

UC Berkeley

UC Berkeley Previously Published Works

Title

Natural experiments in computational social science: principles, methods, and applications

Permalink

<https://escholarship.org/uc/item/0wk7534d>

ISBN

9781802207293

Author

Harding, David J

Publication Date

2025-12-11

DOI

10.4337/9781802207309.00036

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nd/4.0/>

Peer reviewed

Natural Experiments in Computational Social Science: Principles, Methods, and Applications

David J. Harding (0000-0002-2121-0790)

¹Department of Sociology, University of California, Berkeley, Berkeley, California, USA

Abstract

A natural experiment is a research design that uses observational data to draw causal inferences by leveraging naturally occurring random variation in the treatment assignment. This chapter presents an overview of natural experiment research designs and instrumental variable methods for analyzing data from natural experiment studies, aimed at an introductory level and presented using directed acyclic graphs, econometric equations, potential outcomes, and an example based on the random assignment of judges to criminal cases to estimate the effects of imprisonment. It covers the logic and terminology of natural experiments, common estimators, and interpretation of estimates under effect heterogeneity. It then explains the three assumptions of an instrumental variables analysis – relevance, exclusion restriction, and monotonicity – and their implications for the design and analysis of natural experiments. Finally, this chapter reviews three specific types of natural experiment designs: regression discontinuity, difference-in-differences, and synthetic controls.

Keywords: Causal Inference, Identification, Instrumental Variables, Regression Discontinuity, Difference-in-Differences, Synthetic Controls

Introduction

A natural experiment is a research design using observational data that attempts to draw causal inferences by leveraging naturally occurring random variation in the treatment assignment. Unlike an experiment in which the investigator controls assignment to treatment, in a natural experiment variation in treatment is generated by institutional processes or processes originating in nature. Other research designs for observational data rely on the assumption that all sources of variation in the treatment that also impact the outcome are properly measured and conditioned (an assumption that is known by different terminology in different disciplines: selection on observables in economics, no unmeasured confounding in epidemiology, and no hidden bias or conditional ignorability in statistics). In contrast, a natural experiment avoids such an assumption but makes other assumptions about the putatively random source of variation in treatment assignment. For researchers, developing a natural experiment requires identifying sources of variation in treatment assignment that arguably satisfy those assumptions, some of which are not empirically testable. The statistical analysis of data from a natural experiment can be understood as isolating variation in treatment due to the source of randomization (the “instrument”) and discarding variation in treatment that is confounded by unobserved variables that also affect outcomes. Such an analysis is often termed an “instrumental variables” (IV) analysis.

Contemporary analysis of natural experiments can be understood using the conceptual tools of the wider literature on causal inference: directed acyclic graphs (DAGs) and potential

outcomes. A DAG is a tool for assessing and understanding causal identification by graphically representing our assumptions about causal relationships between variables (Pearl 2009, Elwert 2013). Potential outcomes, or “counterfactuals,” denote the outcomes that each unit would have experienced had it been exposed to the treatment and had it been exposed to the control condition. A causal effect is the difference in potential outcomes for the same unit (Rubin 1974). We observe only one of the potential outcomes for a specific unit, the one for treatment actually received. For example, for treated units we only observe the potential outcome under treatment. We can understand causal inference as identifying cases that can reasonably proxy for the missing potential outcomes, units that received a different treatment but are otherwise comparable in the sense that they have the same potential outcomes. A natural experiment is one way to find such reasonable comparison cases by leveraging naturally occurring variation in treatment assignment that is independent of potential outcomes.

Natural experiments are becoming increasingly relevant in the digital age and for computational social science. Natural experiments often require large sample sizes to reliably detect causal effects, and the increasing prevalence of large datasets in the digital age will allow computational social scientists to leverage more natural experiments. Furthermore, digital platforms and digital interventions provide opportunities for natural experiment research designs (for example, Hale et al 2018, Ternovski and Yasseri 2020, Fitzpatrick et al 2021).

A Motivating Example

Throughout this chapter, we discuss a motivating example that aims to estimate the causal effect of imprisonment, compared to an alternative punishment like probation supervision in the community, on various outcomes like employment or recidivism. In most courthouses, judges are randomly assigned to criminal cases, and judges have different tastes or preferences for sentencing people convicted of felonies to prison. This means that the “harshness” of one’s randomly assigned judge can serve as an “exogenous” source of variation in imprisonment, the treatment of interest. That is, the judge’s harshness is the instrument. This research design was originally developed by Kling (1995) and used in many subsequent studies (e.g., Loeffler 2013, Mueller-Smith 2015, Harding et al., 2017).

The Logic of a Natural Experiment

Figure 1 illustrates the logic of a natural experiment using a DAG. We wish to estimate the effect of the treatment (D) on the outcome (Y), an effect denoted by β_1 . In the imprisonment example, D is imprisonment, and Y might be recidivism at some later point in time. One research design would be to identify all covariates (X) that affect both D and Y , measure them on a sample of those sentenced to prison and those not sentenced to prison, and then estimate the effect of imprisonment on earnings as the association between D and Y after conditioning on the X s (for example, by regression or matching). However, we worry that there are unobserved covariates (U) that cannot condition directly, leaving the association between D and Y a biased estimate of β_1 because it reflects both the causal effect of interest and the association between D and Y that is due to the U s, common causes of D and Y . For example, U might be defendant’s wealth, which could affect the ability to buy effective legal counsel and one’s need to commit future crime.

The random assignment of judges provides a natural experiment in that the harshness of the randomly assigned judge serves as an instrument (Z) for imprisonment; judge harshness affects whether a convicted defendant is sentenced to prison, as denoted by δ_1 . This means that the same defendant who would be sentenced to prison by a harsh judge would be sentenced to probation by a lenient judge, at least for some subset of the defendants. An instrumental variables analysis estimates the causal effect β_1 by first estimating the effect of the instrument on the outcome (which will provide the product of δ_1 and β_1 , or $\delta_1\beta_1$), and then scaling that by the effect of the instrument on treatment, δ_1 . That is, $\beta_1 = \frac{\delta_1\beta_1}{\delta_1}$. The effect of the instrument on the outcome is termed the “reduced form,” and the effect of the instrument on the treatment is termed the “first stage.” Here, the effect of imprisonment on the outcome is equivalent to the effect of judge harshness on the outcome divided by the effect of judge harshness on imprisonment. This ratio is one possible estimator for the treatment effect when there are no X s that need to be conditioned: the covariance of Z and Y divided by the covariance of Z and D . When both the instrument and treatment are binary, this is equivalent to the “Wald estimator,” the difference in outcomes for the two groups defined by the instrument divided by the differences in probabilities of treatment between those groups:

$$(\text{Wald estimator}) \beta_1 = \frac{Y_{Z=1} - Y_{Z=0}}{D_{Z=1} - D_{Z=0}}$$

When there are observed confounders that are related to the instrument, we must condition on those confounders in estimating both reduced form and the first stage. If judge harshness were associated with either imprisonment or the outcome (for reasons other than imprisonment), we would need to condition on whatever variables were generating those associations.

We can also represent the natural experiment in equations, as is typical in econometrics (e.g. Angrist and Pischke 2008).

$$\begin{aligned} (1) \quad Y &= \beta_0 + \beta_1 D + \beta_2 X_2 + \cdots + \beta_j X_j + e \\ (2) \quad Y &= \gamma_0 + \gamma_1 Z + \gamma_2 X_2 + \cdots + \gamma_j X_j + u \\ (3) \quad D &= \delta_0 + \delta_1 Z + \gamma_2 X_2 + \cdots + \delta_j X_j + v \end{aligned}$$

Equation 1 provides the effect of treatment if we could assume no unobserved confounders. Equation 2 is the reduced form, and Equation 3 is the first stage. One method for estimating the treatment effect, β_1 , using the natural experiment, is to first estimate Equation 3 via Ordinary Least Squares (OLS), and then generate a new version of the treatment variable, D , that is the predicted value from this equation, $\tilde{D}$. Note that $\tilde{D}$ can be understood as the “part” of D that is related to Z (and X). The rest of the variation in D , represented by the error term, v , has been purged from $\tilde{D}$. If we substitute $\tilde{D}$ into Equation 1 for D and estimate the equation via OLS, we will estimate the treatment effect of interest, β_1 .¹ This procedure, known as 2-Stage Least Squares (2SLS), provides some additional intuition about how natural experiments work. We can understand the predicted value from

¹ This simple procedure will produce incorrect standard errors, because D is a regressor that has been generated from the data. Stat packages for IV analysis will produce the correct standard errors. When one’s outcome is binary and one does not want to use OLS due to concerns about violations of assumptions, it is conventional to estimate the instrumented treatment equation (or “second stage”) using a probit model because its assumptions about the normality of the residuals.

the first stage as containing only the variation in the treatment that is related to the instrument, thus removing variation due to unobserved confounders. Note that we must include the same X variables in both equations. $\tilde{D}$ will contain variation due to the X s, so they must be conditioned when we estimate the treatment effect.

2SLS (and other estimators) allows for the consideration of multiple treatments and multiple instruments by changing D , Z , and their corresponding coefficients to matrices. One must have at least as many instruments as treatments. When there are more instruments than treatments, the system of equations is said to be “over-identified.” When there are fewer instruments than treatments, the system is said to be “under-identified,” and estimation of treatment effects is not possible.

These tools can also be used to recover unbiased causal effect estimates from randomized experiments with non-compliance, when treatment and/or control cases do not comply with their randomly assigned treatment, typically biasing estimated effects toward zero. For example, families who were randomly assigned to the control group in a private school voucher lottery experiment may find other ways to pay for a private school for their child. In this context, the original randomization is the instrument, and the treatment is the treatment actually received. We can understand a 2SLS analysis of an experiment with non-compliance as scaling the effect of the instrument on the outcome by the non-compliance rate.

Interpretation in the Presence of Effect Heterogeneity

When treatment effects are not the same for all units (i.e., when there is “effect heterogeneity”), then an instrumental variables analysis estimates the “Local Average Treatment Effect”, or LATE (Angrist, Imbens, and Rubin 1996) or “Complier Average Causal Effect.” The LATE is the effect of the treatment among those whose treatment status was changed by the instrument, a group termed the “compliers.” In Figure 1, this is the group for whom $\delta_1 \neq 0$ (now δ_1 may vary across units). In the imprisonment example, this means that the random assignment of judges estimates the effect of imprisonment only for defendants for whom being sentenced by a harsh rather than lenient judge meant the difference between prison and probation (the compliers). Some defendants would be sentenced to probation by all judges, presumably due to the less severe nature of their crimes or their shorter criminal history. Such defendants are termed “never takers” because they never take the treatment. Other defendants would be sentenced to prison by all judges, the “always takers.”

It is important to note that in general it is not possible to identify in the data which units are compliers, always takers, and never takers. This is because the definitions of these groups involve unobserved counterfactual states, the treatment they would have received when exposed to values of the instrument to which they were not actually exposed. For example, to identify whether an individual defendant is a complier, always taker, or never taker, we would need to know how that defendant would be sentenced by both a lenient judge and a harsh judge. However, when both the treatment and instrument are binary, it is possible to recover the average characteristics of compliers, always takers, and never takers on observed covariates (Angrist and Pischke 2008). Finally, the LATE interpretation of the IV estimator also requires the monotonicity assumption discussed below.

For some applications of IV, LATE is an important quantity of interest. In policy analyses, the compliers are, in fact, the group that future policy changes would impact. In the

imprisonment example, the compliers are on the policy margin between prison and probation, and therefore, the group for whom a change in sentencing rules would impact. Note also that in many cases, the reduced form estimate can also be an important quantity of interest when policy change is most likely to change the instrument rather than the treatment directly. For example, the reduced form provides the effect of a policy effort to encourage judges to shift from being harsh to lenient.

Assumptions

While a natural experiment design has the benefit of avoiding the “no unobserved confounding” assumption, it requires the researcher to make other assumptions. We describe the three central assumptions of an IV analysis here but note that there are other assumptions that apply across many causal inference research designs that should also be satisfied, such as the Stable Unit Treatment Value Assumption (see Morgan and Winship 2015).

The Exclusion Restriction

The exclusion restriction states that the instrument must be associated with the outcome only through its effect on the treatment. In Figure 1, this means there are no other paths between the instrument, Z , and the outcome, Y , other than $Z \rightarrow D \rightarrow Y$. However, if there are such paths that go through observed covariates, the X s, then those can be blocked by conditioning on those X s. When Z and Y are associated through pathways other than through the treatment, the estimated treatment effect will be biased. Recall from above that we can think of the IV estimator as the ratio of the effect of the instrument on the outcome divided by the effect of the instrument on the treatment. If the association between the instrument and the outcome includes not only the effect of the instrument on the outcome through the treatment but also the association due to those other pathways, the numerator of the IV estimator will be biased.

Alternatively, one can understand the exclusion restriction in terms of equations. The assumption that the instrument affects the outcome only through the treatment implies that the instrument can be “excluded” from Equation 1 (the origin of the term exclusion restriction). This means that Z and the error term in Equation 1 are not associated, i.e., $Cov(Z, e) = 0$. A nonzero covariance would imply that the treatment effect estimate, β_1 , is biased because D and Z are correlated, as Z is subsumed in the error term, e .

Consider how the exclusion restriction can be violated. One way is causal pathways through which the instrument affects the outcome other than through the treatment, assuming these cannot be conditioned. In the imprisonment example, we might worry that a harsh judge affects recidivism through pathways other than the sentence. Perhaps a harsh judge behaves differently when interacting with defendants, leading to a perception that the justice system is unfair and illegitimate, weakening a defendant’s respect for the law and increasing the chances of recidivism.

A second general pathway is an unobserved common cause of both the instrument and the outcome. This concern is why the case for the validity of the exclusion restriction often rests on the argument that the assignment of the instrument is “as good as random” (or, more specifically, random with respect to potential outcomes). In the imprisonment example, prosecutors and defense attorneys may be aware of the harshness of various judges, and they might actively seek to manipulate judge assignment to cases, perhaps by making sure that defendants who are at particularly high risk of recidivism are assigned to harsher judges.

Therefore, the researcher needs to argue that judges are indeed randomly assigned and to understand in detail the institutional procedures through which this happens.

A third general pathway is generated by the process of sample selection. This can occur when the set of units that are eligible for the treatment has an inherent association between the instrument and the outcome. In the imprisonment example, only those who are convicted (or plead guilty) are eligible for sentencing. When the sentencing judge is also the trial judge, defendants may make their decision about a plea bargain based on the harshness of the judge. Defendants facing a lenient judge may be more willing to risk a trial and a sentence imposed directly by the judge. Those who do go to trial and are acquitted will not be in the sample of defendants who are convicted. If such defendants are systematically different from others in a way that impacts outcomes (e.g., risk aversion), an association between judge harshness and the outcome will be present in the sample by virtue of how it was constructed.

Unfortunately, it is not possible to verify the exclusion restriction empirically because it concerns the associations between observed and unobserved variables. One might think that one could estimate whether there is any association between the instrument and the outcome once the treatment is conditioned, but this is not a valid test. This is because the treatment is itself likely to be a common effect of both the instrument and unobserved covariates, and therefore conditioning on it leads to what is called “collider stratification bias” (Elwert and Winship 2014). Indeed, the very motivation for a natural experiment is the presence of an unobserved confounder of treatment.

One approach to potential violations of the exclusion restriction is to conduct a sensitivity analysis that essentially simulates new estimates of the treatment effect under different assumptions about the magnitude of the violation (Small 2007). A sensitivity analysis is most useful when specific additional pathways based on substantive knowledge are proposed, so the magnitude of the resulting association can be calibrated.

Relevance and the “Weak Instrument” Problem

The Relevance assumption states that the instrument does indeed affect the treatment, that is $\delta_1 \neq 0$ for some subset of the sample. Although this assumption can be empirically tested, the challenge comes when the effect of the instrument is small but nonzero. When the effect of the instrument on the treatment is too small, it is said to be a “weak instrument.” A weak instrument will produce an estimate of the treatment effect that is biased toward the estimate one would have obtained from OLS regression conditioning on observed covariates (Bound et al., 1995; Staiger and Stock, 1997). To get some intuition for the cause of this problem, recall that the IV estimator can be thought of as the ratio of the effect of the instrument on the outcome divided by the effect of the instrument on the treatment (the “strength” of the instrument). When that latter quantity is very small, then small violations of the exclusion restriction, including those due to sampling error, will be magnified by dividing by a small effect of the instrument on the treatment.

For this reason, measures of instrument strength are an important part of any IV analysis and the evaluation of a natural experiment design. One convention is to calculate the F statistic for the instrument(s) in the first-stage equation, with F-statistics over 10 considered to be evidence of a sufficiently strong instrument. While it can be tempting to add additional instruments, adding additional weak instruments will make the problem worse. Many other metrics for instrument strength are available in the literature, and this is an area of ongoing

research, particularly the analysis of natural experiments with many weak instruments, including alternative estimators like Limited Information Maximum Likelihood.

The strength or weakness of one's instrument(s) also bears directly on the standard errors on the estimated treatment effect. Recall that we can think of the IV estimator as isolating the part of the treatment that is affected by the instrument and discarding the rest of the variation in treatment. If one's instrument is weak, then the analysis will "discard" much of the variation in treatment. In other words, $\tilde{D}$ will have low variance. In a regression, the standard error of a coefficient is inversely proportional to the unique variance of the predictor. $\tilde{D}$ with lower variance will thus have a larger standard error.

Monotonicity

The monotonicity assumption is required under effect heterogeneity for the LATE interpretation. This assumption means that the instrument always affects treatment assignment in the same direction or not at all. That is, there are no units for whom the effect of the instrument on the treatment operates in the opposite direction as the compliers. Such units, were they to exist, are termed "defiers," and the monotonicity assumption is also sometimes referred to as "no defiers." In the imprisonment example, we must assume that all defendants are more likely to be sentenced to prison by the harsh judges. In other words, there are no defendants who would be sentenced to probation by otherwise harsh judges and to prison by otherwise lenient judges. This assumption could be violated if "harshness" were multidimensional. Perhaps some judges are particularly harsh on those convicted of drug crimes and particularly lenient on those convicted of property crimes. If the monotonicity assumption is violated, the resulting treatment effect estimate will be a weighted average of countervailing treatment effects for compliers and defiers, rendering the estimate uninterpretable. One solution to this problem is to directly model this variation in the meaning of harshness, for example by interacting judge identifiers with defendant characteristics and using those interactions as additional instruments (Mueller Smith, 2015). This will only work, however, if all relevant interactions are observed, and the interactions are not weak instruments.

Other Natural Experiment Research Designs

The above framework is general for designing and analyzing natural experiments, but researchers also use more specific research designs with their own assumptions, strengths, and weaknesses.

Regression Discontinuity

A Regression Discontinuity (RD) design emerges from institutional processes in which assignment to treatment is based on crossing a threshold (or "cut-off") on a continuous variable, termed the assignment variable, running variable, or forcing variable (Lee and Lemieux 2010). The intuition of this design is that we can compare cases just on either side of the cut-off, which are arguably very similar, after conditioning on the assignment variable. The "bandwidth" around the cut-off is an important decision that should be subject to robustness analysis.

In a “sharp” RD design, cases below the cut-off do not receive the treatment, and cases above do (or vice versa). An example would be a placement test for an academic enrichment program. In a “fuzzy” RD design, crossing the cut-off merely changes the probability of receiving treatment, but the cut-off does not perfectly determine treatment assignment (Figure 2). An example of a fuzzy design would be the legal drinking age in a study of the effects of alcohol consumption on a health outcome. While some people below the legal drinking age still manage to drink alcohol, and some people above the legal drinking age abstain from alcohol, crossing the legal drinking age increases drinking. A sharp design can be analyzed with a single regression model while a fuzzy design must be analyzed using an IV estimator. Crossing the cut-off can be thought of as an instrument for receiving the treatment because it changes the probability of treatment but does not determine it perfectly. A fuzzy RD design is doubly local in that it focuses on units near the cut-off on the assignment variable and estimates the LATE, the effect among those at the cut-off whose treatment assignment was in fact changed by the cut-off.

The exclusion restriction in an RD design is the assumption that there are no other changes that occur at the cut-off that affect outcomes. There is a particular concern that institutional actors who know the rules that create the “discontinuity” in the treatment at the cut-off will manipulate which units receive the treatment. For example, teachers who know the cut-off may grade a placement exam more generously for students they believe should be in the enrichment program. To assess the validity of this assumption, researchers typically make sure that other covariates do not show a discontinuity at the cut-off and also look for “heaping” or “bunching” of cases on one side of the cut-off (McCrary 2008).

Difference-in-Differences and Synthetic Controls

A difference-in-differences (DiD) design is useful when the researcher has individual or group-level outcome data from both before and after an intervention on both a group that was exposed to the intervention and a comparison group that was not. Rather than assume any change pre-to-post intervention in the treatment group can be attributed to the treatment, this design uses the trend in the comparison group to construct a counterfactual trend for the treatment group (Angrist and Pischke 2008). A common application is a change in policy that affects one group but not another. In order to attribute the effect to the policy change, one must assume that no other relevant changes also occurred around the same time as the intervention.

The treatment effect can be calculated as the difference between the treatment group trend and the comparison group trend (Figure 3). The key assumption is that the change over time in comparison group outcomes is a valid proxy for the change over time in the outcomes that would have occurred in the treatment group had the intervention not occurred. This “parallel trends” assumption can be assessed with more pre-intervention time periods. We are looking for trends in the outcome over many time periods before the intervention to be the same for the comparison group and the treatment group. Although there is no explicit instrument in DiD, implicitly one must assume that there is some instrument operating in the world that causes one group to receive treatment and the other not and is otherwise unrelated to potential outcomes.

In some applications, a single unit is the treatment group. Still, no other single unit is a reasonable comparison group, for example when a state makes a policy change, and no other state satisfies the parallel trends assumption. A “synthetic control” design attempts to solve

this problem by selecting multiple possible comparison units from a pool and weights them so that their collective (weighted) pre-intervention trend is parallel to that of the treatment group (Abadie et al., 2010).

Conclusion

Natural experiments are an important research design for causal inference questions in which we cannot feasibly or ethically conduct true randomized experiments and for which an assumption of no unobserved confounders is indefensible. However, natural experiments are not without their own assumptions and weaknesses. A good instrument – one that strongly affects the treatment, is only associated with the outcome through its effect on treatment, and always affects the treatment in one direction – is hard to find and depends critically on detailed knowledge of how institutions work to assign treatment to some units but not others. Because an IV analysis uses only variation in the treatment that is due to the instrument(s), it can require more data than other designs. This is especially the case for RD, which also focuses narrowly on cases near the cut-off. Unless one can assume homogenous treatment effects, IV does not estimate an average treatment effect, but instead estimates the LATE, the effect for those units whose treatment assignment was impacted by the instrument. This is an instance of a larger trade-off between causal identification and generalizability that recurs throughout causal inference. Finally, like experiments, natural experiments can leave the processes or mechanisms through which effects occur unexamined unless one accompanies them with qualitative research on processes or has instruments for possible mediating variables.

Further readings

We refer the reader to Angrist and Pischke (2008) for an econometric presentation of natural experiment research designs and accompanying statistical analyses. Gerber and Green (2012) provide a comprehensive discussion of social science experiments and natural experiments. Cunningham (2021) provides an accessible introduction to causal inference with many examples, including code in Stata, R, and Python. Finally, we refer the reader to Morgan and Winship (2015) for an intermediate level introduction to causal inference for social scientists.

References

- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493-505.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434), 444-455.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly Harmless Econometrics*. Princeton University Press.
- Bound, J., Jaeger, D. A., & 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(430), 443-450.

- Cunningham, S. (2021). *Causal inference: The Mixtape*. Yale University Press.
- Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research* (pp. 245-273). Springer, Dordrecht.
- Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40, 31-53.
- Fitzpatrick SL, Mayhew M, Catlin CL, Firemark A, Groß I, Nyongesa DB, O'Keeffe-Rosetti M, Rawlings AM, Smith DH, Smith N, Stevens VJ, Vollmer WM, & Fortmann SP. (2021). Evaluating the Implementation of Digital and In-Person Diabetes Prevention Program in a Large, Integrated Health System: Natural Experiment Study Design." *The Permante Journal*, 26(1), 21-31.
- Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.
- Hale SA, John P, Margetts H, & Yasseri T. 2018. How digital design shapes political participation: A natural experiment with social information. *PLOS ONE* 13(4): e0196068.
- Harding, David J., Jeffrey D. Morenoff, Anh P. Nguyen, and Shawn D. Bushway. "Short-and long-term effects of imprisonment on future felony convictions and prison admissions." *Proceedings of the National Academy of Sciences* 114, no. 42 (2017): 11103-11108.
- Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, 96(3), 863-876.
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), 281-355.
- Loeffler, C. E. (2013). Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment. *Criminology*, 51(1), 137-166.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698-714.
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference*, 2nd edition. Cambridge University Press.
- Mueller-Smith, M. (2015). The criminal and labor market impacts of incarceration. Unpublished Working Paper. Department of Economics, University of Michigan.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479), 1049-1058.

Staiger, D., & Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3), 557-586.

Ternovski, J & Yasseri, T. (2020). Social complex contagion in music listenership: A natural experiment with 1.3 million participants. *Social Networks*, 61: 144-152.

Figures

Figure 1: Directed Acyclic Graph for a Natural Experiment Research Design

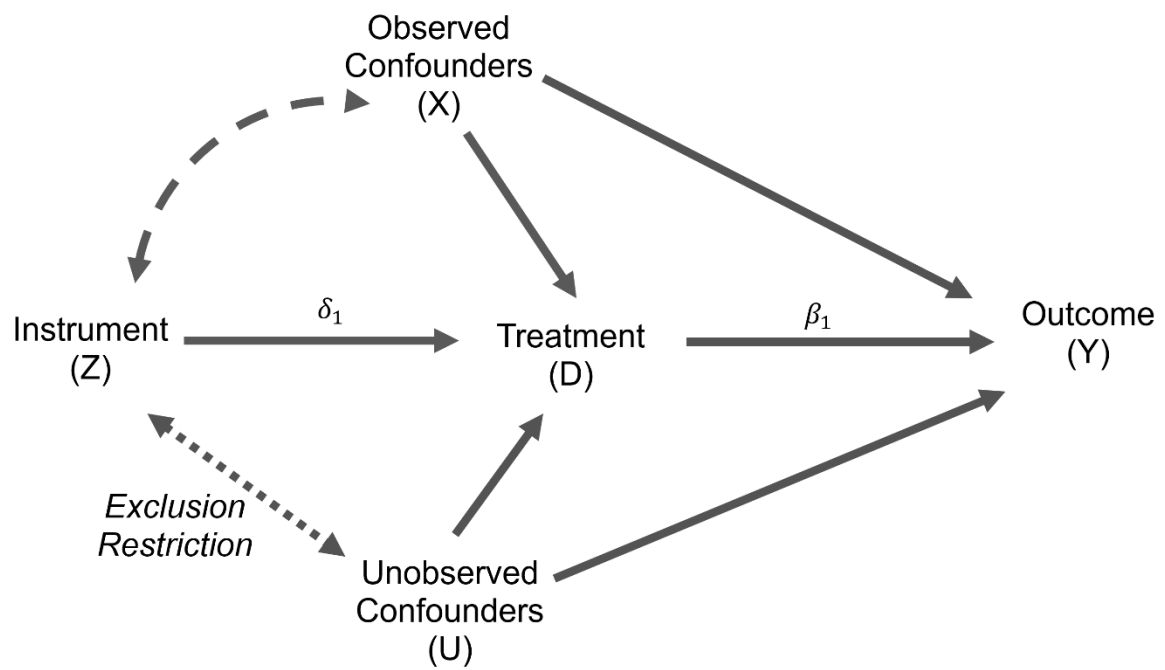


Figure 2: Regression Discontinuity Designs, Sharp (Top Panel) and Fuzzy (Bottom Panel)

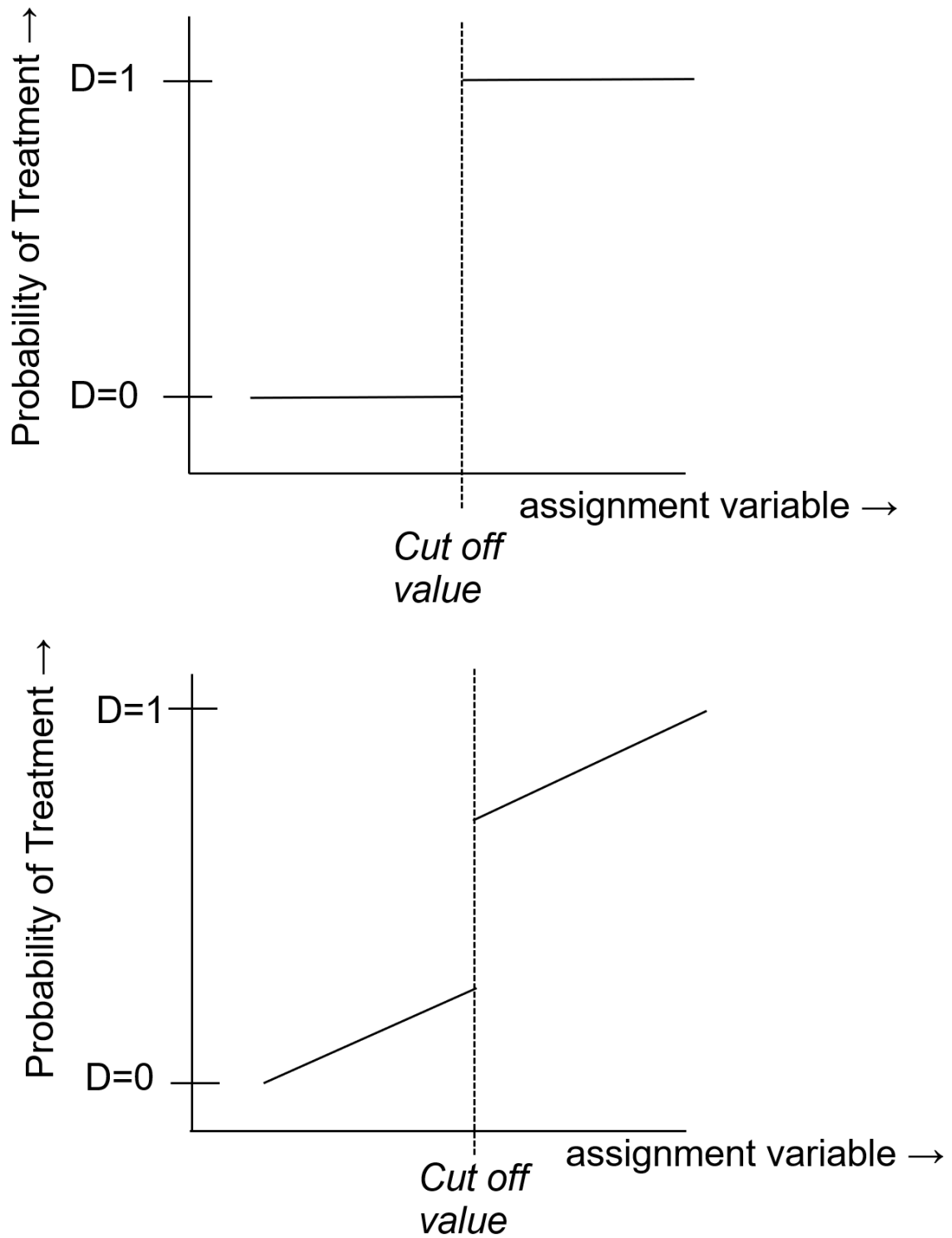


Figure 3: Difference-in-Differences Research Design

