Credible Causal Inference for Empirical Legal Studies

Daniel E. Ho¹ and Donald B. Rubin²

¹Stanford Law School, Stanford University, Stanford, California 94305; email: dho@law.stanford.edu
²Department of Statistics, Harvard University, Cambridge, Massachusetts 02138; email: rubin@stat.harvard.edu

Abstract

We review advances toward credible causal inference that have wide application for empirical legal studies. Our chief point is simple: Research design trumps methods of analysis. We explain matching and regression discontinuity approaches in intuitive (nontechnical) terms. To illustrate, we apply these to existing data on the impact of prison facilities on inmate misconduct, which we compare to experimental evidence. What unifies modern approaches to causal inference is the prioritization of research design to create—without reference to any outcome data—subsets of comparable units. Within those subsets, outcome differences may then be plausibly attributed to exposure to the treatment rather than control condition. Traditional methods of analysis play a small role in this venture. Credible causal inference in law turns on substantive legal, not mathematical, knowledge.

Keywords

research design, policy evaluation, matching, regression discontinuity
1. MOORE’S LAW OF PARKING

It would be easy to dismiss the parking studies of Underhill Moore. From 1933 to 1937, the famed Yale Law professor sought to quantify the causal effect of law. He worked with a cadre of research assistants to count over 13,000 instances of parked cars spanning 15 New Haven areas, dispatched police officers to place tags for dollar fines on over 3,400 cars, and painted large white ovals to simulate a would-be roundabout in the middle of an intersection. While the research entailed the minutiae of defining when a car had parked (when the wheels stopped moving), the goals were lofty: nothing short of a “general theory of human behavior” in relation to law (Moore & Callahan 1943, p. 2). Moore himself admitted that the venture walked a fine line between the avant-garde and the absurd: “They ridicule my project. They do not understand it . . . I am writing for [those] groping for ways of applying the scientific method to the social sciences . . . . [Y]ears from now a kindred soul may find in my crude researches some clue to the solution” (Douglas 1950, p. 188).

Despite the rather obvious mismatch of legal theory and empirical data, as a matter of methodology, the ridicule is misplaced. Moore’s research, like much of the first wave of empirical legal studies in the 1920s and 1930s, grappled with thorny methodological challenges to drawing inferences about the causal effects of laws, all while modern foundations for experiments were only beginning to take shape (Schlegel 1995). It was not until 1925 that Fisher offered randomization as the “reasoned basis for inference” for experiments (Fisher 1925, 1935). How then could one infer the causal effect of a parking regulation? As William O. Douglas and collaborators contemporaneously noted, “[P]roblems will center around the development of more adequate techniques for controlling errors and the production of data from which inferences as to the causal connection of these various factors . . . will emerge” (Clark et al. 1930). More generally, how could the first empiricists quantitatively assess the impact of law?

Moore’s approach was pioneering, if not downright modern. He reasoned that when an “experimental situation[] could not be manufactured at will,” one could “take advantage of the terms of the ordinance itself” (Moore & Callahan 1943, pp. 88–89). This insight was crucial. Simply examining streets with or without limits would be comparing the incommensurable. Moore’s solution capitalized on the geographic or temporal arbitrariness of when a parking time limit applied. On Crown Street, the limit applied on one side of the street for one month and on the other side the next. On Church Street, the 15-minute limit applied only until 7:00 PM. And so Moore collected data immediately before and after 7:00 PM to isolate the impact of the parking regulation. To assess whether the differences in the time of day affected inferences, Moore further checked for similarity of traffic flows and driver activities, the latter monitored by research assistants following subjects to destinations.

Figure 1 presents the data that Moore collected for one street, with minutes parked on the x-axis (on a log scale). (The bins are in 1-, 5-, or 10-minute increments as presented by Moore.) The dark green outlined histogram presents parking durations for cars parked from 5:30–6:30 PM, when the 15-minute limit applied. The light green filled histogram presents parking durations for cars parked from 7:00–8:00 PM, when no limit applied. Parking durations shifted considerably. Roughly 36% of cars parked for 15 minutes or less when the limit was inapplicable, compared with 57% when the limit applied. On average, the effect of the time limit was to decrease the time in the space by 40 minutes, although a large number of drivers still failed to comply with the time limit in place.

Of course, from a modern perspective, Moore’s methodology is lacking in certain respects. Demand for parking may differ sharply after 7:00 PM. The time limit may still affect parking behavior after 7:00 PM. And differences
New Haven parking on Church Street for 35 days in 1936. The light green histogram represents duration of parking starting from 7:00–8:00 PM when the 15-minute parking limit (gray dashed line) was inapplicable. The dark green outlined histogram represents duration of parking starting from 5:30–6:30 PM when the 15-minute parking limit was applicable. The x-axis is on a log scale and bin widths are as Moore & Callahan reported (1 minute up to 30 minutes, 5 minutes from 30–100 minutes, and 10 minutes from 100–450 minutes). Source: Moore & Callahan (1943, pp. 104–6).

In this article, we review modern developments in the statistics of causal inference, focusing in particular on matching methods and regression discontinuity. What unifies such approaches is the prioritization of research design to create—without reference to any data on outcomes—subsets of comparable units. In what might be considered a vindication (or even “kindred soul”) of Moore, modern approaches emphasize design over methods of analysis.

Our article proceeds as follows. Section 2 discusses the broad shift toward credible, design-oriented inference in social science. Section 3 explains the widely used potential outcomes framework that clarifies the central issues of causal inference. We use as a running example data first analyzed in an important study by Berk & de Leeuw (1999) (BdL) of the causal effect of maximum-security incarceration on prison misconduct, which we detail in Section 4. Section 5 uses this data set to illustrate the chief problem of “model sensitivity” that plagues much conventional regression-based practice. Section 6 details what we mean by a focus on research design: collecting, organizing, measuring, and preparing the data without reference to outcome data. Sections 7 and 8 apply matching methods and regression discontinuity to BdL’s prison data, which we compare to experimental results in Section 9. Both approaches provide estimates much closer to experimental findings than do naive regression-based approaches. Section 10 concludes.

---

Figure 1

---

1 In modern terminology, these defects would refer to issues of continuity/smoothness of outcomes with respect to the forcing variable, the “stable unit treatment value assumption,” and sampling variability.
2. CAUSAL INference in Empirical Legal Studies

Causal inference has always been central to the enterprise of empirical legal studies. How does no-fault insurance law affect auto injury compensation? Do defendants with court-appointed counsel fare worse than those with retained counsel? How does discretionary jurisdiction affect the business of the Supreme Court? All these were questions that led the likes of Roscoe Pound, Felix Frankfurter, and James Landis to turn to quantitative data collection in the 1920s and 1930s (Kritzer 2010). Yet their efforts met with frustration. Said William O. Douglas at the conclusion of a project on the causes of bankruptcy: “All the facts which we worked so hard to get don’t seem to help a hell of a lot” (Schlegel 1995, p. 230).

More recently, a similar frustration has surfaced in cognate disciplines about the limits of conventional (regression-based) causal inference. Clearer conceptualization of causal inference has led to an increasing skepticism about the “age of regression” (Morgan & Winship 2007; see also Berk 2004; Donohue & Wolfers 2006; Gelman & Meng 2004; Leamer 1978, 1983; Manski 1995; Pfaff 2010; Sobel 2000; Strnad 2007). “Without... strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive” (Sekhon 2009). Or, as Douglas might note: “All the [regressions] we worked so hard to get don’t seem to help a hell of a lot.”

Yet something else is afoot. Ayres (2008) dubbed the use of large-scale microdata and field experiments the age of “super crunching.” Two leading economists have coined it the “credibility revolution” (Angrist & Pischke 2010). And one researcher forecasts “dramatic transformation” in social science with a deeper understanding of causal inference (Sobel 2000). The unifying feature of this movement is the attempt to hew as closely as possible to an experiment. The law has not remained untouched by this movement. Experimental approaches have reinvigorated our understanding of discrimination (Ayres 1991, Pager 2003), corporate governance (Guttentag et al. 2008), the legal profession (Abrams & Yoon 2007), and health care (King et al. 2007), to name just a few (for others, see, e.g., Angrist 1990, Gerber & Green 2000, Gibson 2008, Green & Winik 2010, Ho & Imai 2006). Even when there is no randomized intervention (when the study is “observational”), approaches directly appealing to an experimental template have crystallized the key issues for empirical inference. Matching methods, for example, have been applied to race and sex (well defined only in certain contexts) (Boyd et al. 2010, Greiner et al. 2006), corporate governance (Litvak 2007), labor and employment (Dehejia & Wahba 2002, Morantz 2010), environment (List et al. 2006), regulation (Galiani et al. 2005), constitutional law (Persson & Tabellini 2002), election law (Brady & McNulty 2007), civil rights (Epstein et al. 2005), and education (Ho 2005a,b).

The credibility revolution: number of articles discussing matching and regression discontinuity in top 23 economics, political science, sociology, statistics & probability, and social science (mathematical methods) journals. All top 10 journals based on 2009 impact factor in Journal Citation Reports categories for which full-text searches were available from 1975–2010 were chosen, with duplicated journals omitted. Search strings were for “matching methods,” “regression discontinuity,” “propensity score,” “Thistlethwaite/5 Campbell,” and “matching and ‘potential outcomes’” in JSTOR, ProQuest, and journal-specific Web sites.

Yet scholars not following these developments may be baffled. How should legal scholars understand and assess these approaches? What principles can legal empiricists incorporate from this rapidly growing literature? Are they in fact more credible? We provide a first guide and review to begin to answer these questions for a general legal audience.

3. POTENTIAL OUTCOMES

We begin by articulating a widely used framework for causal inference, often called the “Rubin Causal Model” (Holland 1986) owing to a series of seminal papers by Rubin (Rubin 1974, 1976, 1978, 1979; Rosenbaum & Rubin 1983b, 1984). The idea is deceptively simple, yet it clarifies the key conceptual issues of causal inference and can be explained without math. Specifically, we are interested in the effect of a single intervention, which we refer to as the “treatment,” compared with the baseline of “control.” For example, one crucial question for prison administration is the causal effect of maximum-security imprisonment (treatment) versus minimum-security imprisonment (control) on the outcome of misconduct.

Each unit then has two “potential outcomes,” one under treatment and one under control. The fundamental problem of causal inference is that we never observe both (Epstein & King 2002, Holland 1986, Rubin 1978). If a prisoner is sent to a maximum-security prison, we cannot observe the “counterfactual” outcome of how she might have fared in a minimum-security prison. Implicit in this framework is that (a) there are no hidden versions of the treatment and (b) treatment of one unit does not affect the potential outcomes of another unit (sometimes referred to as the stable unit treatment value assumption, SUTVA). The former would be violated, for example, if two types of maximum-security prisons were available to one prisoner, each with different effects on that prisoner’s behavioral misconduct. The latter would be violated, for example, if the prison assignment of one gang member affected the behavior of a member of an opposite gang.

Figure 3 visualizes how this framework applies to a data set. The box can be considered the data set, with units in rows, pretreatment covariates in the left columns, potential outcomes in the middle two columns, and an
indicator for treatment in the last column. Green cells indicate that data are observed, and white cells represent "missing data." For example, we observe the outcomes under control for the upper half of the data set and the outcomes under treatment for the lower half of the data set.

This framework highlights several points. First, causal inference is a matter of inferring missing data. Because we never observe counterfactual outcomes, causal inference is inherently uncertain (and hence a probabilistic venture). Second, causal inference is difficult to conceive of without an intervention. More succinctly, “No causation without manipulation” (Rubin 1975, Holland 1986). This poses a particular challenge for empirical inference in areas such as antidiscrimination law, where immutable characteristics per se cannot be manipulated in a clearly defined way (Greiner & Rubin 2010). Lastly, the framework highlights the importance of an experimental template for the research. If resources were no constraint, researchers should be able to articulate how one might design an experiment to study the question of interest.

Why is the intellectual idea of an experiment so crucial, even in observational research? The key feature of an experiment is that treatment is randomly assigned to units. Randomization over a large number of units ensures that treatment and control units are comparable in all respects other than the treatment. ‘Balance’ exists along all possible covariates. Randomizing prisoners to security levels would ensure that maximum-security prisoners, for example, are similar in age, sex, and criminal history to non-maximum-security prisoners. We can thereby properly infer the missing potential outcomes and hence estimate the “treatment effect.”

In observational settings, differences in the outcomes may be "confounded" by other factors. Prison authorities, for example, may intentionally sort prisoners by risk profile into facilities of different security levels. Thus, differences in behavior between maximum- and non-maximum-security prisoners would be confounded by risk profile. Observational research can be seen as replicating the hypothetical experiment by achieving balance on these confounding (pretreatment) covariates. The crucial assumption in most observational studies then boils down to unconfoundedness (alternatively known as exogeneity, conditional exogeneity, ignorability, or selection on observables): that, given covariates, the treatment is random, so researchers can attribute differences to the treatment. The credibility of unconfoundedness, as discussed below, is a qualitative judgment that depends crucially on substantive knowledge.

4. APPLICATION: MAXIMUM-SECURITY PRISONS AND MISCONDUCT

As a running example to fix ideas, we use a prison data set first analyzed by BdL. The data set contains information for 3,918 California inmates admitted to prison in 1994. We (and BdL) are interested in the causal effect of maximum-security imprisonment on misconduct. We use maximum security as shorthand for facilities that have “inside or outside cell construction with a secure perimeter, and both internal and perimeter armed coverage” (CDC 2000, ch. 6, art. 5, §62,010.6). (Variation within such Level IV facilities still exists, but we follow BdL and focus only on the impact of Level IV.)

To assess the credibility of a causal inference, understanding the treatment assignment process is crucial. California’s procedure in 1994 for assigning inmates into prison worked in three steps (BdL, CDC 2000, Petersilia 2008). First, after a defendant was sentenced, the California Department of Corrections (CDC) classified inmates by security risk.² CDC used inmate background, prior escape, and prior incarceration information to

²Of course, CDC here does not refer to the Centers for Disease Control and Prevention, but we use the acronym to be consistent with BdL.
Table 1  Example of inmate classification. For example, a sentence length of 10 years would result in 27 points \[ = (10−1) \times 3 \] added to the classification score and no high school degree adds 2 points. Given this background and prior incarceration behavior, the inmate would be assigned a score of 53. The example is only illustrative, as other factors (favorable prior behavior, undocumented prior behavior) are taken into account [CDC 2000, ch. 6, art. I, § 61010.11.2 (Form 839)]

<table>
<thead>
<tr>
<th>Factors</th>
<th>Calculation</th>
<th>Value</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence length ( (x) )</td>
<td>((x - 1) \times 3)</td>
<td>10</td>
<td>27</td>
</tr>
<tr>
<td>Under age 26</td>
<td>+ 2</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Not married</td>
<td>+ 2</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>No high school degree</td>
<td>+ 2</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Unemployed</td>
<td>+ 2</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>No military service</td>
<td>+ 2</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Number of escapes from minimum custody</td>
<td>(x \times 4)</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Number of escapes from medium custody</td>
<td>(x \times 8)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of escapes with force</td>
<td>(x \times 16)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Prior incarceration behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of serious disciplinaries</td>
<td>(x \times 4)</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Number of assaults on staff</td>
<td>(x \times 8)</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Number of assaults on inmates</td>
<td>(x \times 4)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of possessions of deadly weapon</td>
<td>(x \times 4 - 8)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of inciting disturbances</td>
<td>(x \times 4)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of assaults causing serious injury</td>
<td>(x \times 16)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total score</strong></td>
<td></td>
<td></td>
<td>53</td>
</tr>
</tbody>
</table>

calculate a “classification score” ranging from 1–80. *Table 1* sketches how major factors were incorporated for a hypothetical inmate, resulting in an overall score of 53. Sentence length was the primary factor, with each additional year resulting in 3 more points, but age, marital status, high school degree, employment, military service, and prior escape attempts were also included. Any prior physical assault on prison staff would add 8 points to the score. Across the sample, the mean classification score was 31 (SD = 12).

Second, in most cases CDC exclusively used the classification score to assign inmates to facilities of given security levels. The CDC *Operations Manual* provided that a classification score of 52 or higher would lead to maximum-security confinement (CDC 2000, ch. 6, art. 1, § 61010.11.4). The first row of *Table 2* shows that inmates in non-maximum-security facilities had an average score of 24 (SD = 13), compared with 66 (SD = 12) in maximum-security facilities.

Third, in certain “administrative placements,” CDC deviated from the score based on other attributes. Sex offenders, for example, were more likely to be placed in maximum-security prisons (BdL). Overcrowding could result in alternate placement. Other special case factors included *(a)* whether behavior indicated that an inmate was “capable of successful placement” at a lower level facility, *(b)* the existence of documented “enemies” at institutions, *(c)* family ties, *(d)* medical conditions, and *(e)* work skills [CDC 2000, ch. 6, art. I, § 61010.11.2 (Form 839)]. Because the CDC score was the primary determinant of inmate placement, we also refer to it as the “forcing” variable, namely the variable that “forces” the treatment of maximum-security confinement. Administrative placements, however, mean that the classification score only
Table 2  Summary statistics of prison incarceration data. The first two columns present statistics for prisoners in non-maximum-security prisons (“controls”); the third and fourth columns present statistics for maximum-security prisoners (“treatment”); the fifth and sixth columns present statistics for all subjects in the data set. The last column presents the p-value testing for the difference in means or proportions between treatment and control groups. For the inmate classification score (an ordinal measure), the statistics are means and SDs, while for strike-three offense and behavioral misconduct (binary measures), the statistics are counts and proportions (of the subgroup).

<table>
<thead>
<tr>
<th></th>
<th>Non–maximum security</th>
<th>Maximum security</th>
<th>All</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean/count</td>
<td>Mean/count</td>
<td>Mean/count</td>
<td>SD/prop.</td>
</tr>
<tr>
<td>Inmate classification score</td>
<td>24/13</td>
<td>66/12</td>
<td>31/12</td>
<td>0.000</td>
</tr>
<tr>
<td>Strike-three offense</td>
<td>208/0.07</td>
<td>523/0.72</td>
<td>731/0.19</td>
<td>0.000</td>
</tr>
<tr>
<td>Behavioral misconduct</td>
<td>890/0.28</td>
<td>246/0.34</td>
<td>1136/0.29</td>
<td>0.002</td>
</tr>
<tr>
<td>Total number</td>
<td>3188/0.81</td>
<td>730/0.19</td>
<td>3918/1.00</td>
<td></td>
</tr>
</tbody>
</table>

The last row of Table 2 shows that roughly 81% of inmates were placed in non-maximum-security prisons (control), whereas 19% were placed in maximum-security prisons (treatment). Table 2 presents two further key variables. One key pretreatment covariate was whether an inmate was sentenced under California’s “three strikes” law. (The full set of covariates used in the intake procedure was, unfortunately, not available to us.) Under California law, a third serious felony led to sentence enhancements. These, in turn, increased the probability of maximum-security-level assignment. In the BdL data, roughly 72% of maximum-security inmates were three-strike inmates, compared with only 7% of non-maximum-security inmates. The outcome of interest is whether the inmate was cited for any instance of behavioral misconduct while imprisoned (e.g., failure to obey an order, drug trafficking, or assault). Roughly 29% of inmates—34% of maximum-security and 28% of non-maximum-security inmates—engaged in misconduct.

The last column reports results from tests of differences between the treatment and control groups in the classification score (the probabilistically forced treatment). The last row of Table 2 shows that roughly 81% of inmates were placed in non-maximum-security prisons (control), whereas 19% were placed in maximum-security prisons (treatment).

Table 2 presents two further key variables. One key pretreatment covariate was whether an inmate was sentenced under California’s “three strikes” law. (The full set of covariates used in the intake procedure was, unfortunately, not available to us.) Under California law, a third serious felony led to sentence enhancements. These, in turn, increased the probability of maximum-security-level assignment. In the BdL data, roughly 72% of maximum-security inmates were three-strike inmates, compared with only 7% of non-maximum-security inmates. The outcome of interest is whether the inmate was cited for any instance of behavioral misconduct while imprisoned (e.g., failure to obey an order, drug trafficking, or assault). Roughly 29% of inmates—34% of maximum-security and 28% of non-maximum-security inmates—engaged in misconduct.

The last column reports results from tests of differences between the treatment and control groups in the classification score (the forcing variable), strike-three offense (the covariate), and behavioral misconduct (the outcome). We provide these tests only for expository purposes; as we emphasize below, the design phase should not examine final outcome data. Although the raw difference in the outcome is statistically significant (p-value = 0.002), treated inmates also had statistically significantly higher classification scores and third strikes. Three-strike inmates were likely prone to more dangerous conduct, hence confounding the raw difference.

Figure 4 plots the classification score on the x-axis against the probability of misconduct on the y-axis. Each dot represents the proportion of prisoners that engaged in behavioral misconduct at a given classification score and security level. The solid gray and hollow green dots represent inmates in non-maximum- and maximum-security prisons, respectively, and are proportional to sample size. For example, the large gray dot in the bottom left represents 125 non-maximum-security inmates with a classification score of 1, 25% of whom engaged in misconduct. The vertical line represents the threshold of the classification score used to assign inmates to maximum-security prison. Ninety-two percent of inmates with scores of 52 or above were assigned to maximum-security prison, while 98% with scores below 52 were assigned to non-maximum-security prison.

---

Some use the terms “sharp” and “fuzzy” regression discontinuity to distinguish whether the threshold of the forcing variable deterministically or probabilistically assigns treatment.
Figure 4
Outcome of behavioral misconduct in prison against the inmate classification score. The x-axis presents the inmate classification score, and the y-axis presents the proportion of prisoners at each score engaging in behavioral misconduct while imprisoned. Filled gray dots indicate prisoners in non-maximum-security prisons, and hollow green dots indicate prisoners in maximum-security prisons. Dots are proportional to sample size, so the large dot in the bottom left represents 125 prisoners in non-maximum-security prisons, 25% of whom engaged in misconduct. The vertical gray dashed line represents the threshold of the classification score, which substantially increases the probability of maximum-security confinement.

Because the classification score is only used probabilistically, inmates could still be assigned to either level of security across the entire range of the classification score. For example, the small green dot at the left represents two inmates with a score of 1 confined to maximum-security prison, one of whom engaged in misconduct. Overall, 67 inmates with scores of 52 or above were placed in non-maximum-security prisons, and 51 inmates with scores below 52 were placed maximum-security prisons. As discussed below, the visualization in Figure 4 can help considerably in grasping how a causal effect is identified by regression, matching, and regression discontinuity approaches.

BdL originally used the data to study (a) the effectiveness of CDC risk sorting and (b) the causal effect of maximum-security imprisonment on prison misconduct. Logically, even the direction of the causal effect is unclear. Stronger security measures may deter misconduct (Zimring & Hawkins 1973), or such facilities may induce marginal inmates to acquire deviance from the worst inmates, thereby increasing misconduct (Bayer et al. 2009). Based on a (logit) regression model that capitalized on the discontinuity in the assignment process and laudable sensitivity analyses, BdL concluded that “the balance of evidence supports an interpretation in which assignment to [maximum security] reduces the odds of misconduct” (BdL, p. 1052). Although the analysis below diverges from these findings, these approaches have been rapidly developing in the past decade and so are reconducible to most researchers. We use BdL because it is a landmark study that deserves recognition not only for its central insights in research design, but also for applying them into...
pioneering field experiments that enable us to assess the validity of observational approaches.

5. INCREDIBLE INFERENCE: CONVENTIONAL REGRESSION-BASED PRACTICE


To illustrate this fact, we apply naive regression-based approaches to the prison data. Each of the panels in Figure 5 overlays model-based (pointwise) 95% confidence intervals to summarize the results from a range of regression models against the prison data. For example, the top left panel presents the (logit) model reported by BdL. The gray band plots the predicted probability of misconduct for non-maximum-security prisoners, and the green band plots the predicted probability of misconduct for maximum-security prisoners. These curves, if correctly specified, allow us to impute counterfactual outcomes. The difference between the two is the estimated average treatment effect: Maximum security decreases misconduct by 13%, plus or minus 4%.

But the model imposes two strong and unwarranted assumptions. First, it assumes that the probability effectively has a linear relationship with the classification score (more precisely, linearity in the log odds). Second, it assumes that the relationship between the classification score and misconduct is homogeneous across treatment and control groups. These curves, if correctly specified, allow us to impute counterfactual outcomes. The difference between the two is the estimated average treatment effect: Maximum security decreases misconduct by 13%, plus or minus 4%.

How would a researcher determine the “best” model? How much smoothness should be assumed? Should we impose homogeneity? Linearity? Even with just one covariate, a staggering set of specification choices presents itself. One “nonparametric” way forward would be to estimate the probability at each classification score for treatment and control groups, resulting in 160 parameters. But as other covariates are added, the number of parameters grows exponentially. Adding 360 possible months of sentence length, the number of parameters becomes 57,600 (360 × 160); adding ages of 15–64 years, the number becomes 2,880,000 (50 × 57,600); adding employment status, sex, prior strikes, and marital status leads to over 69 million parameters. Conventional results—which often impose strong and unwarranted functional form assumptions—can be fragile. When groups differ sharply, regression may not credibly “control” for confounding factors.
And regression does not amount to research design.

6. CREDIBLE INFERENCE: DESIGN TRUMPS ANALYSIS

Our central message is that research design trumps methods of analysis. By research design we mean “contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data” (Rubin 2008). Methods of analysis, in contrast, involve the development of a model for outcomes (e.g., linear regression, generalized linear models, machine learning algorithms). Just as experiments elaborate a procedure...
without knowing values of the outcome, observational studies can be designed according to key principles.

First, outcome data should be set aside at the design phase. Classical \( p \)-values from statistical tests are inappropriate when models are fit multiple times. The possibility of inadvertently choosing a model with a particular result threatens credibility.

Second, the crucial element of design is to use all covariate information to achieve balance along all important pretreatment covariates between treatment and control groups. In Section 7, we show how to balance by matching prisoners on exact classification scores. In Section 8, we note how prevailing practice of regression discontinuity has design and analysis reversed.

Third, the researcher must make a qualitative assessment of the substantive credibility of the “identifying” assumptions. What is the so-called “identification strategy”? Do the data contain enough covariates to make matching credible? Are the covariates properly pretreatment covariates? Are subjects able to manipulate treatment assignment?

There is no substitute for substantive knowledge. Consider a compelling study of the effect of classroom size on educational outcomes by Angrist & Lavy (1999). The study capitalized on “Maimonides’s rule” in the Israeli public school system that sets a strict cap on classroom size at 40 students. Because it is plausibly random whether the enrollment at the beginning of the school year is just below or above 40, we can credibly assess the impact of class size by comparing class sizes of 20 and 21 resulting from enrollments of 41 to a class size of 39. Although highly credible in the context of Israeli public schools, in other jurisdictions where parents can switch schools upon discovery of a large classroom, the comparison can be contaminated (Angrist & Pischke 2010, p. 14; Urquiola & Verhoogen 2009). Credibility hence depends on deep, substantive knowledge of the legal system being examined.

7. MATCHING

Matching reduces the role of strong and unwarranted functional form assumptions by trimming the data set down to treatment and control groups that are balanced along pretreatment covariates. The key assumption is that, conditional on covariates, treatment is random. The credibility depends entirely on (a) whether enough relevant (pretreatment) covariates have been collected and (b) whether sufficient balance has been achieved between treatment and control groups.

In the prison data, is the treatment plausibly random given a specific classification score? How administrative placements are made is crucial here. If sex offenders are the only prisoners with scores below 52 placed in maximum-security prison, we may be effectively comparing the propensity for misconduct of sex offenders and non–sex offenders at low scores. Do CDC officials differ systematically in the use of administrative placements? If so, how are officials assigned? If overcrowding at maximum-security-level prisons results in prisoners with high scores being placed in non-maximum-security prisons, is the timing of overcrowding random or might waves of gang violence explain overcrowding shocks (when gang membership may generally lead to more behavioral misconduct)? To ground the assumptions, substantive knowledge and research are required.

Assuming that given a score prison assignment is random, the best practice is to report how much balance has improved after matching. Matching exactly on classification score solves that problem in our data. In other instances, multiple and continuous covariates can make exact matching impossible, in which case “propensity score” matching provides a way forward. For inexact matches, researchers should do everything to achieve the best balance using substantive knowledge. For example, in a study of a drug’s impact on birth defects, matching women on age requires scientific knowledge. A two-year difference between a 21- and a 23-year-old may be trivial, but a one year difference between a
Figure 6
Quantile-quantile plot of balance of inmate classification score between treatment and control groups. Panel (a) plots the raw data, showing that maximum-security prisoners have far higher classification scores. Panel (b) plots the matched data, showing balance on the classification score (points falling along the 45-degree line). The latter exhibits slight sampling variability by sampling proportional to weights from exact matching.

41- and a 42-year-old could invalidate the study. Fortunately, legal academics are precisely the ones who harbor the deepest knowledge about the legal system under study and are thus often in the best position to evaluate balance.

**Figure 6a** plots the quantile-quantile plot of the control group on the y-axis and the treatment group on the x-axis. If there is balance, the dots should line up on the 45-degree line. The raw difference, however, is stark. The right panel presents the same plot for the matched data set. Unsurprisingly, because units are exactly matched, balance is good.

**Figure 7** presents the difference in misconduct probability at each classification score where there are both treatment and control units. For example, the leftmost dot represents the 25% difference at score 0 between the 1 of 2 maximum-security prisoners and 25% of 125 non-maximum-security prisoners who engaged in misconduct. The intervals represent 95% confidence intervals, and dots are weighted by sample size of the smallest group, with 1,910 units in the matched sample. These conditional effects show no pattern, and most contain the origin. To calculate an overall effect, we can use a weighted average across these categories (weighted by the number of treated units at each score, represented by the hollow green circles), resulting in an effect estimate of 0.03, plus or minus 0.08 (the gray interval). In other words, comparing inmates with identical classification scores, maximum security causes from a 5% decrease to an 11% increase in misconduct. Although the interval is fairly informative, we cannot reject the null hypothesis that security level has no impact on behavior.

### 8. REGRESSION DISCONTINUITY

The key assumptions of regression discontinuity (RD) are that (a) treatment assignment is discontinuous at a threshold of the forcing variable, which cannot be precisely manipulated, and (b) all other covariates are smooth (or balanced) at the threshold. Under those assumptions, units just below and above the threshold are

---

4The confidence interval is constructed with a $\chi^2$ approximation and, if anything, may be conservative for the small samples at each classification score.
Treatment effects with exact matches at each score

![Figure 7](image)

**Figure 7**
Treatment effects for each classification score where there are both treatment and control units. The gray dots represent the difference in proportions and are proportional to the minimum number of treated or control units at that score. The hollow green circles are proportional to the number of treated units at that score. The vertical gray lines represent 95% confidence intervals. The horizontal light gray band represents the 95% confidence interval of the average treatment effect on the treated, which includes the origin. The vertical gray dashed line represents the treatment threshold.

plausible comparison groups. Outcome distributions that differ sharply can be attributed to the treatment.

How credible is the discontinuity assumption with the prison data? Prison assignment sharply changes when the classification score reaches 52. The top panel of Figure 9 (discussed more at length below) shows that at the threshold of 52, the probability of assignment to maximum-security prison jumps from 0.2 to 0.9. Do CDC officials manipulate the intake process to target prisoners based on their potential outcomes? Qualitative assessments of both the intake scoring method and the administrative placements are crucial here. For example, 16 points are added to the score if a prison assault "caused serious injury . . . the extent of which are [sic] life threatening in nature and require hospital care or cause disability over an extended period (medical attention beyond first-aid or . . . treatment and release)" (CDC 2000, art. 1, ch. 1, § 61010.11.2). To what degree does this standard permit subjective scoring to place individuals just above or below the threshold based on expected behavior? Similarly, administrative placements may be used to target placement when the score belies expected behavior. Not only are there subjective special case factors (e.g., whether the inmate "has strong family ties to a particular area where other placement would cause an unusual hardship"), but the criteria themselves suggest optimization on potential outcomes (i.e., whether the inmate’s “behavior record indicates he or she is capable of successful placement at an institution level lower than that indicated [by the] score”) (CDC 2000, art. 1, ch. 1, § 61010.11.3). If so, this precise manipulation invalidates regression discontinuity.

It is important to note here the substantive difference in the identification assumptions between matching and RD. Matching essentially identifies effects using administrative placements. If the classification score were
Classification and three strikes

Inmate classification score

Proportion with three strikes

0 0 4 0 6 0 8 0

0.0

0.4

0.8

Testing for difference around threshold

Bandwidth around threshold

Difference

–0.4

0.4

95% confidence interval

–40 –30 –20 –10 0 10 20

Figure 8

Covariate balance of whether inmate is a three-strike inmate. Panel (a) presents the classification score on the x-axis and the proportion of inmates with three strikes on the y-axis. Panel (b) plots the confidence interval of the difference in proportions by varying the bandwidth around the threshold of a score of 52 points (vertical gray dashed line). The larger the bandwidth, the sharper the discontinuity of proportion of three-strike inmates. The vertical light gray bands represent the bandwidth range for which there is relative balance of the third strike covariate.

Conventional practice, in that sense, has the process reversed. Research design (and covariate balance) should be implemented before any analysis. The primary goal in design is to determine the bandwidth around the threshold that results in comparable groups, without resorting to outcomes (cf. Lee & Lemieux 2010).

To illustrate the design phase, Figure 8a shows that the proportion of three-strike inmates increases with the classification score. An increase in the score from 40 to 50 is associated on average with a roughly 12% increase in the probability of a third strike. Moreover, there is a sharp spike from 33% to 72% third strikes at scores 54 to 55. Designs that fail to account for this discontinuity may falsely attribute differences in misconduct to security facility.

Assuming that the treatment cannot be manipulated precisely, conventional RD practice might be to fit numerous regressions to the outcome data (varying polynomial terms, the covariate set, and bandwidths). This ignores two crucial issues. First, it ignores covariate balance. If the design is right, balance should be verified (and the appropriate bandwidth chosen) prior to examination of any outcome data. Second, it imports all the problems of model sensitivity before implementing research design (Rubin 1977).
Figure 9

Treatment discontinuity and outcome continuity. Panel (a) plots the probability of maximum-security confinement (treatment) on the y-axis against inmate classification score on the x-axis. Short vertical lines represent each data point (randomly jittered for visibility). Panel (b) presents the probability of misconduct (outcome) on the y-axis against inmate classification score on the x-axis. Panel (c) plots the 95% confidence interval of the treatment effect (more precisely, the “intention to treat” effect), varying the bandwidth around the threshold of 52 points.

32 Ho • Rubin

Access provided by Stanford University - Main Campus - Robert Crown Law Library on 12/01/17. For personal use only.
size. The vertical gray bands choose one plausible bandwidth from scores of 49 to 56, where the confidence interval includes the origin. In practice, the design phase should develop this bandwidth by examining balance across all important covariates using substantive knowledge.

Having achieved balance on the key covariates, we may examine the outcome data. Our analysis is straightforward. Figure 9b plots the score against the proportion of behavioral incidents. Contrast the discontinuity of the treatment with the continuity of the outcome at the threshold. A sharp jump in the outcome would have been evidence of a treatment effect, but no perceptible change occurs at the threshold. Based on this estimate, the causal effect is indistinguishable from zero, with a wide 95% interval from −17% to 27%.

The vertical gray bands overlay the bandwidth chosen based on the third strike covariate. Within the bandwidth, a (local logistic) regression can be used to adjust for remaining imbalance. The Figure 9b inset magnifies the bandwidth range and overlays pointwise 95% confidence bands, showing that there is little evidence of any treatment effect. Based on this regression adjustment, the overall 95% interval of the causal effect contains the origin, ranging from −10% to 29%.

9. COMPARISON TO EXPERIMENTAL RESULTS

How do these methods compare to a randomized experiment? For most social science applications, few such validations exist (but cf. Dehejia & Wahba 1999; Lalonde 1986; Heckman et al. 1998a,b). Fortunately, Berk et al. (2003) performed a valuable field experiment in cooperation with the CDC, allowing for potential validation of observational approaches. Berk et al. (2003) randomized inmates to the existing intake procedure and to a new proposed one, in which the latter increased the security level for some inmates. This randomization thus allows us to identify the causal effect of security level on the subgroup of inmates whose assignment was affected by the difference in protocols.

To be sure, the field experiment was limited in certain respects. First, the randomization was over the intake procedure, not the treatment of security level. Second, the experimental intake procedure generally increased security from low to medium levels, with little effect on maximum-security confinement. The experiment is hence uninformative about the effect of maximum security per se. Therefore, we focus on (a) the overall (so-called “intention-to-treat”) effect of the experimental intake procedure, essentially comparing misconduct rates between all inmates scored on old and new intake procedures, and (b) the subgroup effect on inmates whose assignment was in fact affected by randomization (Camp & Gaes 2005). These effects may, of course, diverge from the population treatment effect.6

Figure 10 presents point estimates and 95% confidence intervals from the various observational approaches and the experiment. The first line presents naive (logit) regression estimates from BdL and Section 5 in green. The second line presents estimates from exact matching in Section 7. The third line presents estimates from regression discontinuity, either using the simple difference in means within the bandwidth or a model-based adjustment within the bandwidth from Section 8. The overall experimental estimate in the fifth line is −1%, plus or minus 1%, and the subgroup effect (of facilities with individual cells versus open dormitories) in the sixth line is −4%, plus or minus 8%, both

6If there are heterogeneous treatment effects, then matching, regression discontinuity, and the experiment may properly identify effects for the subsets of (a) inmates affected by administrative placements and population overrides, (b) inmates just above and below the classification score of 52, and (c) inmates for whom the experimental procedure changed ultimate placement, respectively, but these effects could nonetheless diverge.
Effect

1. Naive logit
2. Exact matching
3. Regression discontinuity (means within bandwidth)
4. Regression discontinuity (logit within bandwidth)
5. Experimental finding (overall)
6. Experimental finding (subgroup, all misconduct)
7. Experimental finding (subgroup, serious misconduct)

Figure 10
Comparison of observational approaches to experimental findings. Dots plot point estimates (weighted by effective sample size), and lines represent 95% confidence intervals. The naive logit (green) is the overall average treatment effect (based on asymptotic posterior simulation) of the regression reported in Berk & de Leeuw (1999, p. 1048, table 1, model 1) and Section 5. Exact matching is the average treatment effect on the treated, based on the weighted difference of exact matches on the classification score. Regression discontinuity is the intention-to-treat effect, either (a) the mean difference between subjects above and below the threshold within the bandwidth or (b) the treatment effect (based on asymptotic posterior simulation) of the logit regression within the bandwidth. The overall experimental findings are calculated based on sample sizes and effects from Berk et al. (2003, p. 228, table 1, p. 232). The subgroup experimental findings are from Camp & Gaes (2003, pp. 434, 436, tables 1 and 2 therein).

Indistinguishable from 0. The last line presents experimental subgroup estimates of an increase of 3%, plus or minus 8%, on serious misconduct. Naive regression-based estimates, finding a reduction of 13% (plus or minus 4%), deviate considerably from experimental estimates.7 Intervals from matching, regression discontinuity, and the experiment, on the other hand, all contain the origin.

As a last comparison, Bench & Allen (2003) randomized 200 inmates to maximum- or medium-security prisons in Utah. Although the measurement of misbehavior differs, the study is perhaps closest to the ideal experiment. It found that “there is no meaningful difference in the number of...disciplinaries” between treated and control groups (Bench & Allen 2003, p. 377).

In the end, neither the California nor the Utah study is a gold-standard experiment. In Utah, 10% of the inmates changed security assignments midstream. In California, randomization was not over the treatment of interest.

The distinction between experiments and observational studies is one of degree, not of kind. A well-designed observational study, in which fluctuations in the availability of prison beds, for example, affect inmate placement, can be more informative than a broken experiment. The comparisons of Figure 10 provide considerable evidence that quasi-experimental approaches, by reducing the role of unwarranted functional form assumptions, are more likely to recover the true causal effect.

10. CONCLUSION: A RETURN TO MOORE

Causal inference is hard. As we have reviewed, recent advances should allow researchers to more credibly assess the impact of legal institutions. Once designs are stripped of technical garb, research should empower the broader legal academic community—precisely the community with the comparative advantage—to assess the credibility of the inference.

We reiterate the important lessons, if in pithy format:
1. Conceptualize the experimental template.
2. Design research with outcomes last.
3. Collect and balance covariates.
4. Visualize the data.

Although formalization of the approaches we have discussed is relatively recent, in one way the emphasis on research design calls for a return to the early legal empiricist Underhill Moore. True, his parking studies had no standard errors, failed to assess pre- and post-time trends around the threshold, and did not employ conventional models for analysis. But these are second order. Moore got design. Credible design occurs prior to outcomes and, as the BdL data show, can require deep substantive knowledge of the law itself. In that sense, Moore’s law of parking was ahead of the curve, if not the curb.

APPENDIX: WHERE TO GO FROM HERE

Our exposition here is only an informal review of a vast, rapidly growing, and sophisticated literature. In this Appendix, we provide some guidance on where researchers can go to study these approaches in greater depth.


For developments of sensitivity and bounds analyses, see Manski (1990, 1995) and Rosenbaum & Rubin (1983a). For randomization inference, see Imbens & Rosenbaum (2005), Ho & Imai (2006), and Donohue & Ho (2007).

There are, of course, many examples of panel approaches to assessing the impact of law. For examples of difference-in-differences, see Card & Krueger (1994), Autor et al. (2004), and Rubinfeld (2010). Bertrand et al. (2004) make a crucial point about variance estimation in assessing one-time policy interventions.


DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Shira Oyserman, Xiangnong Wang, and Olga Zverovich for terrific research assistance, Gail Long at the California Department of Corrections and Rehabilitation for help with the
history of inmate classification, George Wilson at the Stanford Law Library for help in tracking
down CDC operations manuals, Stephen Galoob for comments, and Richard Berk and Jan de
Leeuw for sharing data.

LITERATURE CITED
http://www.economics.harvard.edu/faculty/imbens/software_imbens
Rev. 93:113–32
Abrams DS, Yoon AH. 2007. The luck of the draw: using random case assignment to investigate attorney
ability. Univ. Chicago Law Rev. 74:1145–77
Angrist JD. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security adminis-
trative records. Am. Econ. Rev. 80:1284–86
Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables (with
Angrist JD, Lavy V. 1999. Using Maimonides’ rule to estimate the effect of class size on scholastic achievement.
Q. J. Econ. 114:533–75
Princeton Univ. Press
Angrist JD, Pischke J-S. 2010. The credibility revolution in empirical economics: how better research design
is taking the con out of econometrics. J. Econ. Perspect. 24:3–30
Autor DH, Donohue JJ III, Schwab SJ. 2004. The costs of wrongful-discharge laws: large, small, or none at
all? Am. Econ. Rev. 94:440–46
104:817–72
98:299–311
rections. Q. J. Econ. 124:105–47
Software. http://www.lrz.de/~sobecker/pscore.html
83:367–82
Berk RA, de Leeuw J. 1999. An evaluation of California’s inmate classification system using a generalized
Criminol. Public Policy 2:215–42
Berk RA, Newton PJ. 1985. Does arrest really deter wife battery? An effort to replicate the findings of the
Minneapolis Spouse Abuse Experiment. Am. Sociol. Rev. 50:253–62
Pap. Ser. 07.04, Univ. Chicago, Chicago, IL
Bertrand M, Duflo E, Mullainathan S. 2004. How much should we trust differences-in-differences estimates?
Q. J. Econ. 119:249–75


Contents

The Legislative Dismantling of a Colonial and an Apartheid State
Sally Falk Moore ................................................................. 1

Credible Causal Inference for Empirical Legal Studies
Daniel E. Ho and Donald B. Rubin ........................................ 17

Race and Inequality in the War on Drugs
Doris Marie Provine ............................................................. 41

Robert J. MacCoun and Peter Reuter ..................................... 61

The Triumph and Tragedy of Tobacco Control:
A Tale of Nine Nations
Eric A. Feldman and Ronald Bayer ....................................... 79

Privatization and Accountability
Laura A. Dickinson ............................................................. 101

The Conundrum of Financial Regulation: Origins, Controversies,
and Prospects
Laureen Snider ................................................................. 121

Corporate and Personal Bankruptcy Law
Michelle J. White ............................................................... 139

Durkheim and Law: Divided Readings over Division of Labor
Carol J. Greenhouse ............................................................ 165

Law and American Political Development
Pamela Brandwein .............................................................. 187

The Legal Complex
Lucien Karpik and Terence C. Halliday ................................. 217

U.S. War and Emergency Powers: The Virtues
of Constitutional Ambiguity
Gordon Silverstein ............................................................. 237

The Political Science of Federalism
Jenna Bednar ................................................................. 269
The Rights of Noncitizens in the United States

*Susan Bibler Coutin* .................................................. 289

Innovations in Policing: Meanings, Structures, and Processes

*James J. Willis and Stephen D. Mastrofski* ........................................ 309

Elaborating the Individual Difference Component
in Deterrence Theory

*Alex R. Piquero, Raymond Paternoster, Greg Pogarsky, and Thomas Laughran* ........ 335

Why Pirates Are Back

*Shannon Lee Dawdy* .......................................................... 361

The Evolving International Judiciary

*Karen J. Alter* ......................................................................... 387

The Social Construction of Law: The European Court of Justice
and Its Legal Revolution Revisited

*Antonin Cohen and Antoine Vauchez* .................................................. 417

Indexes

Cumulative Index of Contributing Authors, Volumes 1–7 .................. 433
Cumulative Index of Chapter Titles, Volumes 1–7 .............................. 436

Errata

An online log of corrections to *Annual Review of Law and Social Science* articles may be found at http://lawsocsci.annualreviews.org