV, Bollen (2012) criticizes that only a few sociological papers using IV regressions test for typical problems in their estimations, e.g. weak instruments. Of the eleven articles using IV in the ASR released from 2013 on, after the article by Bollen (2012), five articles neither show nor discuss the strength of their instrument, and four articles show a formal test for weak instruments but do not discuss it in the text.

Furthermore, the discussion concerning the exogeneity of instruments is oftentimes insufficient. For example, Lyons, Vélez and Santoro (2013) investigate the connection between immigrant concentration and violence and instrument immigrant concentration with a measure of immigrant concentration in 1990, arguing that “given that prior immigration was measured a decade prior to our measures of violence, it is a predetermined variable that is exogenous by

² For this overview, I simply searched for the terms “instrumental variable”, “regression discontinuity design” and “difference-in-differences” in the ASR online archive and excluded studies that did not conduct such an analysis (e.g., articles would show up in this analysis if they simply described a study using this method in the literature section).

construction” (Lyons, Vélez and Santoro, 2013: 615). However, not every predetermined variable is necessarily a good instrument. It could be the case that there are time-constant unobservables that correlate both with crime at the time of the survey as well as with immigrant concentration in 1990, like the local GDP, that causes an endogeneity problem, but this is not discussed at all. Thus, even after the release of an influential paper (Bollen, 2012) on IV estimation, some issues remain in the application of this method. This also applies to some degree (based on the few data points) to difference-in-differences: of the seven papers published in ASR, two fail to discuss the parallel trends assumption, which is central for the validity of the model specification. Thus, these methods still seem to be not well known in sociology. Furthermore, there seems to be some confusion concerning the application of these methods and awareness of their identifying assumptions in practice.

This article provides guidelines for both. I aim to inform researchers about these empirical methods as well as provide a practical guide on how to use them. I give an overview of three methods to identify causal effects that can be applied to a wide variety of research questions: difference-in-differences (DiD), regression discontinuity design (RDD), and instrumental variable estimations (IV). I describe each method by starting with a potential research question and the pitfalls of simply using OLS to investigate this topic. Then, I explain the intuition and the estimation equation for the respective method. Next, I discuss the identifying assumptions and interesting extensions to the approach. Lastly, I discuss recent exemplary studies that use the estimation techniques for a problem that could also be of interest to sociologists.

Furthermore, I provide Stata and Rsyntax for the three estimation methods in the online appendix to this article. The respective do-files simulate data and intuitively introduce the empirical application of the respective method in Stata and R. I also offer a document with a short overview of testable and non-testable assumptions underlying the methods and the Stata and R commands for the estimations. As anecdotal evidence from colleagues suggests (and this

issue is potentially widespread), there is a disconnect between reading theoretically about these methods and applying them in practice. The appendix to this article aims at closing this gap and thus make these methods easily accessible for anyone interested.

Because my overview of the methods is in no way universal, I also provide recommendations for further reading in every case. In writing this article, I heavily relied on Angrist and Pischke (2009) and recommend this book to anyone interested in further reading for all three methods described in this paper and will not mention this book specifically in the further reading-sections. In addition, two books have recently been published that provide a good introduction to causal methods:: *Causal Inference – The Mixtape* by Cunningham (2021) and *The Effect* by Huntington-Klein (2021). Both books are also available freely online at the authors’ web pages and are highly recommended. For more technical introductions, I recommend Abadie and Cattaneo (2018) and Athey and Imbens (2017) on the current state of program evaluation methods in economics.

There are multiple approaches to understanding causality. This paper draws heavily from the economics literature which, at least during the last years, primarily relied on the potential outcomes framework (Rubin, 2005). In general, it understands causal effects as the difference between an observed outcome and an unobserved outcome if a given intervention had not taken place (e.g. a schooling reform). The main challenge is thus to get a grasp on the unobserved counterfactual situation.

I at least want to mention that there several approaches to think about causality, like Lewis’ (1973) “closest possible worlds” and the manipulability theories (Woodward, 2005). A very prominent, recent framework are directed acyclical graphs (DAGs) that are now widely used (Pearl, Glymour and Jewell, 2016; for an intuitive introduction see Pearl and Mackenzie, May 2018). DAGs can be used to show causal paths and thus display problems of omitted variables and other potential pitfalls in the analysis. In the section on Instrumental Variable Regression,

I also use a DAG for intuition. However, because the large part of the literature relies on the potential outcomes framework, this guide also mainly sticks to it for simplicity with regards to terminology. Nonetheless, DAGs play an important role in the current literature on causal analysis and should be kept in mind. For interested readers, Imbens (2020) provides a comparison of the potential outcome and DAG frameworks.

Instrumental Variable (IV) Regression

Intuition and Estimation

Suppose we are interested in the effect of education on life satisfaction, more specifically whether education has a positive effect on life satisfaction. In this case, simply regressing life satisfaction on years of schooling is problematic. There are likely omitted variables, like health problems, that affect life satisfaction and could affect schooling. There could also be reverse causality: happy people could simply select into higher education and not vice versa. Thus, simply running an OLS regression will not yield the causal effect that we are interested in. However, under certain circumstances, an IV estimation provides a good way to estimate the causal effect of interest.

The basic idea of an IV estimation is to create exogenous variation in an otherwise likely endogenous variable. For example, in the case of schooling, which is likely endogenous, compulsory schooling reforms that shorten or lengthen the duration of schooling are typical instruments (e.g. Oreopoulos, 2007). We can use such instruments to generate exogenous variation in the endogenous regressor in a first stage regression:

$$x_i = \sigma_0 + \sigma_1 z_i + u_i$$

In this example, x is the endogenous variable (e.g. schooling) and z is the exogenous instrument (e.g. a dummy for an educational reform that takes the value 1 for those affected by the reform). We can now, for the full sample, predict values for x and use these as the regressor in the second stage regression:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \epsilon_i$$

This way, we use exogenous variation in an otherwise endogenous variable to estimate the causal effect of x on y . One can simply use OLS to estimate both stages, this version is referred to as the two-stage least squares (2SLS) estimator.³ Most statistical programs also contain a pre-build package to estimate IV regressions without having to calculate both stages manually. For example, in Stata, this method is implemented via the `ivregress`-command (see also the Online Appendix to this paper with example code).

Figure 1 shows how IV works graphically. A confounder (unobserved) prevents the simple regression of y on x to estimate the causal effect. However, an exogenous instrumental variable z creates exogenous variation in x that can be used to estimate the causal effect of x on y , i.e. variation that is not subject to biases by omitted variables.

Figure 1: The graphical intuition behind instrumental variable regressions.

In the example discussed above, we can instrument schooling with a compulsory schooling reform dummy and then use the predicted values from this first stage regression as our measure for education in the second stage regression, thus generating exogenous variation in schooling.

³ It is also possible to use GMM instead of OLS, which is more efficient in the presence of heteroscedasticity Hansen (1982). Furthermore, one can also use maximum likelihood methods (Anderson (2005)). However, I will not more specifically explain this version since 2SLS is the most intuitive estimator and common in the literature.

Thus, we just use variation in the duration of schooling that occurs due to the schooling reform, which is likely exogenous. This yields the causal effect of interest.

It is also easy to use control variables in the IV just like in a standard OLS model. In the language of the IV literature, control variables are referred to as exogenous regressors (even though they are not necessarily exogenous) or instruments for themselves. However, we need to control for the full set of control variables in both stages of the regression to effectively isolate the variation in x that is generated solely through the instrument. Using relevant control variables can both increase the efficiency (if they are relevant for Y and in the first stage estimation) as well as the consistency of the estimation (if they account for potentially endogenous variation in the instrument).

It is also possible to use more than one instrument in the first stage regression. In this case, the model is referred to as overidentified. This boosts the precision of our second-stage estimation but, in practice, it is hard to find more than one exogenous instrument, as the instrument has to affect the endogenous variable, but not the second stage outcome directly and has to be exogenous. These are high bars, even for finding one instrument.

Furthermore, it is also straightforward to instrument more than one endogenous variable. This results in multiple first stage regressions, one for each endogenous regressor. For identification, however, one requires at least as many instruments as endogenous regressors.

Note that the standard errors of the second stage estimation need to be corrected (Angrist and Pischke, 2009: 138–40). However, most software packages (including Stata) do this by default.

Identifying assumptions

Several conditions need to hold for IV estimations to produce causal effects. First, the instrument has to be exogenous. In the case of using policies as instruments, for example, the instrument is not exogenous if a second reform (e.g. regarding the example above, if one reform

affects the length of schooling and a second reform affects school curricula) happens simultaneously and both are likely correlated with the endogenous variable (in the previously mentioned example both could correlate with schooling).

The second assumption is relevance. This simply means that the instrument is highly correlated with the endogenous variable. A violation of this assumption leads to inconsistent estimates (Bound, Jaeger and Baker, 1995). The assumption is typically tested with an F-test of the coefficient of the instrument in the first-stage regression (in a case with only one instrument, the F-statistics is equivalent to the square of the t-statistics). When the F-statistic is small (below 10 is considered too small as a rule of thumb), the instrument is considered weak and not suitable for the estimation. A high correlation of the instrument and the endogenous variable in the first stage estimation is generally associated with increased precision of the second stage estimation, so the instrument should be relatively strong. Note that, when the model contains more than one endogenous regressor that is instrumented, one needs to account for this when testing for weak instruments. Stock and Yogo (2005) and Sanderson and Windmeijer (2016) provide solutions for this problem.

Third, the instrument should have no direct impact on y but only through x . In the above-described example, it is likely that the schooling reform affects wages through an increase in the duration of schooling, but not directly. All these assumptions should be tested (in the case of relevance) and discussed (in all cases) in a paper using an IV regression to identify causal effects

The fourth assumption is monotonicity. This means that no individual counteracts the supposed effect of the instrument, i.e. individuals do react adversely to the instrument and not in one direction only. In the case of schooling and a schooling reform, this assumption would be violated if any individual got less schooling due to a schooling reform that increases schooling duration compared to a counterfactual case in which there was no reform. These are called

defiers in contrast to compliers who act as expected (i.e. increase their schooling duration due to the reform).

Limitations and practical problem

The most obvious problem is finding exogenous instruments that are also reasonably strongly correlated with the endogenous variable. Researchers should spend a good share of their paper discussing the exogeneity of their instrument.

Furthermore, IV estimation results are biased when the instrument is weak (Angrist and Pischke, 2009, chapter 4.6). Thus, a weak correlation between the instrument and the endogenous variable does not only result in an imprecise estimation of the second stage but also inconsistency.

IV estimations only identify a local average treatment effect (LATE), i.e. an effect just for the compliers to the instrument (Imbens and Angrist, 1994). This means that the effect is not necessarily externally valid; this should be discussed in a paper. For example, the impact of schooling on wages identified through an IV-estimation using a schooling reform provides likely not the generalizable, causal effect of schooling on wages for the general population, but just for the subgroup of individuals who react to this specific reform (i.e. IV estimations provide a local estimate).

Extensions

Given one has an overidentified model (i.e. more instruments than endogenous regressors), it is possible to use the Sargan-test (Sargan, 1958) to test whether instrumental variables are uncorrelated with residuals from the regression. However, while researchers should know of this test, I do not necessarily recommend using it. Often, in overidentified models, the instruments used follow a comparable rationale, e.g. they are all firm characteristics. Thus, it is likely the case that all are exogenous or none is endogenous. In this case, the Sargan-test is

relatively useless. I consider it far more important to argue for the exogeneity of potential instruments logically.

As previously described, IV does not allow for a direct impact of the instrumental variable on the second stage outcome, but only channeled through the endogenous variable. Oftentimes, this assumption is problematic. Conley, Hansen and Rossi (2012) provide a method to assess the potential bias if there is a direct relation between the instrument and the outcome of interest. This method allows assessing the magnitude of the bias if one assumes that the instrument also affects the second stage outcome directly by assuming that a specific share of the overall effect is driven by this partial correlation. Thus, this method allows to simulate the bias of the estimation if the instrument also affects the outcome of interest directly, not only through the endogenous variable. I recommend using this method to test the robustness of the results.

Recently, a literature on marginal treatment effects (MTEs) emerged (Cornelissen *et al.*, 2016). In general, the effect identified by an IV estimation with a continuous instrument can be regarded as a weighted average of various individual LATEs. In many cases, there is likely treatment effect heterogeneity across the unobserved “resistance to treatment”, i.e. the degree to which the instrument affects treatment probability. For example, Cornelissen *et al.* (2018) estimate the effect of a preschool program in Germany on children’s outcomes that was introduced staggered across Germany. Their instrument is the availability of a childcare slots (which were increased by a policy program regionally), their main outcome is a measure of school readiness and the endogenous variable is attending early childcare. The idea of MTEs is, intuitively speaking, to estimate the unobserved propensity to attend childcare in absence of the intervention using matching methods. One can now estimate MTEs across this distribution. The authors conclude that children that are least likely to attend childcare benefit the most from it, in contrast to children who are most likely to attend formal childcare anyways. Thus, there

is substantial treatment effect heterogeneity that sheds light on results that could be especially important for policy implications.

Example study: Cygan-Rehm and Wunder (2018)

Cygan-Rehm and Wunder (2018) study the causal effect of working hours on health. As working hours are endogenous, they use statutory working hour requirements for public-sector employees as an instrument for actual working hours. Furthermore, they use individual fixed effects to account for time-constant unobserved heterogeneity. They find that more working hours causally affect subjective and objective health measures, especially for individuals with small children.

This article is a good example of the usage of IV regressions. The instrument is relevant, with an F-value of 20 in the baseline specification of the first stage, and monotonicity is likely given (it is unlikely that individuals work less when statutory working hours increase). Furthermore, it is hard to think of a story of how the instrument is endogenous, and actual working hours are likely the only channel through which statutory working hours could affect health.

Further reading

Angrist and Pischke (2015: 98–146) provide probably the most intuitive approach to IV regression overall in their undergraduate book. For a good overview of IV designs in sociology, I recommend Bollen (2012). Furthermore, I also recommend the article by Mogstad and Torgovitsky (2018), which is more advanced. Both provide relatively in-depth overviews on IV.

Angrist (1990) is an applied example of IV-estimation from a Nobel laureate of 2021. In this paper, Angrist estimates the effect of veteran status on civilian earnings. As veteran status is not randomly assigned, he uses the draft lottery during the Vietnam War to generate exogenous

variation in veteran status. This paper is an early, accessible and interesting approach that shows the potential of IV-estimation.

Difference-in-Differences (DiD)

Intuition and Estimation

Suppose we are interested in the effect of female managers on the gender wage gap in establishments. We could simply run a regression in which we regress the within-firm gender wage gap on the share of female managers in the firm. However, our estimate of the effect is likely biased. For example, firms that hire female managers could just be female-friendly overall due to some unobserved factor and thus exhibit a smaller gender wage gap. Alternatively, firms that already have a relatively small gender wage gap could just be more attractive for high-potential females which leads to more female managers. In both cases, a simple regression does not yield the causal effect. However, a difference-in-differences (DiD) design provides a good tool to estimate the causal effect we are interested in.

The intuition behind a DiD estimation is relatively simple and resembles an experimental design. For a basic DiD estimation, we need two groups (one treatment, one control) and two time periods (one before the treatment, one after). We are now interested in the causal effect of the treatment on an outcome Y . To obtain this causal effect, we simply subtract the difference in the outcome variable between the two groups before treatment from the difference after the treatment. This difference gives us the DiD-estimation. In regression terms, it can be written as:

$$Y_{it} = \beta_0 + \lambda(treat_i \cdot post_t) + \beta_1 treat_i + \beta_2 post_t + \epsilon_{it}$$

Where $treat$ is 1 for the treatment group and $post$ is 1 for observations after the treatment occurred. DiD resembles an experiment in which we eliminate systematic differences in the outcome between the groups by subtracting the pre-treatment difference. Note, however, that

we do not necessarily need panel data for a DiD-estimation; repeated cross-sections work as well assuming that the composition of treatment- and control groups does not change over time.

Figure 2 visualizes the idea behind DiD. The y-axis is an outcome of interest and the x-axis is the time dimension. At some point in time, indicated by the red vertical line, the treatment happens and only affects the treatment group (the solid line), while nothing changes for the control group. The causal effect of the treatment is the difference between the treatment and control group in the outcome of interest after the treatment minus the difference between the groups before the treatment. The treatment effect is indicated by the orange arrow in Figure 2. The figure also shows (as the orange dotted line) the supposed course of the outcome over time for the treatment group if the treatment had not happened. The critical assumption here is that, were there no treatment, treatment, and control group would develop in parallel over time (this assumption is further discussed below). While Figure 2 presents a case with multiple periods (five in this case), it is also possible to estimate it with only two periods, one prior and one after the treatment. However, to illustrate the parallel trends assumption, this example contains multiple periods.

Figure 2: Visualization of the intuition behind difference-in-differences estimations.

Coming back to the question of female managers and the gender wage gap, we need a reform that, for example, sets a quota for females in management for firms of a given size, e.g. firms with 500 employees. Given that we have data on firms both prior and after the reform, we could view firms with more than 500 employees as the treatment group and firms below 500 employees as the control group. We can then calculate the difference in the mean gender wage gaps between these groups after the reform and subtract the difference between these groups before the reform. This gives us the causal effect of the reform on the gender wage gap which allows us to make statements about the impact of female managers on gender wage inequality.

DiD can also easily be estimated by using more than two periods or with fixed effects of any kind. For example, when using multiple years, one should include survey-year dummies instead of the post-dummy and thus use more information in the estimations (thus modeling the overall time trend more flexibly).

DiD is widely applied in economics for the evaluation of policies. For example, it has been used to analyze the impact of minimum wages on employment (by comparing two adjacent US states, one of which introduced a statutory minimum wage, while the other did not, Card and Krueger, 1994) – an application that also contributed to David Card winning the Nobel Prize - and the impact of migration on local labor markets (Card, 1990).

Identifying assumptions

For consistency of DiD, several conditions need to hold for a causal interpretation. First, in absence of the treatment, the mean outcome of Y would have developed in parallel for both groups. This assumption is generally referred to as the parallel or common trends assumption. It would be harmed if, for example, the trends in the outcome between the groups, e.g. the gender wage gaps in between firm size groups in our example, develop differently over time. Typically, this assumption is made plausible by investigating the trends in Y over time for the treatment and control group before the treatment. In the before-mentioned example in Figure 2,

the treatment and control group move in parallel before the treatment. The assumption is that this would have continued in absence of the treatment. However, at this point, it is important to note that the assumption in itself is untestable, as it would require the treatment group without having received the treatment. Investigating parallel trends before the treatment can be used to discuss the plausibility of this trend continuing if there were no treatment.

Second, the stable unit value treatment assumption (SUTVA) needs to hold. That essentially means that there should not be any spillover effects of the treatment to the control group. For example, when evaluating a reform e.g. sector-specific minimum wages, there could be spillovers from the treatment sector to other sectors if these sectors are potential substitutes.

Third, it is important to discuss whether there are any confounding factors. For example, in the case of policy evaluations, researchers should discuss whether any reforms happen contemporaneously with the reform of interest

Fourth, there should be no anticipation effects. This is also typically discussed in the text. Ideally, the treatment is relatively unexpected so that the treatment group cannot react before the treatment is implemented.

Limitations and Caveats

Typical problems in DiD estimations are a violation of the above-described parallel trends assumption. In practice, this is typically tested graphically by plotting the means for both groups overtime before the treatment. This assumption can also be tested more formally by using placebo treatments. For example, if the treatment happened in 2011 one can easily test whether shifting the treatment period to 2009 in a placebo estimation also shows any treatment effect. We should be worried that the parallel trends assumption is violated if there is a non-zero effect in the placebo year.

Another assumption is that the treatment should not affect the group composition. Thus, there should be no time-varying selection into the treatment. This is also a violation of the parallel trends assumption because the effect of the treatment could just be an effect of changing sample composition. Coming back to our example, it could happen that firms with 501 employees before the treatment simply fire employees (or do not hire new employees) to drop below the 500-employees threshold. In a typical DiD paper, authors should present balancing tables showing means e.g. socio-demographic characteristics for the treatment and control group before and after the treatment occurs or use these variables as outcomes in the estimation; there should be zero effects.

Control variables can also be simply included in a DiD-estimation. However, one should be cautious when picking control variables that are not pre-determined: they could also be affected by the treatment. Controlling for them would then bias the estimation of the total causal effect of the treatment. Thus, one should motivate every control variable in the estimation against this backdrop.

Furthermore, it is not easy to decide how to calculate standard errors for DiD estimations. Bertrand, Duflo and Mullainathan (2004) as well as Angrist and Pischke (2015: 205–08) provide some guidelines for this issue. Bertrand, Duflo and Mullainathan (2004) generally recommend collapsing the data into pre-and post-periods with a small number of cases and block bootstrapping with a large number of groups.

Extensions

DiD can be easily extended to a triple-differences design (DiDiD) when additional groups are available. For example, suppose the imaginary reform discussed previously is a state-level law in the US and some adopt it while others do not. Now we could use variation within states (pre and post and between different firm size groups) and variation between states (states that

implement the law and states that do not). The estimation equation, in this case, would look like this:

$$Y_{it} = \beta_0 + \gamma(treat_i \cdot post_t \cdot state_i) + \beta_1 treat_i + \beta_2 post_t + \beta_3 state_i \\ + \beta_4(treat_i \cdot state_i) + \beta_5(treat_i \cdot post_t) + \beta_6(state_i \cdot post_t) + \epsilon_{it}$$

In this case, γ yields the causal effect of interest and should not differ from the estimated λ in the DiD estimation. This also serves as a further robustness check of the parallel trends assumption and can also be applied to test the robustness of the estimate against potential confounding events.

Furthermore, DiD can be easily combined with matching when the parallel trends assumption fails. This way, one can use matching to identify treatment and control cases specifically based on parallel trends before treatment and then use this sample in the DiD estimation (e.g. Marcus, 2013).

Additionally, it is possible to estimate a two-way fixed effects model (with fixed effects for time and observation units) when there are multiple treated units with variation in treatment timing (imagine for example the staggered introduction of a schooling reform across federal states in a country, e.g. Marcus and Zambre, 2019). In this case, the treatment effect is the weighted average of the individual DiD-estimates. The potential heterogeneity in these estimates creates several pitfalls; Goodman-Bacon (2021) provides a recent and comprehensive introduction to these models. Cameron and Miller (2015) further provide guidance on clustering standard errors in this case.

Example: Nollenberger and Rodríguez-Planas (2015)

Nollenberger and Rodríguez-Planas (2015) investigate the effect of an expansion of full-time subsidized childcare slots for three years olds in Spain during the 1990s on maternal labor market participation. Regional variation on municipality levels concerning the reform allows

the authors to estimate a DD-analysis, thus comparing mothers whose youngest child is three years old living in regions where the reform was introduced to regions where the reform was still not in effect. Furthermore, the authors choose a triple differences (DiDiD) design as their baseline specification, where they also include mothers whose youngest child is two years old and who are not affected by the reform. Their results suggest that the reform increased the maternal labor market participation of mothers of three-year-olds by 2.8 percentage points which correspond to a relative increase of around 10%.

Nollenberger and Rodríguez-Planas (2015) provide a model example of how to conduct a DiD estimation and test the robustness of it. They graphically investigate the parallel trends assumption and use a placebo treatment to validate their finding. Furthermore, they show that it is also important to be aware that differential time trends between treatment and control groups could bias the results and solve this problem with the DiDiD analysis and DiD with linear time trends by region. They also show an alternative DiDiD specification in which they include mothers of four and five-year-olds instead of two-year-olds as an additional group. Furthermore, in their appendix (Table A.4 in the paper), the authors also show that the reform did not change the sample composition, which is also crucial. Overall, this study serves as a good template for any DiD analysis.

Further reading

Angrist and Pischke (2015: 178–205) provide a comprehensive overview on DiD in *Mastering Metrics*. Gangl (2010) also provides a comprehensive overview of this method in his article. Furthermore, Havnes and Mogstad (2011) also provide an empirical study of the effects of an expansion of publicly subsidized childcare in Sweden on maternal labor supply that is a model example of how to conduct a DiD-analysis. However, the definition of treatment and control groups in their case is less accessible compared to Nollenberger and Rodríguez-Planas (2015),

because it rests on living in municipalities that increased their childcare supply above or below the overall median.

Regression Discontinuity Design (RDD)

Intuition and estimation

A core interest in sociology is gender inequalities. One of the most prominent examples is the motherhood wage penalty, i.e. women earning less after the birth of a child. It is hard to pin down the reason for this. Is it a decay of human capital due to the time out of work after childbirth that leads to cumulated wage losses? Is it employer discrimination?

Let us focus on the role of the time out of work after childbirth for this one and take the example of Germany. If we are interested in empirically estimating the effect of time out of employment on maternal wages, we encounter some serious concerns. For example, selection is a huge problem. The length of the employment interruption is not random and mothers with high earnings potentials are likely to return fast. However, perhaps we can exploit a legislative change. In December 2006, Germany passed a law that changed the system of parental leave. Before this law, parents of children could decide whether they want two years of paid parental leave with 300€ per month (the default) or one year with 450€ per month. The law changed this system to a subsidy of 66% of the mean net monthly wages of the parent taking parental leave and limited the period for one recipient to twelve months. Thus, it incentivized shorter parental leave periods. The new system was in place only for parents whose child was born from January 1 2007 on. Thus, there is a sharp cutoff rule for this law that was also enforced. This makes this reform an ideal candidate to investigate the impact of times out of the labor force on mother's labor market outcomes using a regression discontinuity design.

First popularized in Psychology (Thistlethwaite and Campbell, 1960), RDDs exploit a discontinuous change in a potential treatment status along with a distribution of a continuous variable and then compares the outcome of interest for observations just below or above the

threshold (in our example: mothers giving birth just before and after the reform). Net of trends around the cutoff, the difference between observations just below and above the cutoff is the causal estimate of the treatment effect. A typical estimation equation for a RDD can be written as:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 c_i + \epsilon_i$$

Where D is a binary variable indicating whether an observation is above the critical threshold and c is the continuous variable the cutoff is based on. In the case of our example, the running variable would be the day of birth of the child, D would be 1 for children born from January 2007 onwards and 0 before. We can identify the effect of the policy change on maternal labor market outcomes e.g. 1 year after birth when, for example, comparing mothers with children born in December 2006 to mothers with children born in January 2007. The coefficient for β_1 then yields the causal effect of the policy change, i.e. the causal effect of the parental leave legislation on maternal wages. In this example, the cutoff is sharp: the policy change does not apply to any mothers with children born in 2006, but to all mothers with children born from 2007 onwards. This is why this version of RDD is also referred to as *sharp* RDD. RDD can also be estimated when the treatment probability only increases at a cutoff, but not perfectly from 0 to 1. This case is called *fuzzy* RDD and will be discussed in a separate subsection.

Figure 3 shows the general intuition behind RDD visually. The x-axis denotes any running variable (e.g. date of birth) with a treatment that only applies from a certain threshold in the running variable (e.g. the eligibility for a certain subsidy). The y-axis shows the values of an outcome of interest. The causal effect of the treatment can be identified by comparing the outcomes of observations just below and above the threshold. One should also account for trends in the outcome caused by the running variable by control for these (more on that in the next section).

Figure 3: The intuition behind a RDD estimation.

In this simple example, we assume that Y is linear and continuous in c , the running variable. However, this is not necessarily the case. We could also, in the regression, model Y as any function of c (and we need to do so for correct identification, more in *Limitations and Caveats* to this method). Furthermore, it could also be the case that the slope changes around the threshold. This is, in general, no problem: we could simply model different functions below and above the threshold and still identify the discontinuity.

Identifying assumptions

The central assumption to estimate a causal effect via RDD is continuity. This assumption requires that were there no threshold, the outcome Y is a continuous function of the variable- at least around the threshold value, in the observation window used in the analysis. Thus, the central assumption in RDD is that we correctly specify the functional form of Y across the running variable.

Furthermore, RDD requires that there is no selection of individuals just around the cutoff, i.e. the distribution of individuals around the cutoff is as good as random concerning individual

characteristics. This is also one of the attractive properties of a sharp RDD design because it allows for non-random-selection except around the cutoff (Lee, 2008).

An implication of RDD, as discussed by Imbens and Lemieux (2008), is that there is no value of the running variable for which there are both treatment and control observations. In contrast to e.g. DiD, this prevents RDD to be used with matching, because there is no common support that is required for matching.

Limitations and Caveats

For a causal interpretation, an RDD estimation requires that there is no manipulation around the cutoff. In our case, manipulation is unlikely, because the reform was passed in December 2006 and affected parents with children born in January 2007 and later. However, it is theoretically possible, if the policy is announced well before it comes into effect, that parents plan the time of gestation so that they are or are not affected by the reform. In this case, observations just above and below the cutoff would not be comparable, because there are likely systematic differences between the groups. In this case, the estimate of the RDD design is also likely biased. McCrary (2008) proposed a test for this manipulation around the cutoff that is widely used; however, the newer estimator by Cattaneo, Jansson and Ma (2020) is more robust (Kuehnle, Oberfichtner and Ostermann, 2021) and should be used. Furthermore, as previously described for DiD estimations, RDD papers typically contain balancing tables that show the sample composition in terms of socio-demographic characteristics around the cutoff.

Additionally, the correct identification of the treatment effect relies on correctly identifying the discontinuity at the threshold. This requires the trends around the cutoff to be correctly specified. One can either do this parametrically or non-parametrically. The former requires exactly specifying the correct functional form of the running variable and leads to noisy estimates (Gelman and Imbens, 2019). In the case of misspecification, the estimated treatment

effect is biased. Using a non-parametric approach with local polynomial regressions is thus now a standard solution (Hahn, Todd and van der Klaauw, 2001).

Like in the DiD-setting, placebo treatments can and should be used to assess the credibility of the estimation results. In the case of RDD, this can be done by simply choosing an arbitrary threshold where there should be no treatment effect and rerunning the estimation.

Another concern regarding the correct identification of the discontinuity is what window around the threshold is used in the estimation. This is always a trade-off between efficiency and potential biases. A narrower window around the cutoff potentially provides a better estimation of the treatment effect but is typically associated with increasing standard errors because the number of observations is relatively small. Concerning the trends around the cutoff, a smaller window should be generally less sensitive to the identification of the trend in the running variable. For example, when using observations just below and above the threshold, trends should generally not have a huge effect on the comparison of these observations. But, as previously mentioned, this also leads to a loss of observations and thus less power.

Using control variables in RDD is possible, but one should test that control variables are not affected by the treatment to avoid biases. This is comparable to using control variables in a DiD-estimation.

Extensions

In some cases, the treatment in a RDD coincides with other events that could affect the outcome of interest that happen regularly, e.g. any reform could happen parallel to the start of the schooling year. In this case, one can easily combine RDD with DiD and thus account for the event that happens regularly (e.g. Cygan-Rehm, Kuehnle and Riphahn, 2018; Dustmann and Schönberg, 2012). Furthermore, given that the connection between the running variable and the outcome of interest follows the same pattern every period, this method makes exactly identifying the functional form not critical.

Fuzzy RDD

In this paper, I focus on sharp RDD, the ideal case in which the treatment status is perfectly determined by the threshold. Another version of RDD is the so-called fuzzy RDD. In the case of a fuzzy RDD, the threshold is merely associated with an increased probability to be treated. A fuzzy RDD design then uses an IV approach to instrument the endogenous variable of interest with the exogenous treatment in the first stage and then calculate the second stage as is the case in IV. For further reading on fuzzy RDDs, I recommend Angrist and Pischke (2009: 259–67) and Chapter 6.2.7 of Cunningham (2021).

Example: Kuehnle and Wunder (2016)

Kuehnle and Wunder (2016) investigate the impact of the daylight savings time transition on individual life satisfaction. They investigate both the switch from winter to summertime and from summer to winter time in their analysis. They use variation in the timing of the interviews around the transition periods to identify effects on life satisfaction. Kuehnle and Wunder use the SOEP and BHPS in their analysis and just look at the weeks around the respective daylight saving time transition and compare the mean life satisfaction of individuals who were surveyed within the two weeks around the transition to daylight savings time. They find negative effects of the spring transition in both the UK and Germany (which effectively shortens one night by one hour) in the first week after the transition of no effects of the autumn transition (which they can, due to the interview timing, only identify in the BHPS).

This paper is very intuitive and convincing. The treatment is plausibly exogenous and sorting around the threshold would require respondents that differ systematically in their levels of life satisfaction to be surveyed systematically before or after the transition and the authors can plausibly show (with balancing tables) that this is not the case. Furthermore, the paper also uses placebo treatments that arbitrarily shift the timing of the transition and show no effect.

Further reading

Angrist and Pischke (2015: 178–205) also discuss RDDs in their undergraduate book. The JEL article by Lee and Lemieux (2010) also provides an intuitive introduction to RDD for applied research. Furthermore, I recommend Lee (2008) who investigates the causal effect of incumbency on electoral outcomes for US house elections and also provides a very comprehensive introduction to the method in the paper. van der Klaauw (2008) provides an accessible overview and discussion on RDDs in economics.

Conclusion

The identification of causal effects has gained rising importance in the social sciences during recent years. However, causal analyses in quantitative sociology are still rare and often not conducted in the best manner possible. This article provides an overview of three causal methods that can easily be applied in a variety of cases: instrumental variable regressions, difference-in-differences models, and regression discontinuity designs. I present the intuition behind the methods, the identifying assumptions, typical problems and caveats in applications, exemplary studies, and further reading recommendations for each method. Furthermore, the online appendix to this paper contains Stata syntax with simulated data that shows the empirical application of each respective method step by step and can serve as a template for empirical analyses.

However, even when applied rightly, researchers should still be cautious and transparent in the presentation of the results. A recent paper by Brodeur, Cook and Heyes (2020), for example, shows that especially publications that use IV regressions seem to be susceptible to p-hacking and that there seems to be substantial publication bias. Brodeur *et al.* (2016) provide additional evidence for a substantial amount of p-hacking and/or publication bias for three top economics journals. The implication is that a valid identification strategy alone does not solve all problems that potentially arise in quantitative empirical research and that researches should be careful to avoid these biases.

The goal of this paper is to be a short introduction to these methods of causal analyses for a wide audience. I hope that this guide is a useful tool for any sociologist interested in estimating causal effects and leads to the avoidance of typical errors conducted in such analyses.

References

- Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, **10**, 465–503.
- Anderson, T. W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, **127**, 1–16.
- Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review*, **80**, 313–336.
- Angrist, J. D. and Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, **114**, 533–575.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics. An empiricist's companion*. Princeton, NJ: Princeton Univ. Press.
- Angrist, J. D. and 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**, 3–30.
- Angrist, J. D. and Pischke, J.-S. (2015). *Mastering 'metrics. The path from cause to effect*. Princeton, NJ, Oxford: Princeton University Press.
- Athey, S. and Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, **31**, 3–32.
- Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, **119**, 249–275.
- Bollen, K. A. (2012). Instrumental Variables in Sociology and the Social Sciences. *Annual Review of Sociology*, **38**, 37–72.
- Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, **90**, 443–450.
- Brodeur, A., Cook, N. and Heyes, A. (2020). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, **110**, 3634–3660.
- Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, **8**, 1–32.
- Cameron, C. A. and Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, **50**, 317–372.
- Card, D. (1990). The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relations Review*, **43**, 245.
- Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, **84**, 772–793.
- Cattaneo, M. D., Jansson, M. and Ma, X. (2020). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, **115**, 1449–1455.
- Conley, T. G., Hansen, C. B. and Rossi, P. E. (2012). Plausibly Exogenous. *Review of Economics and Statistics*, **94**, 260–272.
- Cornelissen, T., Dustmann, C., Raute, A. and Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, **41**, 47–60.
- Cornelissen, T., Dustmann, C., Raute, A. and Schönberg, U. (2018). Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy*, **126**, 2356–2409.
- Cunningham, S. (2021). *Causal Inference: The Mixtape*. New Haven, Connecticut: Yale University Press.
- Cygan-Rehm, K., Kuehnle, D. and Riphahn, R. T. (2018). Paid parental leave and families' living arrangements. *Labour Economics*, **53**, 182–197.

- Cygan-Rehm, K. and Wunder, C. (2018). Do working hours affect health? Evidence from statutory workweek regulations in Germany. *Labour Economics*, **53**, 162–171.
- Dustmann, C. and Schönberg, U. (2012). Expansions in Maternity Leave Coverage and Children's Long-Term Outcomes. *American Economic Journal: Applied Economics*, **4**, 190–224.
- Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology*, **36**, 21–47.
- Gelman, A. and Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, **37**, 447–456.
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.
- Hahn, J., Todd, P. and van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*. *Econometrica*, **69**, 201–209.
- Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50**, 1029.
- Havnes, T. and Mogstad, M. (2011). Money for nothing? Universal child care and maternal employment. *Journal of Public Economics*, **95**, 1455–1465.
- Huntington-Klein, N. (2021). *The Effect. An Introduction to Research Design and Causality*. London: Taylor & Francis.
- Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, **58**, 1129–1179.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**, 467.
- Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142**, 615–635.
- Jaeger, D. A., Joyce, T. J. and Kaestner, R. (2018). A Cautionary Tale of Evaluating Identifying Assumptions: Did Reality TV Really Cause a Decline in Teenage Childbearing? *Journal of Business & Economic Statistics*, **57**, 1–10.
- Kearney, M. S. and Levine, P. B. (2015). Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing. *American Economic Review*, **105**, 3597–3632.
- Kuehnle, D., Oberfichtner, M. and Ostermann, K. (2021). Revisiting gender identity and relative income within households: A cautionary tale on the potential pitfalls of density estimators. *Journal of Applied Econometrics*, **81**, 813.
- Kuehnle, D. and Wunder, C. (2016). Using the Life Satisfaction Approach to Value Daylight Savings Time Transitions: Evidence from Britain and Germany. *Journal of Happiness Studies*, **17**, 2293–2323.
- Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, **142**, 675–697.
- Lee, D. S. and Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, **48**, 281–355.
- Legewie, J. (2012). Die Schätzung von kausalen Effekten: Überlegungen zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, **64**, 123–153.
- Lewis, D. (1973). *Counterfactuals*. Hoboken: Wiley.
- Lyons, C. J., Vélez, M. B. and Santoro, W. A. (2013). Neighborhood Immigration, Violence, and City-Level Immigrant Political Opportunities. *American Sociological Review*. *American Sociological Review*, **78**, 604–632.
- Marcus, J. (2013). The effect of unemployment on the mental health of spouses - evidence from plant closures in Germany. *Journal of health economics*, **32**, 546–558.
- Marcus, J. and Zambre, V. (2019). The Effect of Increasing Education Efficiency on University Enrollment. *Journal of Human Resources*, **54**, 468–502.

- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142**, 698–714.
- Mogstad, M. and Torgovitsky, A. (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics*, **10**, 577–613.
- Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference. Methods and principles for social research*. New York, NY: Cambridge University Press.
- Nollenberger, N. and Rodríguez-Planas, N. (2015). Full-time universal childcare in a context of low maternal employment: Quasi-experimental evidence from Spain. *Labour Economics*, **36**, 124–136.
- Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics*, **91**, 2213–2229.
- Paola, M. de, Scoppa, V. and Lombardo, R. (2010). Can gender quotas break down negative stereotypes? Evidence from changes in electoral rules. *Journal of Public Economics*, **94**, 344–353.
- Pearl, J., Glymour, M. and Jewell, N. P. (2016). *Causal inference in statistics. A primer*. Chichester, West Sussex: Wiley.
- Pearl, J. and Mackenzie, D. (May 2018). *The Book of Why. The New Science of Cause and Effect*. New York: Basic Books.
- Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, **100**, 322–331.
- Sanderson, E. and Windmeijer, F. (2016). A weak instrument Formula: see text-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, **190**, 212–221.
- Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, **26**, 393.
- Stock, J. H. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In Andrews, D. W. K. and Stock, J. H. (Eds.). *Identification and Inference for Econometric Models*: Cambridge University Press, pp. 80–108.
- Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, **51**, 309–317.
- van der Klaauw, W. (2008). Regression–Discontinuity Analysis: A Survey of Recent Developments in Economics. *Labour*, **22**, 219–245.
- Watson, B., Guettabi, M. and Reimer, M. (2019). Universal Cash and Crime. *Review of Economics and Statistics*, 1–45.
- Woodward, J. (2005). *Making things happen. A theory of causal explanation*. Oxford: Oxford Univ. Press.