How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data*

Luke Keele† William Minozzi‡

January 19, 2012

Abstract

Political scientists are often interested in estimating causal effects. Identification of causal estimates with observational data invariably requires strong untestable assumptions. Here, we outline a number of the assumptions used in the extant empirical literature. We argue that these assumptions require careful evaluation within the context of specific applications. To that end, we present an empirical case study on the effect of Election Day registration on turnout. We show how different identification assumptions lead to different answers, and that many of the standard assumptions used are implausible. Specifically, we show that EDR likely had little effect on turnout in any state including Minnesota and Wisconsin. We conclude with an argument for stronger research designs.

Word Count: 8385

*We thank Michael Hanmer for generously sharing code and data. For comments and suggestions, we thank Curt Signorino, Shigeo Hirano, Robert Erikson, Mike Ting, Walter Mebane, Michael Hanmer, Betsy Sinclair, Jonathan Nagler, Don Green, and seminar participants at Columbia University, the University of Michigan, and the University of Rochester. A previous version of this paper was presented at the 2010 Annual Meeting of the Society of Political Methodology, Iowa City, IA. and APSA 2010

†Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 Phone: 814-863-1592, Email: ljk20@psu.edu

‡Assistant Professor, Department of Political Science, 2189 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-247-7017, Email: minozzi.1@osu.edu

In recent years, there has been a renewed interest in methods for estimating causal effects using observational data. This interest has led to a greater focus on the assumptions needed for various statistical estimators to produce estimates that can be interpreted as causal. Often, however, assumptions are mistaken for estimators. For example, some assume matching can yield estimates of causal effects, when matching estimators rely on the same specification assumption as regression models.[1]

In this essay, we focus on the role of assumptions in the estimation of causal effects. We start with an outline of the key assumptions behind a number of popular approaches to the statistical estimation of causal effects using observational data. We begin with a discussion of approaches that depend on the specification of a statistical model. Here, we outline the key assumption needed to make causal inferences based on estimates from regression models, matching estimators, and the differences-in-differences estimator. We also describe some basic methods for probing the specification assumptions needed for these approaches. Next, we highlight the partial identification approach, where one uses weak, but credible, assumptions but can only bound the causal effect estimate. We conclude with coverage of two types of natural experiments: instrumental variables and regression discontinuity designs.

We then present an empirical case study of the effect of Election Day Registration (EDR) in Minnesota and Wisconsin. We argue that the quality of any assumption is hard to assess outside of the context of a specific empirical application. This application allows us to closely examine the plausibility of the assumptions needed for each approach. Next, we apply the various approaches and demonstrate how different assumptions lead to different conclusions. We present estimates from a partial identification analysis, logistic regression, differences-in-differences, and a regression discontinuity design. We also use a set of basic techniques to demonstrate that some estimates may be an artifact of the strong assumptions needed for identification. Moreover, we present estimates from a regression discontinuity design in Wisconsin that indicate that EDR did not increase turnout. Our results challenge a long

[1] See Barabas (2004) for one example of this confusion.

held presumption that widespread adoption of EDR would boost participation in elections.

1 Assumptions and Identification

We start with two preliminary tasks. First, we outline notation with the potential outcomes framework (see, e.g. Rubin 1974). The potential outcomes framework, often referred to as the Rubin Causal Model (RCM) (Holland 1986), has come to be an important tool for understanding the assumptions needed for the estimating of causal effects in both experimental and observational settings. In the potential outcomes model, each individual has two *potential outcomes* but only one *actual outcome*. Potential outcomes represent individual behavior in the presence and the absence of a treatment, and the actual outcome depends on the treatment actually occurs. We denote a binary treatment status with $D \in \{0, 1\}$. We use the the values of 0 and 1 refer to the values that D might take. While D can take on many values, we focus on the binary case for clarity.[2] The potential outcomes are $Y_D \in \{0, 1\}$, and the actual outcome is $Y \in \{0, 1\}$, where 1 is interpreted as taking an action like voting and 0 is not taking that action.

The potential outcomes model formalizes the idea that the individual-level causal effect of a law is unobservable, which is sometimes called the *fundamental problem of causal inference* (Holland 1986). We focus instead on population-level estimands, such as the *average causal effect* $\delta = E[Y_1] - E[Y_0]$. Limits to credible inferences about such causal estimands come in at least two varieties (Manski 2007). First, there are statistical limits. For example, sampling variability limits the conclusions that one can draw based on a small sample of observations. While the statistical problem is obviously important, we concentrate on a second limit to credible causal inference: the identification problem. This brings us to our second task: outlining the role of identification.

Formally, we observe $E[Y_1|1]$ and $E[Y_0|0]$ instead of $E[Y_1]$ and $E[Y_0]$. Therefore, an *identification problem* exists because there are terms in the causal estimand that are not

[2] We also omit a subscript i that would indicate that these are individual-level variables.

observable. Even if we had unlimited random samples that perfectly represent the population of interest, we still could not estimate the average causal effect without observing both potential outcomes. Resolution of the problem requires an *identification strategy*: a set of assumptions that warrant inferences based on observable quantities. Any research design, at least implicitly, adopts an identification strategy. Identification assumptions thus bridge theoretical and observable quantities. When identification assumptions hold, our estimate of the causal estimated is said to be is identified, which implies that the confidence interval for the estimated parameter shrinks to a single point as the sample size increases to infinity. We now review the various assumptions one might use for identification. All have been used in various parts of the political science literature. In each case, we focus on the assumption most critical for identification. Usually, this is an untestable assumption.

1.1 Standard Operating Procedure: Cross-Sectional Specification Assumptions

The most common approach to identification is to assume that we observe all relevant covariates. That is we must assume that our statistical model is "correctly" specified. When the model is correctly specified, τ will be identified. Critically, this assumption is non-refutable, insofar as it cannot be verified with observed data (Manski 2007). Under this approach, analysts collect all known confounders and use a statistical estimator to make treated and control comparable while the treatment effect is estimated. One statement of this assumption is that of "selection on observables" (Barnow, Cain and Goldberger 1980),[3] to emphasize that this assumption requires *all* covariates that predict both the outcome *and* the treatment. This is not the only requisite assumption for this approach, but it is the critical and untestable assumption needed for identification. Under this strategy, analysts often use a regression model or a matching estimator. While matching estimators rely on a weaker functional form than regression models, matching cannot correct for an "incorrect"

[3]Other names for this assumption include "conditional ignorability" and "ignorable treatment assignment."

specification.

We wait to fully discuss this assumption until the empirical application. The plausibility of any assumption typically depends on the specific application. That said, many argue that specification assumptions either do not hold or there is no way to know if they hold in many if not most applications with observational data (Green and Gerber 2002). If one is unwilling to proceed based on a standard specification assumption, what alternatives exist? We now explore a number of alternatives. In each case, we cannot avoid assumptions, it is just that different assumptions are made.

1.2 Temporal Specification: Differences-in-Differences

The next approach, differences-in-differences (DID), exploits longitudinal variation to alter the specification assumption in a fundamental way. DID might seem to unconditionally dominate cross-sectional analysis of treatment effects unless the key assumption is well understood. Given the longitudinal component, we need some additional notation. Let $t \in \{0, 1\}$ indicate time, where $t = 0$ before the treatment is administered and $t = 1$ after. The DID treatment estimand is $\delta = \{E[Y|D = 1, t = 1] - E[Y|D = 0, t = 1]\} - \{E[Y|D = 1, t = 0] - E[Y|D = 0, t = 0]\}$. This is just the difference between the treated-control differences from before and after treatment went into effect. Rather than modeling treatment assignment, DID uses the temporal component to eliminate differences across units that confound inferences.[4]

What must we assume for τ from a DID estimator to be identified? Identification, here, requires the expected potential outcomes for treated and control units to follow parallel paths in the absence of treatment. In formal terms, $E[Y_0(t = 1) - Y_0(t = 0)|D = 1] = E[Y_0(t = 1) - Y_0(t = 0)|D = 0]$. In words, the differences between the treatment and control groups

[4]The most straightforward estimation method for the DID treatment effect is to use conditional sample moments, but regression also suffices. The actual outcome is now $Y_i(t) = \tau D_i \times t + \delta_t + \eta_i + \nu_{it}$, where δ_t is time-specific, η_i is individual-specific, and ν_{it} represents unobservable characteristics. Let μ_{dt} be the conditional sample moment for group d in time t. The DID estimate is $\hat{\tau} = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})$. This quantity can also be estimated with least squares. Let i represent a particular citizen. Then one estimates the linear model $Y_{it} = \beta_0 + \beta_1 t + \beta_2 D_i + \beta_3 t \times D_i + \epsilon_{it}$. Abadie (2005) shows that $plim\ \beta_3 = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})$.

must be constant across time absent treatment. That is we must assume that no other events beside the treatment alters the temporal path of either the treated or control groups. This "parallel paths" assumption is also nonrefutable. If there are covariates that predict deviations from a parallel path, these can also be incorporated into a statistical model. Again one needs the "correct specification" but now one needs all the relevant covariates that predict the temporal paths of the treated and control groups. Thus, DID relies on a specification assumption, but a weaker one than articulated in 1.1.

1.3 Addressing Bias From Unobserved Confounders

The next approach acknowledges that causal inferences based on specification assumptions are often not credible due to hidden confounders. The response is to develop techniques that address this limitation. We outline three techniques that may clarify whether an association estimated under a specification assumption might be causal or instead reflects a hidden bias due to an unobserved confounder.

John Stuart Mill (1867) emphasized the need to ensure that treated and control units were identical in all respects save treatment. Ronald Fisher (1935, pg. 18) later dismissed this goal as "totally impossible" and advocated random assignment to generate comparable treated and control groups. In an observational study, however, it is often useful to restrict the analysis to a more homogeneous subset of the available data which can reduce sensitivity to biases from unobserved confounders. This does imply the use of a smaller sample which can lead to imprecise estimates, but uncertainty due to unobserved confounders is far greater in magnitude than sampling uncertainty (Rosenbaum 2005*a*). In observational data, increasing the sample size limits sampling variability but does nothing to reduce sensitivity to unobserved bias. As Rosenbaum (2010, pg. 102) notes increasing the sample size with heterogeneous units may increase precision around a biased point estimate, potentially excluding the true effect from the confidence interval! In many cases, a specification assumption may be invoked, but the analyst can opt to use a more homogenous subset of the data to

eliminate heterogeneity. As we demonstrate later, natural experiments reduce heterogeneity in a formal way.

Cook and Shadish (1994, 95) write, "Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations." Rosenbaum (2005*b*) develops this idea more formally into the concept of *pattern specificity* where one uses a pattern of confirmatory tests rather than rely on a single test. Thusly, comparing the treatment group to different control groups can illuminate the role of unobserved covariates. If a common pattern of effects emerges, the effects are more credibly due to the treatment. Second, causal theories do more than predict the presence of an effect; they also predict the absence of an effect in the absence of treatment (Rosenbaum 2002*b*). For example, if one compares turnout in two states before the implementation of an electoral reform and still detects a "causal" effect, any post-reform effects are suspect. For studies of turnout, these *placebo tests* are critical. Combining a specification assumption with a series of confirmatory tests according to a specific causal pattern may be enough to convince an audience that an estimated association deserves a causal interpretation.

Finally, one can evaluate the robustness of our inferences by conducting a sensitivity analysis. In a sensitivity analysis, we quantify the exact degree to which the identification assumption must be violated in order for our inference to be reversed. Although such analyses are not currently a routine part of statistical practice in political science, they are powerful tools for understanding the magnitude of possible hidden confounders. There are standards methods of sensitivity analysis for the specification assumption (Imbens 2003; Rosenbaum 2002*a*). We demonstrate many of these techniques in the empirical example.

1.4 Partial Identification and Bounds

A more radical approach is to abandon point identification. Most identification strategies produce *point* identification-a single parameter describes the causal effect. The partial identification approach instead focuses on producing a range of estimates that depend only on

weak assumptions. Under the partial identification approach, the analyst acknowledges that there is a fundamental tension between the credibility of assumptions and the strength of conclusions (Manski 1995).

Manski (1990) argues for first using the weakest possible set of assumptions that are based on the evidence from the data alone. Using the weakest set of assumptions produces a set of bounds (called no-assumption bounds) for the estimate of the causal effect. This strategy isolates ranges of values for the unobservable counterfactuals and therefore produces ranges of average causal effects.[5] Instructively, these bounds are always uninformative which means that they always bracket zero. In short, without stronger assumptions one cannot rule out the possibility that there is no effect. The no-assumption bounds are *not* a confidence interval, but an *identification region*. The notion of an identification region is prior to the notion of a confidence interval, which represents statistical uncertainty.

To make the inference informative, one adds assumptions about the nature of treatment response or assignment. The assumptions must be based on substantive insights about the process under study. These additional assumptions narrow the bounds on the treatment effect. By adding the assumptions individually, it allows one to observe exactly which assumption provides an informative inference. Assumptions can also be combined for sharper inferences. Next, we outline two common assumptions often used to narrow no-assumptions bounds.

First, assume the treatment is not counterproductive, so that it has a monotone response (Manski 1997).[6] This is tantamount to assuming that we know the sign of the average causal effect. Given this assumption, we can fix the lower no-assumption bound to zero. This assumption is often referred to a monotone treatment response (MTR). Independent of any assumption about response, one can make an assumption about assignment to treatment. For example, one can assume monotone treatment selection (MTS), which means that average

[5]Recall that the average causal effect is $\delta = E[Y_1] - E[Y_0]$. Using the law of iterated expectations, δ can be decomposed into conditional average potential outcomes. Specifically, the law of iterated expectations implies that $\delta = E[Y_1] - E[Y_0] = \pi E[Y_1|1] + (1-\pi)E[Y_1|0] - \pi E[Y_0|1] - (1-\pi)E[Y_0|0]$.

[6]Monotone response requires $E[Y_1|X, D] \geq E[Y_0|X, D]$ for all X and D.

potential outcomes are higher for individuals under the treatment than for those who do not receive the treatment.[7] Under MTS, we assume that treated units are selected to maximize the outcome. MTS decreases the upper no-assumption bound down to the estimate based on a naive comparison of the treated and control. Finally, we could combine both assumptions to narrows the bounds to between 0 and the naive comparison estimate.

This strategy has three strengths. First, the role of the assumptions in the analysis is completely transparent. The use of statistical models often obscures which assumption is critical, but here the set of assumptions necessary for identification of the treatment effect is transparent. Second, the treatment effect estimate can easily be assessed according to the plausibility of the identifying assumption. Finally, we avoid any type of specification assumption. We need not assume that we have correctly specified either a model for the outcome or treatment. In short, we can proceed under very weak assumptions, though this will come at cost since we can never rule out that there is no effect.

1.5 Natural Experiments

The final approach we discuss is that of natural experiments. We define a natural experiment as a real-world situation that produces plausible as-if randomized assignment to a treatment. Some in economics credit natural experiments with having produced a "revolution" in the study of observational data (Angrist and Pischke 2010). We review two forms of natural experiments: instrumental variables and regression discontinuity designs. Both of these methods provide an estimate for a subset of the study population. This is an important point: the leverage of both methods is predicated on reducing heterogeneity in the study population by making a more focused comparison. As such, efforts to reduce heterogeneity as outlined in Section 1.3 are in essence an attempt to mimic natural experiments.

Instrumental Variables Another approach to identification is to find an instrument for treatment status. We do not provide a full account of the assumptions needed to identify

[7] Monotone selection requires $E[Y_D|1] \geq E[Y_D|0]$ for $D \in \{1, 0\}$.

estimates as causal in the instrumental variables (IV) context. Sovey and Green (2011) provide a recent review of these assumptions, and Angrist, Imbens and Rubin (1996) fully derive these assumptions. We outline an experimental design where IV is valid, and use this as a heuristic for understanding whether the IV approach is valid in the context. The IV estimator is identified in what is known as the randomized encouragement design. Under this experimental design, subjects are randomly encouraged to take a treatment. Some subjects refuse or fail to the take the treatment, but the object of inference is the effect of the treatment and not the randomly assigned encouragement. If the assumptions hold, the IV estimator provides an estimate of the treatment as opposed to the encouragement. The IV estimand is the average effect among those induced to take the treatment by a randomized encouragement known as the complier average causal effect or the local average treatment effect. An example is useful. In one classic application of the encouragement design, subjects are randomly encourage to exercise. Some subjects choose to exercise while others do not. Later, health outcomes are measured for all participants. The IV estimate will identify the effect of exercise as opposed to the effect of encouragement even though not all subjects exercise. IV estimates the average effect among those induced to take the treatment (exercise) by the randomized encouragement. In the context of natural experiments, one hopes to find some intervention that randomly encourages units to take the treatment. The Vietnam draft lottery is one instrument where IV assumptions are credible (Angrist 1990; Erikson and Stoker 2011). Often, however, the IV assumptions are no more plausible than specification assumptions. Again, careful evaluation of assumptions within the context of a specific application is critical.

Regression Discontinuity Designs Recently the regression discontinuity (RD) design has been revived particularly in economics but also in political science. The promise behind RD designs arises from the relatively weak assumption need to identify treatment effects. Below, we briefly outline the RD design. Interested readers should see Lee and Lemieux (2010) for a detailed introduction.

In a regression discontinuity design, assignment of a binary treatment, D, is a function of a known covariate, S, usually referred to as the *forcing variable* or the *score*. In the sharp RD design, treatment assignment is a deterministic function of the score, where all units with score less than a known cutoff in the score, c, are assigned to the control condition ($D = 0$) and all units above the cutoff are assigned to the treatment condition ($D = 1$).[8] Hahn, Todd and van der Klaauw (2001) demonstrate that for an RD design to be identified the potential outcomes must be a *continuous* function of the score. Under this continuity assumption, the potential outcomes can be arbitrarily correlated with the score, so that, for example, people with higher scores might have higher potential gains from treatment.

The continuity assumption is a formal statement of the idea that individuals very close to the cutoff but on opposites sides of it are comparable or good counterfactuals for each other. Thus, continuity of the conditional regression function is enough to identify the average treatment effect *at the cutoff*. That is, the RD design identifies a *local* average treatment effect for the subpopulation of individuals whose value of the score is at or near c. Estimation of this treatment effect proceeds by selecting a subset of units just above and below the discontinuity and calculating the difference across these two groups. As we discuss in the Appendix, there are a number of different methods for selecting units around the threshold as well for the estimation of the treatment effect.

Lee (2008) provides an important behavioral interpretation for the RD continuity assumption. He demonstrates that when agents are able to precisely manipulate their value of the score continuity of the conditional regression function is unlikely to hold. Here, the score is $S = W + e$, where W represents efforts by agents to sort above and below c and e is a stochastic component. When e is small and agents are able to precisely sort around the threshold, the RD design may not identify the parameter of interest. In this case, treatment is completely determined by self-selection and the potential outcomes will be correlated with

[8] In contrast, in a fuzzy RD design the assignment to treatment is a random variable given the score, but the probability of receiving treatment conditional on the score, $P(D = 1|S)$, still jumps discontinuously at c.

observed and unobserved characteristics of the agents. However, when e is larger agents will have difficulty self-selecting with any precision and whether an agent is above or below the threshold is essentially random. The behavioral interpretation of the continuity assumption allows for easier assessment of the identification assumption in the RD design. For example, the first use of the RD was to study the effect of a scholarship on educational outcomes (Thistlethwaite and Campbell 1960). Students that took the PSAT and received a score above a defined threshold received a scholarship, and those below that score did not. The RD design was used to compare the outcomes for students just above the threshold and those just below. Unless, we have reason to believe that students can very precisely manipulate their score on the exam, the treatment effect will be identified within a local area since we expect students just above the threshold to be good counterfactuals for the students just below the threshold. Another advantage of the RD is that the identifying assumption has a clearly testable implication. In the RD design, variation in the treatment is approximately random within some local neighborhood around the threshold, and when true all "baseline characteristics"–all those variables determined prior to the realization of the assignment variable–should have the same distribution just above and below the cutoff. If there is a discontinuity in these baseline covariates, then the underlying identifying assumption in an RD is unwarranted (Lee and Lemieux 2010).

When we have reasons to believe that e is small relative to W, we maintain that the RD design is the strongest of the approaches that we have reviewed. In the RD design, we have a clear testable implication of the identifying assumption. Second, the RD is based on a design. That is, we do not rely on found data, but instead we rely on the fact that a threshold had to be created and implemented. Recently, RD designs have gained further credibility by recovering experimental benchmarks (Cook, Shadish and Wong 2008)

There is a drawback to the RD identification strategy. The RD estimate is necessarily local: it only applies to some limited range of units above and below the threshold, but not to all units. One goal of our analyses might be to make policy recommendations, and a local

treatment effect by definition may not extrapolate to other units within a single state much less to other states. With observational data, however, we must often trade external validity for internal validity. The RD has the highest level of interval validity but is also necessarily local, which hurts external validity. In this sense, the RD embodies the call for more local estimates that we outlined in Section 1.3.

2 Case Study: Election Day Registration

We now turn to an empirical application to highlight the role of assumptions in the estimation of causal effects with observational data. More specifically, we examine the effect of election day registration on turnout. Election day registration significantly reduces the cost of voting by collapsing voting and registration into the same act. EDR is widely credited with increasing turnout (Brians and Grofman 1999, 2001; Hanmer 2007, 2009; Highton and Wolfinger 1998; Knack 2001; Mitchell and Wlezien 1995; Rhine 1995; Teixeira 1992; Timpone 1998; Wolfinger and Rosenstone 1980). Based on this empirical evidence, political scientists are often willing to propose that if all states adopted EDR, turnout would increase in publications such as the *New York Times* (Just 2011). Thus, we examine a policy intervention that many believe has increased turnout and whose effects are thought to be well understood. The analysis below is not meant to be comprehensive study of EDR, but it is meant to highlight how inferences differ depending on what assumptions are used for identification.

Below, we focus on EDR in Minnesota and Wisconsin for two reasons. First, EDR has long been credited with placing these states among those with the highest turnout (Wolfinger and Rosenstone 1980). Second, while some work has cast doubt on whether EDR increased turnout in states like New Hampshire, Wyoming, and Montana, it is still widely understood to have increased turnout in Minnesota and Wisconsin (Hanmer 2009). Therefore, we use Minnesota and Wisconsin as our treated states. This implies that turnout in the other 48 states must serve as our counterfactual. The difficulty is that other states may have different levels of turnout for many reasons other than the fact that they do not have EDR.

In each approach below, we will have to rely on assumptions to create a valid counterfactual comparison.

We start with a bounds analysis to understand what we can learn with minimal assumptions. Next, we use both the cross-sectional and temporal identification strategies. We then examine these estimates for bias due to unobserved confounders. We conclude with estimates from a natural experiment. Specifically, we use a regression discontinuity design within the state of Wisconsin, which produces surprising results.

2.1 Bounds

We start with a bounds analysis to understand what we can learn from the data without strong assumptions.[9] In the bounds analysis, we compare turnout in Minnesota and Wisconsin, the two treated states, to turnout in the rest of the country. The advantage with the bounds approach is that we rely on a set of very weak assumptions. The disadvantage is that bounds leave us with a great deal of uncertainty about whether EDR was actually effective. We calculate the bounds in 1980. Minnesota first used EDR in 1974 while Wisconsin first used it in 1976. We use data from 1980 to allow for any delay in the onset of the EDR effect. We start with the no assumption bounds in the first row of Table 1. Recall these are bounds on the identification region and do not reflect any statistical uncertainty.[10] These bounds reveal what we can learn from the data alone without *any* assumptions about identification. The no assumption bounds range from -63 to 37. That is without any assumptions, we can say the effect of EDR ranges from depressing turnout by 63% or increasing it by 37%. It is instructive to understand how little can be learned without assumptions.

To make the bounds more sharp, we must add assumptions. We next assume that monotonicity holds. This is tantamount to assuming that we know the sign of the treatment effect. In the context, here, the monotone treatment response (MTR) assumption implies that EDR does not depress turnout. Monotone treatment response is a fairly weak assumption in this

[9] For examples of more complete analyses based on bounds see Hanmer (2007, 2009).

[10] The bootstrap may be used to provide estimates of statistical uncertainty.

context, since it is difficult to imagine how EDR would cause a citizen to not vote. The MTR bounds are in the second row of Table 1. Under this assumption, EDR raised turnout by as much as 37% or as little as 0. We next calculate the bounds assuming monotone treatment selection, or that treated units are selected to maximize the outcome. That is, we assume that Minnesota and Wisconsin enacted EDR under the assumption that it would increase turnout. In the context of EDR, MTS is not wholly plausible. In Minnesota, EDR served as a compromise when policy-makers wanted to implement a statewide voter registration system (Smolka 1977). Thus, EDR was put in place as a legislative compromise and not necessarily because it was thought EDR would increase turnout. MTS, then, is not a given in this context. The third row of Table 1 contains the bounds calculated under monotone selection. Under this assumption, the EDR effect ranges from -63% to 12.7%. If we now combine MTS and MTR, the range of the EDR effect narrows to 0 and 12.7%. Thus under a fairly weak set of assumptions, we can infer that EDR increased turnout by as much as nearly thirteen points, but that effect may also have been zero or anywhere in between. The cost of weak assumptions is much greater uncertainty about the true treatment effect. We now estimate the treatment effect under specification assumptions. This approach will provide what appears to be greater certainty but at the cost of much stronger untestable assumptions.

Table 1: Bound Analysis of the Effect of EDR in Wisconsin and Minnesota, 1980

No Assumption Bounds	[-63, 37]
Monotone Treatment Response (MTR)	[0, 37]
Monotone Treatment Selection (MTS)	[-63, 12.7]
Monotone Treatment Selection (MTR + MTS)	[0, 12.7]

Note: Each set of bounds represents a lower and upper bound on the EDR treatment effect in percentage terms. Thus in 2000, under Monotone Treatment Selection, the increase in turnout due to EDR for all states was between 0 and 5.6 percentage points.

2.2 Specification Approaches

Many analysts use a cross-sectional specification approach to causal inferences about turnout. That is, they assume that they observe all the relevant covariates that predict both the treatment and the outcome. This assumption may be reasonable in some settings, but it is not realistic in the study of EDR. Recasting the specification assumption as a selection on observables assumption helps illuminate the fact that while these models have a reasonablly good specification for turnout, there are few if any covariates that predict why a state adopts EDR. Under the specification approach, analysts typically use a regression model to adjust for differences in state level distributions of education, income and other similar covariates. These are all covariates that are correlated with turnout, but these measures do not provide much leverage over why respondents in one state are treated to EDR and those in another state are not. In short, these models do not account for why some states select into the EDR treatment and other states do not. Importantly, regression models can do little to overcome a poorly specified selection mechanism. Moreover, this approach requires us to make comparisons *across* states. Such comparisons are problematic since cross-state differences in turnout are a function of a wide variety of processes. Things such as electoral competitiveness which can be magnified by the Electoral College, specific statewide elections, political culture, or differences in other state election procedures such as polling hours or absentee balloting can all contribute to differences in turnout at the state level. Unless we are confident that we have fully measured and controlled for all these varied factors, any attempt to isolate cross-state turnout differences due to EDR will be confounded.

Here, we estimated the effect of EDR in Minnesota and Wisconsin in 1980. We used a logistic regression model with specified with measures for income, education, age, age-squared, a dummy variable for whether the respondent is African American, marital status, sex, a dummy variable for Southern states, and time at address before the election. The results are in Table 2. Under a specification assumption, EDR increased turnout 10.1% points.[11]

[11] Our estimate is quite close to a similar analysis based on fewer states in (Hanmer 2009).

Unlike the bounds approach, now that we have adopted a strong untestable assumption, we get a precise estimate which indicates that EDR produced a large increase in turnout.

Table 2: Logistic Regression Estimate of the Effect of EDR in 1980

Treatment Effect Estimate	10.1
95% Confidence Interval	[8.7, 11.7]

Note: Estimate from logistic regression. Confidence intervals are calculated using the bootstrap. Cell entries are in percentages. Estimate is Minnesota and Wisconsin compared to all other states. Model includes a full specification with measures of education, income, race, sex, marital status, age, and a dummy variable for Southern states.

We next use an alternative specification approach: differences-in-differences (DID). When applied to EDR, DID has both advantages and disadvantages. First, the DID estimator is clearly superior to the specification approach used with cross-sectional data. The DID estimator will account for baseline differences in turnout across states, which are quite common. However, the DID estimator also has serious drawbacks particularly when applied to turnout. Recall that with the DID estimator we must assume that the differences between groups are constant across time absent treatment. That is we must assume that no other events beside the treatment alter the temporal path of either the treated or control groups. In the context of voter turnout, this implies that the nothing else in the treated states serves to boost turnout. In short, a competitive senatorial or gubernatorial race in the treated states could boost turnout and that would be attributed to EDR. As another example, if a state that has adopted EDR becomes a battleground state in the next election, it will be impossible to distinguish the effects of EDR from increased mobilization efforts that result from increased competition. Again, the standard turnout model specification is of little help. Measures such as education and income are nearly constant across elections and will provide little predictive power of why the overtime dynamics of turnout might change. Thus while the DID assumption may be plausible in many contexts, this assumption is less plausible in the study of turnout across states where it is not unusual for other factors to alter the dynamics of turnout across elections.

For purposes of comparison, we estimated the treatment effect using DID. Here, we used 1972 as the pretreatment year and 1980 as the post-treatment year. This allows for possible delay in the effect of EDR. We clustered standard errors by state though this may not be enough for correct standard error estimates (Donald and Lang 2007). We also use a slightly different empirical specification. We use measures of education, sex, a dummy variable for African Americans, income, age, and age-squared since not all the measures in 1980 were available in 1972. Table 3 contains the results from the DID estimator. Under the DID estimator, the EDR effect is now 4.1% instead of 10.1%.[12] First, the DID estimate is much smaller, which suggests that the specification approach with cross-sectional data is is badly misspecified. Much of the EDR effect in that model is due to the myriad of differences that cause turnout to differ across states in each election. The DID estimator appears to correct for this, but perhaps not sufficiently.

Table 3: Differences-in-Differences Estimate of the Effect of EDR, 1972-1980

Treatment Effect Estimate	4.1
95% Confidence Interval	[0.76, 7.3]

Note: Cell entries are in percentages. Model includes a full specification using education, income, race, sex, and age. Estimate is Minnesota and Wisconsin compared to the rest of the states. Standard errors adjusted for clustering at the state level. Model includes a full specification including education, income, race, sex, and age.

One advantage of the DID estimator is that one can use the data to perform an informal test of the identifying assumption. To do so, one plots the trend in the treated and control outcome before the treatment goes into effect (Angrist and Pischke 2009). If the two trends are largely parallel, then this is evidence that the assumption holds. Figure 1 contains a plot of the average turnout in Wisconsin and Minnesota compared to the average turnout in the rest of the U.S. Based on Figure 1 the DID assumption appears to be fairly credible as the pretreatment trends are fairly similar.

[12]Our estimators are very similar to those in (Hanmer 2009).

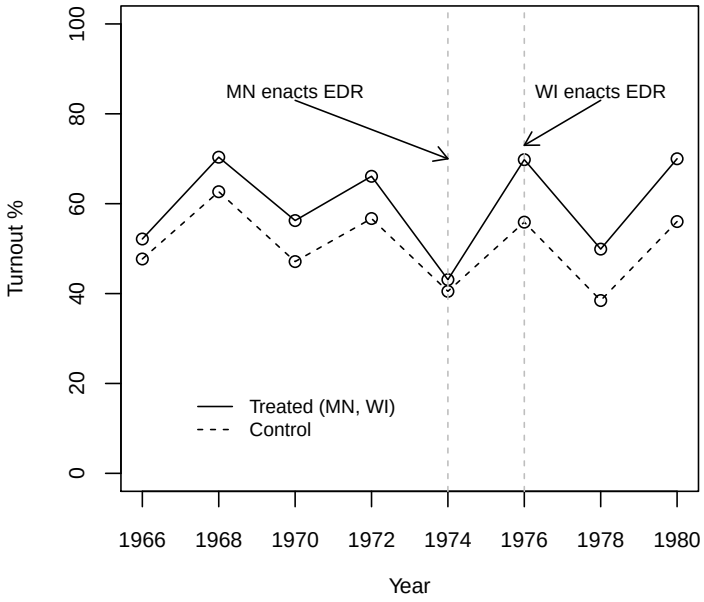

Figure 1: Trends in Presidential Election Turnout for Treated and Control States

2.3 Addressing Bias From Unobserved Confounders

So if we rely on a specification assumption in either form, we would conclude that EDR did increase turnout in Minnesota and Wisconsin. We now probe for bias from unobserved confounders. We do this in three ways. First, we attempt to reduce heterogeneity through a more focused comparison. Second, we use a sensitivity analysis for the specification assumption. Third, we examine the data for patterns specific to our the causal hypothesis.

Thus far, we have compared the two treated states, Minnesota and Wisconsin, to all the other states. In short, we are using a weighted average of other states as a counterfactual for these two states. An alternate strategy that we highlighted in Section 1.3 is to restrict the study population to areas of the U.S. that are more comparable to Minnesota and Wisconsin. Such a restriction should reduce heterogeneity and potentially reduce bias in the estimates. Hanmer (2009) uses this approach by only using Iowa and South Dakota as control states in a differences-in-differences analysis, but neither state has a large metro area comparable

to Milwaukee and Minneapolis-St. Paul. Comparing Minnesota and Wisconsin to Iowa and South Dakota is to mostly compare urban voters to largely rural voters. [13]

Here, we take advantage of the timing of EDR implementation. Minnesota first used EDR in the 1974 election, while Wisconsin did not have EDR in place until 1976. We exploit the timing in two ways. First, we conduct an analysis in 1974 where we compare Minnesota to Wisconsin. This allows for a comparison among the two states that actually implemented EDR first. Here, we did a simple matching analysis in 1974, where we matched on the same covariates used in earlier analyses. We rely on matching since it allows us to more easily implement a sensitivity analysis.[14]

Table 4: Difference in Turnout Rates Across Minnesota and Wisconsin, 1974

Unmatched Estimate[a]	9.6
95% Confidence Interval	[3.9, 15.2]
Matched Estimate[b]	9.7
95% Confidence Interval	[3.9, 15.2]
Γ	1.2
N	866

Note: Cell entries are in percentages. [a] Unmatched estimate is simply a difference in proportions test for all CPS respondents across Minnesota and Wisconsin 1974.
[b] Matched estimate is based on matching respondents across Minnesota and Wisconsin.

In Table 4 we present two different treatment effect estimates for EDR in Minnesota. The first estimate does not adjust for any covariates, while the second estimate uses matching to adjust for education, income, age, marital status, and time at address. We excluded a small number of racial minorities from the analysis. We draw attention to two different aspects of the estimates. First, the covariates do little in that the treatment effect estimate barely

[13]Moreover, a quirk in the CPS data further invalidates using Iowa and South Dakota as counterfactuals for Minnesota and Wisconsin. In 1974, the year Minnesota implemented EDR, all the respondents in the CPS sample from Minnesota and Wisconsin were drawn from Minneapolis-St. Paul and Milwaukee, and in 1972 and 1976, over 90 percent of the sample is from these metropolitan areas.

[14]We use genetic matching (Sekhon and Diamond 2005; Sekhon 2011) to form two essentially equivalent groups based on the observed CPS covariates. Balance was quite good with the smallest p-value from KS-tests for balance being above 0.40.

changes once we match. This demonstrates how much adjustment is provided by simply selecting to comparable states. Second, we perform a sensitivity analysis. Recall that in a sensitivity analysis, we quantify the exact degree to which the identification assumption must be violated in order for our inference to be reversed. Here, that implies that we quantify the degree of misspecification required to reverse our inference. We apply a sensitivity analysis for matching estimators developed by Rosenbaum (2002*b*). In the sensitivity analysis, we manipulate the Γ parameter which measures the degree of departure from random assignment of treatment. Two subjects with the same observed characteristics may differ in the odds of receiving the treatment by at most a factor of Γ. In a randomized experiment, randomization of the treatment ensures that $\Gamma = 1$, that is the odds of treatment are the same across treated and control. In an observational study, Γ may depart from one. For example, if Γ is two for two subjects, one treated and one control, that are identical on matched covariates, then one subject is twice as likely as the other to receive the treatment because they differ in terms of an unobserved covariate (Rosenbaum 2005*b*). While the true value of Γ is unknown, we can try several values of Γ and see how the conclusions of the study change. Specifically, we calculate an upper bound on the p-value for a range of Γ values. If the upper bound on the p-value exceeds the conventional 0.05 threshold, then we conclude that a hidden confounder of that magnitude would explain the observed association. If the study conclusions hold for higher Γ values, the estimate is fairly robust to the presence of a hidden confounder. The third row of Table 4 contains the value of Γ at which the p-value exceeds 0.05. Here, we see that if an unobserved covariate caused two identically matched voters to differ in their odds of treatment by as little 1.2, that would explain the estimated effect. Normally, we would compare this number to the effect sizes of other covariates that predict treatment. However, we have no covariates that help us explain why Minnesota adopted EDR before Wisconsin. As a rule of thumb, we would prefer that the Γ values exceed 1.5 and preferably 2. As such, the sensitivity analysis indicates an unobserved confounder could easily explain the association we observe.

Table 5: Turnout Rates in Minnesota and Wisconsin, 1966-1980

Year	Minnesota	Wisconsin	Difference[a]
1966	57.7	46.6	11.1
1968	73.6	67.1	6.5
1970	60.9	51.6	9.3
1972	69.2	63.0	6.2
1974	46.8	39.4	7.4
1976	72.3	67.3	5.0
1978	54.8	45.0	9.8
1980	71.4	68.6	2.8

Source: *A Statistical History of the American Electorate* (Rusk 2001) [a]The difference is calculated as the Minnesota turnout rate minus the Wisconsin turnout rate.

Next, we examine these results in terms of pattern specificity. That is we ask whether the estimated effects fit a broader pattern that is consistent with the causal hypothesis. First, we perform an informal placebo test. If Wisconsin is a good counterfactual for Minnesota, turnout rates in the two states should be similar before EDR went into effect in Minnesota in 1974. We are unable to perform a more formal placebo test since turnout items were not apart of the CPS in 1970, the first midterm election before EDR went into effect in 1974. Instead, we use actual turnout rates for the two states. Table 5 contains turnout rates for Minnesota and Wisconsin and the difference from 1966 to 1980. Importantly, this data reveals that turnout in Minnesota was higher than turnout in Wisconsin in *all* pretreatment years and that difference is at least six points. This suggests that the effect we estimated in Table 4 is simply an estimate of the fixed effect between the two states.

We can deduce other patterns as well from the data in Table 5. As we outlined before, in 1974 Wisconsin did not yet have EDR, while Minnesota did. If EDR is effective, the difference in 1974 should vanish or shrink when the treatment goes into effect in Wisconsin in 1976. If, instead, the estimated gap stays roughly constant, is likely not due to EDR. This sequencing of treatment relies on a post-treatment comparison, so we must assume that no other interventions alter the trajectory after treatment. It does, however, allows

us to look for a pattern of effects that is consistent with our causal hypothesis. While we observe some shrinking of the difference in 1976 as it drops from to a five percentage point gap the difference is nearly ten points in 1978. Moreover, using the data in Table 5, we can also perform an informal DID analysis. Here, we use turnout in 1970 as the pretreatment baseline and use 1974 as the posttreatment period for the DID estimates. This DID estimate is -1.9%, which of course in the wrong direction.[15]

In sum, the evidence for an EDR effect is not compelling once we probe for evidence of hidden confounders. The matching analysis reveals how weak the specification is once the states are comparable. The sensitivity analysis suggests a hidden confounder could easily explain the estimated difference. Next, we see that the general turnout pattern is not consistent with the causal hypothesis. Perhaps the problem is that our estimates are not local enough. The difference in turnout between Minnesota and Wisconsin due to unobserved confounders is too large to make the observed associations credible estimates of a causal effect. Thus, we move to a natural experiment within a single state.

2.4 Natural Experimental Approach: Regression Discontinuity

Next, we use a natural experiment to estimate the effect of EDR. Here, we would hope to more closely approximate the counterfactuals produced by a randomized experiment. First, we rule out IV as a plausible alternative. To evaluate IV in this context, we must ask how well EDR fits the paradigm of the randomized encouragement design. We argue that the fit is poor. In the case of EDR, we must find an instrument that randomly encourages states to adopt EDR but has no subsequent direct effect on turnout. We know of no instrument that does so. While the IV approach may be successful in some venues, we argue that it tends to be implausible for studying interventions like EDR.

Is it possible to conduct an RD analysis in the context of EDR? In the case of Wisconsin, we can. Before the adoption of EDR in Wisconsin, municipalities with populations of less than 5,000 were not required to use a voter registration system while those above 5,000

[15]We plotted the turnout rates and trend is quite similar in the pretreatment time periods.

had a standard system of registration (Smolka 1977). In 1976, once EDR was adopted, municipalities with populations of less than 5,000 were still not required to maintain a registration system but those with a population greater than 5,000 switched to an EDR registration system. The RD design allows for a nearly ideal design to estimate the causal effect of EDR on turnout for two reasons. First, we can implement the design within a single state which holds all state level confounders constant without specification assumptions. The key difficulty with the other approaches is that we are unable to provide a compelling specification for why turnout differs cross states. The best we have been able to do thus far is to rely on state fixed effects with the DID estimator. Given that turnout may differ across states due to a variety of institutions, history and competitiveness, the best approach may be to hold those factors constant with a within-state design where all state level confounders are held constant by the design. Second, the RD design is credible in this setting since we expect it will be difficult for municipalities to manipulate their population in a precise manner to avoid having to use a registration system. Population for municipalities is determined by the census and not the municipality, so it should be nearly random whether a municipality has a population either just above or below 5,000. Thus, we can estimate the EDR effect under the weaker RD assumption, and the RD estimate provides more specificity than using partial identification.

There is one key limitation associated with the RD design in this context: it identifies an effect but does not identify the effect of interest. Ideally, we would have an RD design where municipalities below 5,000 have voter registration and those above the threshold have EDR or vice versa. This design would isolate the precise EDR effect. What we actually observe is that municipalities below the threshold have no registration compared to municipalities with either a standard registration system or EDR. It is not unusual for a natural experiment to identify an effect, but the effect that is identified may be subtly different from the effect of interest. Sekhon and Titiunik (2011) provide an excellent case study of this phenomenon. Does this mean all is lost?

One strategy is to conduct two separate RD analyses, the first before the adoption of EDR and the second after. If EDR has an effect, the gap in the treatment effect between these two RD estimates should shrink. This is essentially a mixture of RD and DID.[16] Comparing the two RD estimates, however, requires adopting the DID assumptions which, as we have argued, may not be realistic. Moreover, it seems imprudent to adopt one identification strategy due to its weaker assumptions and then add another more implausible assumption. We argue that a better approach is to use the RD estimate from 1974 as an upper-bound on the effect of interest. That is, the gap between no registration and registration should form an upper bound on the EDR effect. If turnout is five points lower for municipalities with a population of 5,000 or more before EDR goes into effect, we cannot expect EDR to reduce this gap by more than five points. Moreover, if there is no difference in turnout before EDR goes into effect we cannot credibly argue that EDR increases turnout. Thus despite the fact that we cannot identify the effect of interest we can at least put an upper bound on the EDR effect. We also estimate the RD treatment effect for 1978 but only as a means of informal comparison.

We start our RD analysis with two plots. In Figure 2, for both 1974 and 1976, we plot the log of population again turnout. As is standard in an RD analysis, we bin population values and plot turnout means within these bins (Lee and Lemieux 2010). We also add the fit from a nonparametric regression model and 95% confidence intervals. In both plots, we observe the same general trend where municipalities with larger populations have lower turnout. The question, however, is whether turnout differs in a local neighborhood around the threshold of 5000. In both years, there is little evidence of a difference in turnout around this threshold.

One important strength of the RD design is that the identification assumption has a

[16]Burden and Neiheisel (2011*a*) exploits the same variation in Wisconsin voting laws but applies the DID estimator. They find a positive effect for EDR but only when they include Milwaukee. As such, their estimate is far less local than that in an RD design. We found if one ignores the RD and uses all the data, Milwaukee exerts a strong influence. This is to be expected, but Milwaukee has no reasonable counterfactual area within the state of Wisconsin. In general, we think a DID estimator, here, fails to exploit the much weaker assumptions in the RD design.

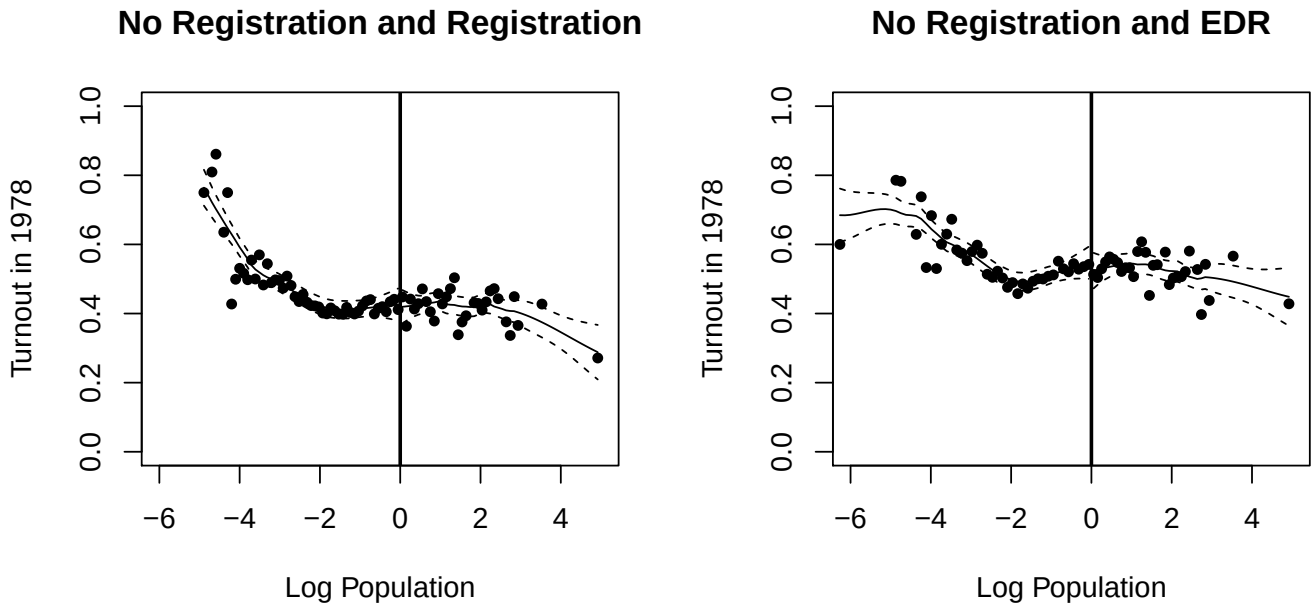

Figure 2: Regression Discontinuity Design: The Effect of Voter Registration System on Turnout.

testable implication. If we apply the RD analysis to other covariates that should be correlated with turnout, we should not find any effect. We compare eight important census covariates aggregated to the municipal level: percentage African American, percentage over the age of 65, percentage with a high school degree, percentage below poverty, per capita income, median household income, median house value, and median rent for 325 municipalities in Wisconsin that had Census data. We plotted each of these covariates against log population along with a nonparametric regression fit to both sides of the discontinuity. Importantly, there is no evidence of that any of these covariates are correlated with the score near the threshold (see the appendix).

Table 6 contains point estimates and 95% confidence intervals for both 1974 and 1978. See the appendix for details on estimation and bandwidth selection. We focus on the estimate in 1974. As we reasoned earlier, this estimate should serve as an upper bound on the EDR effect. That is, EDR should only make the effect of registration smaller. While the point estimate is in the expected direction, the point estimate is relatively small (-1.5 percentage

Table 6: Regression Discontinuity Estimates of the Effect of EDR in Wisconsin, 1974 and 1978

	1974	1978
Estimate	-1.5	-2.4
95% CI	[-5.8, 2.6]	[-6.6, 1.4]

Note: Effect estimates at the threshold of municipal population of 5000. Estimates from local regression fit to both sides of the threshold with a triangular kernel. Confidence intervals based on 5000 bootstrap resamples with BC_a confidence intervals. Bandwidth selection done via MSE-minimization (Imbens and Kalyanaraman 2010). For the 1974, $N = 800$ and for 1978 $N = 790$.

points) and the standard error is too large to rule out that the effect is not zero. Even if we ignore the fact that this estimate is not statistically significant, EDR could at most have an effect of less than two percentage points.[17] The estimate for 1978, while somewhat larger, remains statistically insignificant and negative. If we were to adopt the DID assumptions, the point estimate would be in the wrong direction. Importantly, under the most credible research design we find no evidence that EDR increased turnout. Our estimate for the effect of EDR has declined from ten points under the strongest assumption to less than two points under the weak assumptions of the quasi-experimental approach.

What might explain the large difference between the estimate from the RD design and the estimates from logistic regression and differences-in-differences? Is it simply that the estimands are different? We cannot provide a definitive answer, but we offer that the RD estimate differs since it creates a better counterfactual comparison in two ways. First, this is a within-state design where all the state-level confounders, of which there are many, are held constant. Second, even within the state of Wisconsin the inference is confined to municipalities that are actually comparable. We could use all the data, but does it make sense to compare Milwaukee to towns where the population is less than 500? The fact that EDR appears to be ineffective, here, makes it much less surprising that the National Voter Registration Act of 1993 more commonly known as the Motor Voter Act did little to increase

[17] Our estimate, here, is consistent with other work on the effect of registration, which also finds around a two percentage point effect (Burden and Neiheisel 2011*b*).

turnout.

3 Discussion

Causal inference with observational data must invariably rely on strong untestable assumptions. In this essay, we have delineated the most commonly invoked assumptions and used them in a case study of election day registration. Techniques like matching or differences-in-differences are often invoked as silver bullets which allow one to easily estimate causal effects. But this is simply not true. Differences-in-differences may be very plausible in one context, while less plausible in another. Gordon (2011) provides one example where differences-in-differences is plausible. We argue that differences-in-differences is much less plausible in the context of turnout since changes in the dynamics of elections from one year to the next may invalidate the key assumption. Without carefully specifying the underlying assumptions, inspecting the plausibility of those assumptions, and probing the sensitivity of inferences, it is difficult to make the move from correlation to causation. While assumptions are unavoidable in the study of politics, what needs to be clear is the role that assumptions play in the inference.

In general, we would argue that analysts should rely on a design-based inference. The design-based approach places explicit emphasis on reducing heterogeneity, clarity about identifying assumptions, a concern about endogeneity, and the role of research design (Imbens 2010, pg. 403). The design-based approach emphasizes that without a strong research design or a credible natural experiment, complex statistical modeling cannot give correlations a causal interpretation. The concepts we presented in Sections 1.3 and 1.5 are from this design-based literature. Even when this is true much can go wrong (Caughey and Sekhon 2011). Such are the perils of trying to estimate causal effects with observational data. As we have showed, the magnitude of our statistical estimates varied widely depending on what assumptions we used. No single study, including ours, is likely to be definitive, but when the role of assumptions in transparent the scientific community can more readily evaluate the

credibility of empirical evidence. In general, we are partisans of the design-based movement. In our EDR case study, it was the design-based elements that illuminated the weakness of specification assumptions, and it wasn't until we use the stronger design offered by RD that our estimate became credible.

Finally, while our main goal is to present a methodological argument, we believe our study has substantive implications as well. We demonstrate that EDR appears to have done little to change turnout even in Wisconsin and Minnesota. This may explain why states that later adopted EDR have seen little increase in turnout. In general, this challenges much of what we know about how state institutions affect turnout.

References

Abadie, Alberto. 2005. "Semiparametric Difference-in-Difference Estimators." *Review of Economic Studies* 75(1):1–19.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records." *American Economic Review* 80(3):313–335.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434):444–455.

Angrist, Joshua D. and J̈orn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D. and J̈orn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24(2):3–30.

Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98(4):687–702.

Barnow, B.S., G.G. Cain and A.S. Goldberger. 1980. Issues in the Analysis of Selectivity Bias. In *Evaluation Studies*, ed. E. Stromsdorfer and G. Farkas. Vol. 5 San Francisco, CA: Sage.

Brians, Craig Leonard and Bernard Grofman. 1999. "When Registration Barriers Fall, Who Votes? An Empirical Test of a Rational Choice Model." *Public Choice* 99:161–176.

Brians, Craig Leonard and Bernard Grofman. 2001. "Election Day Registration's Effect on U.S. Voter Turnout." *Social Science Quarterly* 82:170–183.

Burden, Barry C. and Jacob R. Neiheisel. 2011*a*. "The Effect of Election Day Registration on Voter Turnout and Election Outcomes." *American Politics Research* Forthcoming.

Burden, Barry C. and Jacob R. Neiheisel. 2011*b*. "Election Administration and the Pure Effect of Voter Registration on Turnout." *Political Research Quarterly* Forthcoming.

Caughey, Devin and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008." *Political Analysis* 19(4):385–408.

Cook, T.D. and W.R. Shadish. 1994. "Social Experiments: Some Developments Over the Past Fifteen Years." *Annual Review of Psychology* 45:545–580.

Cook, Thomas D., William R. Shadish and Vivian C. Wong. 2008. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27(4):724–750.

Donald, Stephen G. and Kevin Lang. 2007. "Inference With Differences-In-Differences and Other Panel Data." *The Review of Economics and Statistics* 89(2):221–233.

Erikson, Robert S. and Laura Stoker. 2011. "Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes." *American Political Science Review* 105(2):221–237.

Fisher, Ronald A. 1935. *The Design of Experiments*. London: Oliver and Boyd.

Gordon, Sanford C. 2011. "Politicizing Agency Spending Authority: Lessons from a Bush-era Scadal." *American Political Science Review* 105(4):717–734.

Green, Donald P. and Alan S. Gerber. 2002. Reclaiming The Experimental Tradition in Political Science. In *Political Science: The State of the Discipline*, ed. Ira Katznelson and Helen V. Milner. New York: W.W. Nortion pp. 805–832.

Hahn, Jinyong, Petra Todd and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatments Effects with a Regression-Discontinuity Design." *Econometrica* 69(1):201–209.

Hanmer, Michael J. 2007. "An Alternative Approach to Estimating Who is Most Likely to Respond to Changes in Registration Laws." *Political Behavior* 29(1):1–30.

Hanmer, Michael J. 2009. *Discount Voting*. New York, NY: Cambridge University Press.

Highton, Benjamin and Raymond E. Wolfinger. 1998. "Estimating the Effects of the National Voter Registration Act of 1993." *Political Behavior* 20(1):79–104.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *The American Economic Review Papers and Proceedings* 93(2):126–132.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2):399–423.

Imbens, Guido W. and Donald B. Rubin. 2008. *Causal Inference in Statistics and the Medical and Social Sciences*. Vol. Forthcoming Cambridge, UK: Cambridge University Press.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–476.

Imbens, Guido W. and Karthik Kalyanaraman. 2010. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator."

Just, Marion. 2011. "Same-Day Registration." http://www.nytimes.com/roomfordebate/2011/11/07/should-voting-in-the-us-be-mandatory-14/same-day-voter-registration-would-improve-turnout. Last Accessed: Nov 21, 2011.

Knack, Stephen. 2001. "Election-Day Registration: The Second Wave." *American Politics Research* 29:65–78.

Lee, David S. 2008. "Randomized Experiments From Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142(2):675–697.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontiuity Designs in Economics." *Journal of Economic Literature* 48(2):281–355.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review Papers and Proceedings* 80(2):319–323.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65(5):1311–1334.

Manski, Charles F. 2007. *Identification For Prediction And Decision*. Cambridge, Mass: Harvard University Press.

Mill, John Stuart. 1867. *A System of Logic: The Principles of Evidence and the Methods of Scientific Investigation*. New York, NY: Harper & Brothers.

Mitchell, Glenn E. and Christopher Wlezien. 1995. "Voter Registration and Election Laws in the United States, 1972-1992." *ICPSR* 6496:999.

Rhine, S.L. 1995. "Registration Reform and Turnout Change in American States." *American Politics Quarterly* 23:409–427.

Rosenbaum, Paul R. 2002*a*. "Covariance Adjustment In Randomized Experiments and Observational Studies: Rejoinder." *Statistical Science* 17(3):321–327.

Rosenbaum, Paul R. 2002*b*. *Observational Studies*. 2nd ed. New York, NY: Springer.

Rosenbaum, Paul R. 2005*a*. "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Obervational Studies." *The American Statistician* 59(2):147–152.

Rosenbaum, Paul R. 2005*b*. Observational Study. In *Encyclopedia of Statistics in Behavioral Science*, ed. Brian S. Everitt and David C. Howell. Vol. 3 John Wiley and Sons pp. 1451 – 1462.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer-Verlag.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6:688–701.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6:34–58.

Rusk, Jerrold G. 2001. *A Statistical History of the American Electorate*. Washington, D.C.: Congresional Quarterly Press.

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package For R." *Journal of Statistical Software* 42(7):1–52.

Sekhon, Jasjeet S. and Alexis Diamond. 2005. "Genetic Matching for Estimating Causal Effects." Unpublished Manuscript. Presented at the Annual Meeting of the Political Methodology, Tallahassee, FL.

Sekhon, Jasjeet S. and Rocío Titiunik. 2011. "When Natural Experiments are Neither Natural Nor Experiments." *American Political Science Review* Forthcoming.

Smolka, Richard G. 1977. *Election Day Registration: The Minnesota and Wisconsin Experience in 1976*. Washington, D.C.: American Enterprise Institute for Public Policy Research.

Sovey, J. Allison and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

Teixeira, Ruy A. 1992. *The Disappearing American Voter*. Washington D.C.: Brookings.

Thistlethwaite, Donald L. and Donald T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51(6):309–317.

Timpone, Richard J. 1998. "Structure, Behavior, and Voter Turnout in the United States." *American Political Science Review* 92(1):145–158.

Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.

Appendices

A Assumptions

We treated assumptions in a less formal manner in the text. Here, we present the various assumptions with formal notation using the potential outcomes framework.

A.1 Cross-Sectional Specification

The cross-sectional specification assumption can be written as a conditional ignorability assumption:

Assumption 1. *For any unit, the potential outcomes are independent of treatment assignment once we condition on the treatment assignment mechanism.*

$$Y_1,\ Y_0 \perp D \mid \mathbf{X}.$$

where $\mathbf{X}$ represents a matrix of variables that confound treatment with outcomes. This assumption can be written in a variety of other ways.

A.2 Differences-in-Differences

The identifying assumption for the DID estimator of treatment effects is:

Assumption 2. *Conditional on the covariates, expected potential outcomes for treated and control units follow parallel paths in the absence of treatment. In formal terms,*

$$E[Y_0(1) - Y_0(0)|X, D = 1] = E[Y_0(1) - Y_0(0)|X, D = 0].$$

A.3 Partial Identification

We consider two common assumptions to improve upon the no assumption bounds. The first assumption that we adopt is monotone treatment response (MTR) (Manski 1997). Under MTR, we assume

$$Y_{i1} \geq Y_{i0} \text{ or } Y_{i1} \leq Y_{i0}\ \forall\ i = 1, \cdots, n. \tag{1}$$

The second assumption we consider to sharpen the inference is that of monotone treatment selection (MTS). Formally, we express the MTS assumption as:

$$Pr[Y_1 = 1|D = 1] \geq Pr[Y_1 = 1|D = 0]$$
$$Pr[Y_0 = 1|D = 1] \geq Pr[Y_0 = 1|D = 0]$$

A.4 Instrumental Variables

Instrumental variables requires some additional notation. The treatment indicator remains $D \in \{0, 1\}$, but we introduce $Z \in \{0, 1\}$ as the indicator for encouragement. Typically Z is referred to as the instrument. In the IV setting, we seek to estimate the effect of D on Y using Z. For the IV estimand to be identified requires the following assumptions.

Assumption 3. *Random Assignment of the Instrument*

$$Pr(Z = 1) = Pr(Z = 0)$$

Assumption 4. *Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978)*

$$\begin{gathered} \textit{If } Z_i = Z_i', \textit{ then } D_i(t) = D_i(t') \\ \textit{If } Z_i = Z_i' \textit{ and } D_i = D_i', \textit{ then } Y_i(t, d) = Y_i(t', d') \end{gathered}$$

If these two assumptions hold, one can estimate what is known as intention-to-treat (ITT) effects without any further assumptions. This is simply the effect of Z on Y. To estimate the effect of D on Y requires three additional assumptions.

Assumption 5. *Exclusion Restriction*

$$Y_i(1, d) = Y_i(0, d) \textit{ for } d = 0, 1$$

In words, the exclusion restriction states that the effect of Z on Y must be entirely through the effect Z_i has on D_i, or Z must not have any direct effect on Y.

Assumption 6. *Nonzero Average Causal Effect of Z on D*

$$E[D_i(1) - D_i(0)] \neq 0$$

Assumption 7. *Monotonicity (Imbens and Angrist 1994)*

$$D_i(1) \leq D_i(0) \textit{ for all } i = 1, \ldots, N$$

A.5 Regression Discontinuity

Hahn, Todd and van der Klaauw (2001) demonstrate for an RD to be identified the potential outcomes must be a *continuous* function of the score. Under this continuity assumption, the potential outcomes can be arbitrarily correlated with the score, so that, for example, people with higher scores might have higher potential gains from treatment. This continuity assumption can be formally stated as:

Assumption 8. *The conditional regression functions are continuous in s at c:*

$$\lim_{s \to c} E(Y_{i0}|S_i = c) = E(Y_{i0}|S_i = c)$$

$$\lim_{s \to c} E(Y_{i1}|S_i = c) = E(Y_{i1}|S_i = c).$$

Since $Y_i = Y_{i1}$ when $D = 1$, $Y_i = Y_{i0}$ when $D = 0$, and $D_i = \mathbf{1}\{S_i \geq c\}$ where $\mathbf{1}\{\cdot\}$ is the indicator function, assumption 8 implies

$$\lim_{s \to c^+} E(Y_i|S_i = c) = E(Y_{i1}|S_i = c)$$

and

$$\lim_{s \to c^-} E(Y_i|S_i = c) = E(Y_{i0}|S_i = c),$$

which is a formal statement of the idea that individuals very close to the cutoff but on opposites sides of it are comparable or good counterfactuals for each other. Thus, continuity of the conditional regression function is enough to identify the average treatment effect *at the cutoff*. That is, the RD design identifies a *local* average treatment effect for the subpopulation of individuals whose value of the score is at or near c. Without further assumptions, such as constant treatment effects, the effect at c might or might not be similar to the effect at different values of S. Thus, under the continuity assumption, the RD identifies the following treatment effect:

$$\tau = E\{Y_{i1} - Y_{i0}|(S_i = c\} = \lim_{s \to c^+} E\{Y_i|S_i = c\} - \lim_{s \to c^-} E\{Y_i|(S_i = c\}$$

Estimation of this treatment effect proceeds by selecting a subset of units just above and below the discontinuity and calculating the difference across these two groups.

B Bounds Analysis

Below we outline, the method for calculating no-assumptions bounds, bounds under monotonicity, bounds under selection, and bounds under both monotonicity and selection. We first define the relevant quantities need for calculating the bounds. First, is the observed probability of treatment. Recall that we defined the treatment D as 1 if a state has EDR and 0 if not, and we defined the potential outcomes as: Y_0 or $Y(NoEDR)$ and Y_1 or $Y(EDR)$, where $Y = 1$ if a citizen votes and 0 otherwise. Table

Table 7: Observed Data For EDR in 2000

	$D = 0$ No EDR	$D = 1$ EDR
Did Not Vote $Y = 0$	39627	835
Voted $Y = 1$	71675	2807

One way to define the average treatment effect is as follows:

$$
\begin{aligned}
E[\Delta] &= \{Pr(D=1)Pr[Y_1=1|D=1] + Pr(D=0)Pr[Y_1=1|D=0]\} \\
&- \{Pr(D=1)Pr[Y_0=1|T=1] + Pr(D=0)Pr[Y_0=1|D=0]\}
\end{aligned} \tag{2}
$$

With observational data four of the terms in Equation 2 are identified in the data. In the data used here $Pr(D=1) = .032$ and of course $Pr(D=0) = 1 - Pr(D=1) = .968$. Using the observed data, I can calculate two other quantities. The first is

$$Pr[Y_1 = 1|D=1] = .77 \tag{3}$$

which is the probability that a republican wins in the subpopulation that received a presidential visit. The second is:

$$Pr[Y_0 = 1|D=0] = .64 \tag{4}$$

which is the Republican win probability in the subpopulation not receiving a presidential visit. This leaves two counterfactual quantities that are not observed:

$$Pr[Y_1 = 1|D=0]$$
$$Pr[Y_0 = 1|D=1]$$

To calculate Manksi's no-assumption bounds, we set these counterfactual quantities to their maximum and minimum values. To calculate the lower bound on the treatment effect, I set:

$$Pr[Y_1 = 1|D=0] = 0$$
$$Pr[Y_0 = 1|D=1] = 1$$

For the upper bound on the treatment effect, we set:

$$Pr[Y_1 = 1|D=0] = 1$$
$$Pr[Y_0 = 1|D=1] = 0$$

Under these values, the no-assumption bounds are:

$$\{Pr(D=1)Pr[Y_1=1|D=1]\} - \{Pr(D=1) + Pr(D=0)Pr[Y_0=1|D=0]\}$$
$$\leq E[\Delta] \leq$$

$$\{Pr(D=1)Pr[Y_1=1|D=1]+Pr(D=0)\}-\{Pr(D=0)Pr[Y_0=1|D=0]\}$$

As reported earlier in the paper, these bounds are $[-0.63, 0.37]$ (Manski 1990). The next assumption of interest is monotone treatment response (MTR) (Manski 1997). Under MTR, I assume that the individual level treatment effect is positive such that $\Delta \geq 0$ for every individual i. This implies

$$Pr[Y_1=1|D=0] = Pr[Y_0=1|D=0]$$
$$Pr[Y_0=1|D=1] = Pr[Y_1=1|D=1]$$

Therefore under MTR the bounds on the average treatment effect would be $[0, 0.37]$. The next assumption goes under a number of names including outcome optimization, selection, or monotone treatment selection (MTS) (Manski 2007). Under this assumption, we assume that units either select into treatment in order to achieve an optimal outcome or that units are selected for treatment to optimize the outcome. Under MTS we assume:

$$Pr[Y_1=1|D=1] \geq Pr[Y_1=1|D=0]$$
$$Pr[Y_0=1|D=1] \geq Pr[Y_0=1|D=0]$$

MTS puts a bound on the upper bound of the no-assumption bounds. Under MTS, the bounds for the average treatment effect is now $[-0.63, 0.127]$. If we combine the MTR and MTS assumptions the bounds are $[0, 0.127]$. By definition under MTR and MTS, the bounds on the treatment effect are always 0 for the lower bound, with the upper bound being the naive estimate of the average treatment effect.

C RD Analysis

In recent years, a number of methods have been proposed for estimation of RD estimates outside of simple plots. There are two related issues that analysts must contend with when estimating treatment effects in an RD design. First, one must select a local neighborhood around the discontinuity. In the RD design, we believe that observations near the cutoff

are good counterfactuals, the questions is how far must an observation be from the cutoff before we think observations are no longer good counterfactuals. Therefore, the analyst must select some local neighborhood above and below the cutoff. Two methods that are widely used to select the size of the local neighborhood are cross-validation (Imbens and Rubin 2008) and algorithmic MSE-minimization (Imbens and Kalyanaraman 2010). We might also select the neighborhood based on whether baseline covariates are balanced within that area. Unfortunately, in our application we don't have baseline covariates for most observations.

Once the size of the local neighborhood is chosen one can estimate either an unweighted mean difference or use (local) linear regression to estimate a conditional expectation for each side of the discontinuity and take the difference in these conditional expectations. In these methods, all observations in the local neighborhood receive equal weight. One can also use a kernel function to give observations closer to the discontinuity greater weight than those observations farther from the cutoff (Imbens and Kalyanaraman 2010). Inference can proceed either via large sample standard errors or the bootstrap. We found that no matter which method we used, the estimates and our inferences were unchanged. Our inferences were insensitive to a wide range of local neighborhood width choices. We report estimates with bandwidth selected via MSE-minimization and using a triangular kernel function with local regression. For inference we use bias-corrected and accelerated (BC_a) bootstrap confidence intervals. For the 325 municipalities with census covariates, we plot each covariate against the score in Figure 3. We observe no evidence of any obviously correlation between these covariates and the score at the discontinuity.

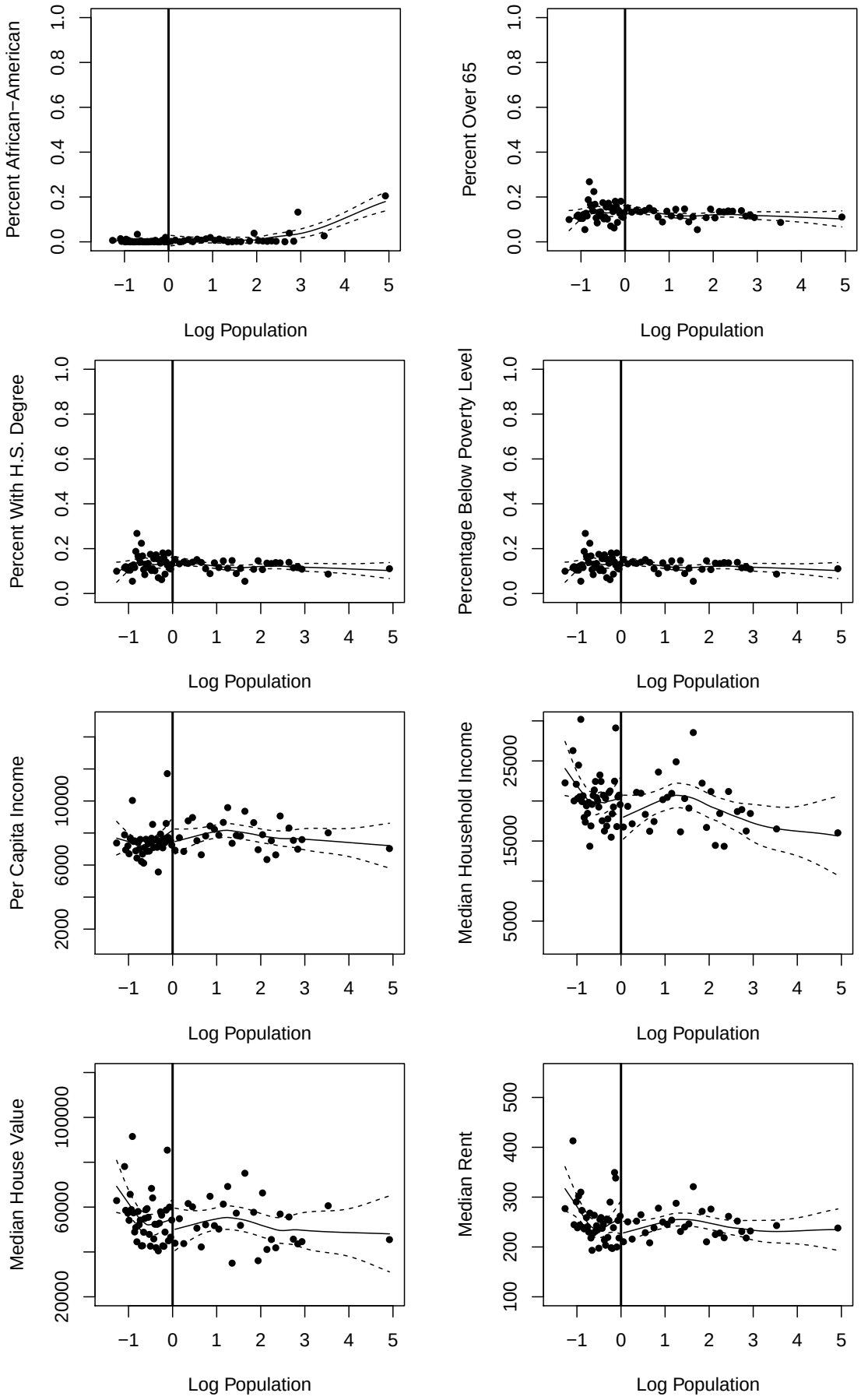

Figure 3: Regression Discontinuity Design: Other Census Covariates