Observational Studies 2 (2016) 174-182 Submitted 10/16; Published 12/16

Understanding Regression Discontinuity Designs As Observational Studies

Jasjeet S. Sekhon **sekhon@berkeley.edu**
Robson Professor
Departments of Political Science and Statistics
UC-Berkeley
210 Barrows Hall #1950, Berkeley, CA 94720-1950

Rocío Titiunik **titiunik@umich.edu**
James Orin Murfin Associate Professor
Department of Political Science
University of Michigan
505 South State St., 5700 Haven Hall,Ann Arbor, MI 48109-1045

Keywords: Regression Discontinuity, Local Randomization, Local Experiment

1. Introduction

Thistlethwaite and Campbell (1960) proposed to use a "regression-discontinuity analysis" in settings where exposure to a treatment or intervention is determined by an observable score and a fixed cutoff. The type of setting they described, now widely known as the regression discontinuity (RD) design, is one where units receive a score, and a binary treatment is assigned according to a very specific rule. In the simplest case, all units whose score is above a known cutoff are assigned to the treatment condition, and all units whose score is below the cutoff are assigned to the control (i.e., absence of treatment) condition. Thistlethwaite and Campbell insightfully noted that, under appropriate assumptions, the discontinuity in the probability of treatment status induced by such an assignment rule could be leveraged to learn about the effect of the treatment at the cutoff. Their seminal contribution led to what is now one of the most rigorous non-experimental research designs across the social and biomedical sciences. See Cook (2008), Imbens and Lemieux (2008) and Lee and Lemieux (2010) for reviews, and the recent volume edited by Cattaneo and Escanciano (2017) for recent specific applications and methodological developments.

A common and intuitive interpretation of RD designs is that the discontinuous treatment assignment rule induces variation in treatment status that is "as good as" randomized near the cutoff, because treated and control units are expected to be approximately comparable in a small neighborhood around the cutoff (Lee, 2008; Lee and Lemieux, 2010). This local randomization interpretation has been extremely influential, and many consider RD designs to be almost as credible as experiments. Although the formal analogy between RD designs and experiments was discussed recently by Lee (2008), the idea that the RD design behaves like an experiment was originally introduced by Thistlethwaite and Campbell, who called a hypothetical experiment where the treatment is randomly assigned near the cutoff an “experiment for which the regression-discontinuity analysis may be regarded as a substitute” (Thistlethwaite and Campbell, 1960, p. 310). Building on this analogy, Lee (2008) formalized the idea in a continuity-based framework; in addition, Cattaneo et al. (2015) formalized this idea in a Fisherian finite-sample framework. See Cattaneo et al. (2017) and Sekhon and Titiunik (2017) for recent discussions on the connections between both frameworks.

The analogy between RD designs and experiments has been useful in communicating the superior credibility of RD relative to other observational designs, and has focused attention on the need to perform falsification tests akin to those usually used in true experiments. All these developments have contributed to the RD design’s rigor and popularity. Despite these benefits, we believe the analogy between RD designs and experiments is imperfect, and we offer a more cautious interpretation in which the credibility of RD designs ranks decidedly below that of actual experiments.

In our view, RD designs are best conceived as non-experimental designs or *observational studies*—i.e., studies where the goal is to learn about the causal effects of a treatment, but the similarity or comparability of subjects receiving different treatments cannot be ensured by construction. Interpreting RD designs as observational studies implies that their credibility must necessarily rank below that of experiments. This, however, does not mean that RD designs are without special merit. Among observational studies, RD designs are one of the most credible alternatives because important features of the treatment assignment mechanism are known and empirically testable under reasonable assumptions.

We justify our view by focusing on three main issues. First, we consider the RD treatment assignment rule, and show that it contains considerably less information than the analogous rule in an experimental assignment. Second, we consider the special role of the score or running variable, in particular the possibility that the score may affect the outcome via post-treatment channels and violate an exclusion restriction that holds by construction in experiments. Finally, we highlight that in order to obtain meaningful conclusions from testing the “empirical implications” of a valid RD design, further assumptions must be made about the data generating process. All these issues support our view that RD designs are observational studies. We do not mean these arguments as a critique of RD designs. Our point is simply that a compelling observational study faces hurdles that are absent in experimental designs, and therefore the analysis and interpretation of RD designs should be done with the same caution as in any other observational study.

2. The RD Treatment Assignment Rule

The fundamental feature of RD designs is that the treatment is assigned based on a known rule. In the so-called sharp RD design where compliance with treatment is perfect, treatment status is deterministic given the score: all units with score below the cutoff are assigned to and receive the control condition, and all units with score above the cutoff are assigned to and receive the treatment condition. Moreover, in the standard RD setup, the cutoff is known. This can be formalized in the rule $T_i = \mathbb{1}\{X_i \geq c\}$, where $i = 1, 2, \ldots n$ indexes the units in the study, T_i is the treatment status, c is the cutoff, and X_i is the score or running variable. Because this rule is at the heart of every RD design,[1] any researcher working with an RD design has rich information about the treatment assignment mechanism.

At first glance, the fact that treatment assignment is based on a known rule might suggest that RD designs are not observational studies. As commonly defined (e.g. Rosenbaum, 2002), a key feature of an observational study is that the treatment assignment mechanism is not under the control of the researcher (or someone else the researcher has access to), which implies that it is fundamentally unknown. For example, an observational study of the effects of smoking on lung cancer may compare smokers and non-smokers and obtain valid inferences under some assumptions, but the probability of smoking always remains unknown.

RD designs are different in this regard because, although the actual assignment of treatment is rarely under the direct control of the investigator, the probability of receiving treatment given the score is known for every unit. In other words, if a unit receives a particular score value, in a sharp RD design we know with certainty whether the probability of receiving treatment was one or zero. Although this has many advantages, it is not enough to lift the status of RD from observational studies to experimental designs. The reason is that the distribution of the score remains fundamentally unknown: although we know that $T_i = 1$ if the score X_i is above the cutoff and $T_i = 0$ otherwise, we know nothing about how the value of X_i was determined. Thus, despite the treatment assignment rule being known, the comparability of treated and subjects is not ensured.

This fundamental lack of knowledge about the distribution of the score makes the RD design inherently different from experiments. In an experiment, units are randomly assigned to treatment or control, which implies that the distribution of all predetermined characteristics and unobserved confounders is identical in the treatment and control groups, ensuring their comparability. In the language of the potential outcomes framework, random assignment of treatment ensures independence between treatment status and potential outcomes. In the absence of complications (such as interference across units and compliance issues), this independence is sufficient to guarantee identification of the (sample) average treatment effect.

In contrast, in RD designs, the treatment assignment rule $T_i = \mathbb{1}\{X_i \geq c\}$ is not enough to ensure the identification of the treatment effect (at the cutoff). This is a direct consequence of the fact that the assignment rule determines T_i, but it does not determine X_i. For example, as shown by Hahn et al. (2001), the main condition to obtain identification of the average treatment effect at the cutoff in a sharp RD design is the continuity of the regression functions of the potential outcomes at the cutoff. Letting Y_{1i} and Y_{0i} denote the potential outcomes under treatment and control for unit i, defining the observed outcome as $Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$, and assuming the observed data $\{Y_i, X_i\}_{i=1}^{n}$ is a random sample from a larger population, the continuity condition says that $\mathbb{E}[Y_{1i}|X = x]$ and $\mathbb{E}[Y_{0i}|X_i = x]$, seen as functions of x, are continuous in x at c.

Crucially, the continuity of the potential-outcome regression functions at the cutoff is not implied or guaranteed by the known and deterministic RD treatment assignment rule; it is an assumption that must be imposed. In other words, the fact that the treatment is assigned according to the rule $T_i = \mathbb{1}\{X_i \geq c\}$ places no restrictions on the proper-

[1] In a fuzzy RD design compliance with the assignment is no longer perfect; in this case, the rule $T_i = \mathbb{1}\{X_i \geq c\}$ still applies, but T_i now refers to treatment assignment instead of treatment status.

ties of functions such as $\mathbb{E}[Y_{1i}|X = x]$ and $\mathbb{E}[Y_{1i}|X_i = x]$. In contrast, the unconfounded random treatment assignment rule in classical experiments guarantees a statistical independence assumption (or a known randomization distribution assumption in finite-sample settings). This fundamental asymmetry between the credibility of identification conditions in experiments versus RD designs—in the former guaranteed by construction, in the latter by assumption—is one of the reasons why the RD should be considered an observational design.

Randomized experiments do need additional assumptions for parameter estimation and hypothesis testing in many cases. Depending on the parameter or hypothesis of interest and the statistic used, researchers usually need to impose additional regularity conditions, in addition to modeling the sampling structure of the data. For example, in the case of the average treatment effect, these regularity conditions, aside from non-interference, include moment conditions on the outcomes (and covariates)—see, e.g., Lin (2013). Such conditions will typically be weaker than the assumptions required for estimation in the continuity-based RD case, where smoothness conditions are required in addition to the continuity assumption (Calonico et al., 2014), neither of which is guaranteed by the design. We also note that in the case of randomized experiments, both parameter estimation standard hypothesis testing can be skipped in favor of permutation tests of the Fisherian sharp null, which require even weaker assumptions (Rosenbaum, 2002).

3. The Intermediate Role of the Running Variable

The existence of the running variable—and our fundamental lack of knowledge about its distribution and determinants—poses another challenge for the analogy between experiments and RD designs, and gives another reason to classify the latter as an observational design. In a nutshell, the source of this second challenge is that the RD running variable is often a very important determinant of the potential outcomes—not only because it may correlate with predetermined characteristics that are related to the outcome, but also because it can have a "direct" or "post-treatment" effect on the potential outcomes. As we discuss in detail in Sekhon and Titiunik (2017), the special status of the RD score breaks the usual connection between the concepts of random assignment, statistical independence, and constant or "flat" regression functions that are taken for granted in experiments. This exclusion restriction was first noted by Cattaneo et al. (2015) in a Fisherian framework, and is relaxed under additional assumptions in Cattaneo et al. (2017).

One intuitive way to motivate the RD-experiment analogy is that a randomized experiment can be understood as particular case of the RD design where the score is a (pseudo) random number, and the cutoff is chosen to ensure the desired probability of treatment assignment. For example, one can randomly assign a treatment among a group of subjects with probability 50% by assigning a uniform random number between 1 and 100 to each subject, and then assigning the treatment only to those subjects whose assigned number exceeds 50. This randomized experiment can be easily recast as a sharp RD design where the uniform random number is the score and the cutoff is 50.

This hypothetical experiment recast as an RD design has two crucial features:

(i) By virtue of random assignment, the score is statistically independent of all predetermined covariates, including all those covariates that affect or are related to the potential outcomes;

(ii) By virtue of the score being an arbitrary number generated solely for the purpose of assigning the treatment, there can be no "post-treatment" effect of the score on the potential outcomes except via the treatment assignment indicator.

The combination of (i) and (ii) implies, for example, that the regression functions $\mathbb{E}[Y_{0i}|X = x]$ and $\mathbb{E}[Y_{1i}|X_i = x]$ are constant in the entire support of the score.

The RD design, in practice, does not generally satisfy either of these conditions. In typical RD treatment assignment rules, the score or running variable is a crucial determinant of the potential outcomes. For example, a party may win an election when its vote share exceeds 50%, and we may be interested in the effect of winning on future victories. Or a municipality may receive federal assistance when its poverty index is below a certain threshold, and we may be interested in the effect of federal assistance on mortality. In such cases, the score is fundamentally related to both predetermined characteristics of the units that may be strongly related to the outcome (e.g., municipalities with high poverty index may also have high unemployment which can affect mortality via lower health insurance coverage), and it can also affect the outcome directly (e.g., increased poverty may reduce access to potable water and increase disease and mortality risk). Both possibilities make the analogy between experiments and RD designs imperfect.

This challenge can be further illustrated by noting that even if we assume that the score is randomly assigned among subjects, the score—and, consequently, the treatment assignment, may fail to be independent of the potential outcomes. The reason is simply that, although the random assignment of the score ensures condition (i), it fails to ensure condition (ii). A randomly assigned score is by construction independent of all predetermined covariates, but it nonetheless may have an effect on the outcome that occurs not via correlation with predetermined characteristics, but via a post-treatment channel. This implies that the random assignment of the score is not enough to guarantee the exclusion restriction that the score affects the potential outcomes only through the treatment assignment indicator.

To understand why this occurs, note that in a true experiment the exclusion restriction holds by construction because the pseudo-random number assigned to each subject plays no role in the data generating process of the potential outcomes. Importantly, the exclusion restriction holds in a true experiment not because of the random assignment per se, but because the score used to implement the randomization procedure is arbitrary (indeed, in most real experiments, this "score" is entirely unknown to the experimental subjects). This is why in a RD design, where the score may often affect the outcome by various post-treatment channels, the random assignment of the score does not—and cannot—guarantee condition (ii).

This brief discussion shows that assuming random assignment of the RD score in a neighborhood near the cutoff does not imply that the potential outcomes and the treatment are statistically independent, or that the potential outcomes are unrelated to the score in this neighborhood. Furthermore, as we show formally in Sekhon and Titiunik (2017), the assumption of local independence between the potential outcomes and the treatment assignment does not imply the exclusion restriction that the score affects the outcome only via the treatment indicator but not directly.

In sum, the RD treatment assignment rule does not by itself place any restrictions on the ways in which the score can influence the potential outcomes—and even in a locally random RD design where the score is randomly assigned near the cutoff, the statistical independence between potential outcomes and treatment assignment that we take for granted in experiments need not follow. This is another reason why we view RD designs as observational studies.

4. The RD Assumptions and Their Empirical Implications

Lee (2008) heuristically argued that a consequence of interpreting RD designs as local experiments is that predetermined covariates in treated and control groups should be similar in a neighborhood of the cutoff. Formally, Lee established continuity of the distribution of observed predetermined covariates at the cutoff. As a consequence, he proposed to test whether the treatment has an effect on predetermined covariates at the cutoff to falsify the RD assumptions—similarly to the way in which balance tests are used in experiments to evaluate whether the randomization was performed correctly. This emphasis on the need to test empirically the comparability of treatment and control groups has been a positive and influential development in the RD literature. By now, falsification tests are a standard part of most empirical RD applications (see, e.g., Caughey and Sekhon, 2011; de la Cuesta and Imai, 2016; Eggers et al., 2015).

Under the assumption of continuity of the potential-outcome regression functions, these "covariate balance" tests should be implemented treating each covariate as an outcome in the RD analysis—that is, estimating average RD treatment effects on the covariates in the same way as these effects are estimated for the true outcome of interest. The standard implementation of continuity-based RD estimation and inference uses local polynomial methods, fitting a weighted polynomial of the outcome/covariate on the score within an optimally chosen bandwidth around the cutoff (see, e.g., Calonico et al., 2014, 2016, and references therein). This implementation allows all predetermined covariates to be arbitrarily related to the score variable, and looks for an effect at the cutoff. Since the covariates are determined before treatment is assigned, researchers are reassured when such RD effects on the covariates cannot be distinguished from zero.

The use of these "covariate balance" tests for falsification is perhaps the most salient practical similarity between RD analysis and experimental analysis. The assumption behind the RD falsification tests on covariates is that continuity of the covariate regression functions implies or at least supports the assumption that the potential-outcome regression functions are continuous. This is a strong requirement because, as with continuity of the potential-outcome regression functions, continuity of the covariate regression functions is not implied by the RD treatment assignment rule. Moreover, continuity of the covariate regression functions is neither necessary nor sufficient for the potential-outcome regression functions to be continuous. Thus, falsification tests based on covariates require assumptions that are not true by construction. Similarly, falsification tests based on the density of the running variable (McCrary, 2008) require that such density be continuous at the cutoff, another condition that is neither necessary nor sufficient for the main RD identification assumption to hold.

It follows that falsification analysis in RD designs is more demanding than in experimental settings. In the case of actual experiments, we know that if the random assignment of the treatment was implemented without errors, the treatment assignment will be independent of all predetermined covariates (as well as of potential outcomes). Thus, the design itself implies that the distribution of predetermined covariates in treatment and control groups is the same, and falsification tests try to corroborate the empirical implication of a balance condition we know to be true. In contrast, in RD designs, neither the identification assumptions on the potential outcomes nor the falsification assumptions on the covariates are known to be true, because these assumptions are not implied by the treatment assignment rule. Precisely for this reason, falsification analysis plays a more crucial role in RD designs than in experiments, as researchers are eager to provide empirical evidence that the invoked RD assumptions are plausible. The paradox is that falsification tests are most needed in those settings where they require more assumptions to be informative. The bottom line is that identification assumptions are a prerequisite for the data to be informative about the parameters of interest, and we cannot use the data to test the assumptions that make the data meaningful in the first place. In general, nonparametric identification assumptions are fundamentally untestable.

This, of course, does not mean that RD falsification tests are not useful. In most applications, it is entirely reasonable to assume that if the potential-outcome regression functions are continuous at the cutoff, most predetermined covariates that are related to the outcome will also have continuous regression functions. This assumption will be particularly plausible for certain covariates, such as the outcome measured before treatment assignment and other variables that are known to be strongly related to the outcome of interest. Our point is simply that this is an assumption that must be made, in contrast to a feature that is true by design.

5. Conclusion

In sum, we believe the RD design is an observational study, and should be interpreted as such. Despite the usefulness of the analogy between RD designs and experiments, RD designs lack the credibility of experiments for the simple reason that the treatment assignment rule does not guarantee the assumptions that are needed for identification of the treatment effects of interest. In particular, the RD assignment rule implies neither continuity of the relevant potential-outcome functions nor local independence between the potential outcomes and the treatment assignment; and the random assignment of the score near the cutoff does not imply local independence between the potential outcomes and the score or treatment assignment. Moreover, falsification tests in RD designs require additional assumptions about the relationship between the selected predetermined covariates and the potential outcomes.

Acknowledgments

We are indebted to Matias Cattaneo, Kosuke Imai, Max Farrell, Joshua Kalla, Kellie Ottoboni, and Fredrik Sävje for valuable comments and discussion. Sekhon gratefully acknowledges support from the Office of Naval Research (N00014-15-1-2367) and Titiunik gratefully acknowledges financial support from the National Science Foundation (SES 1357561).

References

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2016). Regression discontinuity designs using covariates. Working paper, University of Michigan.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cattaneo, M. D. and Escanciano, J. C. (2017). *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38).* Emerald Group Publishing, forthcoming.

Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management,* forthcoming.

Caughey, D. and Sekhon, J. S. (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. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654.

de la Cuesta, B. and Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19:375–396.

Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., and Snyder, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, 59(1):259–274.

Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Imbens, G. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Rosenbaum, P. R. (2002). *Observational Studies.* Springer, New York, NY, second edition.

Sekhon, J. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In Cattaneo, M. D. and Escanciano, J. C., editors, Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38). Emerald Group Publishing, forthcoming.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.