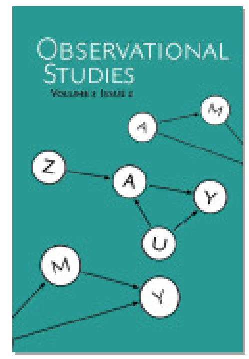
Regression Discontinuity Designs in the Econometrics Literature

Guido W. Imbens



Observational Studies, Volume 3, Issue 2, 2017, pp. 147-155 (Article)

Published by University of Pennsylvania Press
DOI: https://doi.org/10.1353/obs.2017.0003

Observational Studies 2 (2016) 147-155 Submitted 11/16; Published 12/16

Regression Discontinuity Designs in the Econometrics Literature

Guido W. Imbens **imbens@stanford.edu**
Graduate School of Business, Stanford University, SIEPR, and NBER
Stanford, CA, U.S.A.

Abstract

Many decades after being introduced by Thistlewaite and Campbell (1960), regression discontinuity designs have become an important tool for causal inference in social sciences. Researchers have found the methods to be widely applicable in settings where eligibility or incentives for participation in programs is at least partially regulated. Alongside, and motivated by, the many studies applying regression discontinuity methods there have been a number of methodological studies improving our understanding, and implementation, of, these methods. Here I report on some of the recent advances in the econometrics literature.

Keywords: regression discontinuity designs, matching, regression kink designs, local linear regression

1. Introduction

In the late 1990s and early 2000s, regression discontinuity designs (rdd's for short), originally developed many years earlier by Thistlewaite and Campbell (1960), enjoyed a renaissance in social science in general, and in economics in particular. As the rdd method became one of the most popular strategies for identifying causal effects (Angrist and Pischke (2008)) and a standard topic in first year econometrics courses in PhD programs, researchers became increasingly aware of the wide applicability of the methods developed by Thistlewaite and Campbell. Early applications in economics include Black (1999), using geographical boundaries, Van Der Klaauw (2002), using college application thresholds, and Lee (2008), using election thresholds. See Cook (2008) for a historical perspective, including references to earlier discussions in economics that failed to catch on, and for recent general discussions and surveys in the economics literature see Imbens and Lemieux (2008); Van Der Klaauw (2008); Lee and Lemieux (2010); Calonico et al. (2015); Choi and Lee (2016). For general discussions outside of economics see Trochim (1984); Shadish et al. (2002); Skovron and Titiunik (2015). The recent increase in applications in economics has motivated novel theoretical work on rdd methods in econometrics that have improved our understanding of rdd methods, as well as affected empirical practice. Here I want to discuss some of these recent methodological innovations.

2. Basic Set Up

As is common in the econometric causal literature, though not originally in the rdd literature, we set the problem up in the Rubin Causal Model or potential outcome framework

(Rubin (1974); Holland (1986); Imbens and Rubin (2015)). We assume there are, for each unit in a large population, two potential outcomes, $Y_i(0)$ and $Y_i(1)$ for unit i, corresponding to the control and treated outcome, with the unit-level causal effect some comparison of the two, e.g., the difference $Y_i(1) - Y_i(0)$. There is a binary treatment $W_i \in \{0, 1\}$, defining the observed outcome $Y_i^{\text{obs}} = Y_i(W_i)$, and an exogenous forcing variable X_i, as well as possibly additional covariates Z_i. At the threshold, say $x = 0$, the probability of receiving the treatment changes discontinuously. If it changes from zero to one we have a sharp rd design, otherwise a fuzzy rd design. In general the estimand in rdd analyses is the ratio of two discontinuities, first, the discontinuity in the conditional expectation of the realized outcome given the forcing variable, and second, the treatment indicator given the forcing variable, both at the threshold:

$$\tau = \frac{\lim_{x\downarrow 0} \mathbb{E}[Y_i^{\text{obs}}|X_i = x] - \lim_{x\uparrow 0} \mathbb{E}[Y_i^{\text{obs}}|X_i = x]}{\lim_{x\downarrow 0} \mathbb{E}[W_i|X_i = x] - \lim_{x\uparrow 0} \mathbb{E}[W_i|X_i = x]}.$$

In the sharp rdd setting the denominator is excactly one and we simply look at the magnitude of the discontinuity in the conditional expectation of the outcome at the threshold.

3. The Interpretation of Fuzzy Regression Discontinuity Designs

The first, and arguably most important, innovation in the econometrics literature concerns the precise interpretation of the estimand in fuzzy rd designs in settings with heterogenous treatment effects. Although fuzzy regression discontinuity designs had been around at least since Trochim (1984), their analysis was limited to the case with constant treatment effects. Hahn et al. (2001) (HTV from hereon) established an important link to the instrumental variables literature with heterogenous treatment effects. In particular they showed that in the fuzzy rdd the ratio of discontinuities in the conditional mean for the outcome and the conditional mean for the treatment has an interpretation of a local average treatment effect (Imbens and Angrist (1994); Angrist et al. (1996)). The HTV argument shows that the fuzzy rdd estimand is the average effect of the treatment, for the subpopulation of compliers, among those with a value for the forcing variable close to the threshold. Compliers in this subpopulation of individuals with a value for the forcing variable close to the threshold are individuals for whom it matters which side of the threshold they are on. The argument is slightly subtle, because it relies on a clear interpretation of what is stochastic in this setting. Specifically, for an individual close to, but on the left of the threshold, say with $X_i \in (-\epsilon, 0)$, it requires one to think about what would have happened to this individual had they been on the other side of the threshold. We can do this in two ways. In the HTV approach, the forcing variable X_i is taken as potentially manipulable, so that one can think of potential treatment values $W_i^{\text{htv}}(x)$ for different values of the forcing variable, with $W_i = W_i^{\text{htv}}(X_i)$ the realized value. Compliers are in this approach individuals who would have received the treatment had X_i been slightly above the threshold, but not if X_i had been slightly below the threshold:

$$C_i^{\text{htv}} = \begin{cases} a & \text{if } \lim_{x\downarrow 0} W_i^{\text{htv}}(x) = 1,\ \lim_{x\uparrow 0} W_i^{\text{htv}}(x) = 1, \\ c & \text{if } \lim_{c\downarrow 0} W_i^{\text{htv}}(x) = 0,\ \lim_{c\uparrow 0} W_i^{\text{htv}}(x) = 1, \\ n & \text{if } \lim_{c\downarrow 0} W_i^{\text{htv}}(x) = 0,\ \lim_{c\uparrow 0} W_i^{\text{htv}}(x) = 0, \\ d & \text{if } \lim_{c\downarrow 0} W_i^{\text{htv}}(x) = 1,\ \lim_{c\uparrow 0} W_i^{\text{htv}}(x) = 0. \end{cases}$$

Typically the presence of defiers is ruled out. In this perspective the forcing variable is stochastic, rather than a fixed characteristic of the individual, and could have taken on a different value for a given individual from that observed one for that individual.

In some applications it may be difficult to imagine the forcing variable as a causal variable, defining potential outcomes, say in the case where the forcing variable is a fixed immutable characteristic such age. In such cases an alternative, following Bertanha and Imbens (2016), may be to view the threshold, rather than the forcing variable, as manipulable, generating potential treatment values corresponding to the threshold: $W_i^{\text{bi}}(c)$ is in this approach the treatment level for unit i if the threshold were set at c, where we only consider values of c close to actual threshold of zero. In this perspective compliers are defined by the pair of limits of $W_i^{\text{bi}}(c)$, taken from the left and from the right of the actual threshold value zero:

$$C_i^{\text{bi}} = \begin{cases} a & \text{if } \lim_{c\downarrow 0} W_i^{\text{bi}}(c) = 1,\ \lim_{c\uparrow 0} W_i^{\text{bi}}(c) = 1, \\ c & \text{if } \lim_{c\downarrow 0} W_i^{\text{bi}}(c) = 0,\ \lim_{c\uparrow 0} W_i^{\text{bi}}(c) = 1, \\ n & \text{if } \lim_{c\downarrow 0} W_i^{\text{bi}}(c) = 0,\ \lim_{c\uparrow 0} W_i^{\text{bi}}(c) = 0, \\ d & \text{if } \lim_{c\downarrow 0} W_i^{\text{bi}}(c) = 1,\ \lim_{c\uparrow 0} W_i^{\text{bi}}(c) = 0. \end{cases}$$

Again we typically rule out the presence of defiers.

This difference in interpretation has some conceptual implications. If one views the forcing variable as stochastic, it can be used to generate a randomization distribution for the regression discontinuity estimator with approximately independent treatment assignments, similar to that in a randomized experiment. Using only individuals close to the threshold, we have essentially a randomized experiment with assignment for all units close to independent. However, if we view the threshold as potentially manipulable, there is only a single stochastic component driving the randomization properties of the estimator, so that the treatment assigments are closely related, and the fundamental difference with an actual randomized experiment becomes clear.

4. Supplementary Analyses in Regression Discontinuity Designs

A second active area of methodological innovations has been the use of, what Athey and Imbens (2016) call in general discussion of causal inference, supplementary analyses. They define these as analyses where the aim is not to get a better estimate of the object of interest, that is the causal effect. Rather they are analyses that are intended to provide support for the main analyses, by disproving potential arguments against the validity of the main analyses. Depending on the results of the supplementary analyses the credibility of the main analyses is either weakened or strengthened.

One of the major concerns in rdd analyses is that the forcing variable may have been manipulated. In many cases there are substantial costs or benefits for the agents associated with being just to the left or right of the threshold associated with the change in incentives. If agents have some ability to change the actual, or even just the recorded, value of the forcing variable, they would in that case have a strong incentive to do so. A classic example is that of tests scores used to decide on student's eligibility of attractive educational options, or to decide on required remedial programs. If there is discretion in the grading of the tests, and the graders are aware of both the importance of the test, and of the value of the threshold, and if the graders have preferences over the outcomes for the students, they may change grades for some individuals sufficiently close to the threshold.

It can be challenging to address this type of manipulation through the statistical analysis, although there are some interesting approaches involving shape restrictions on the underlying potential outcome distributions (Diamond and Persson (2016)). Much easier is the task of establishing whether such manipulation is taking place. If there is, one would expect a discontinuity in the marginal density of the forcing variable because for individuals on one side of the threshold there, and for individuals on the other side of the threshold there is no, incentive to manipulate the score. McCrary (2008) developed a test for the null hypothesis of no discontinuity in the density of the forcing variable that should be performed any time someone does a rdd analysis. See also Otsu et al. (2015) for an alternative version of the test. Note that, for the purpose of estimating the difference in the conditional means of the outcome on the right and the left of the threshold, there is formally no need for the marginal density of the forcing variable to be continuous at that point. The reason that the test is important is that the argument that underlies the identification strategy, and in particular the notion that individuals directly to the left and the right of the threshold are comparable other than through the receipt of the treatment, is difficult to reconcile with finding that there are substantially fewer people just to the right than to the left of the threshold.

A second supplementary analyses in the rdd setting involves checking the continuity of the conditional expectation of exogenous variables around the threshold. Again this continuity is not required for consistency, but a discontinuity in such conditional expectations is difficult to reconcile with comparability of the individuals to the left and the right of the threshold, and would suggest the possibility of unobserved imbalances on the right and the left. Such analyses are similar to placebo analyses in studies of causal effects under unconfoundedness, where often tests for zero effects on lagged outcomes are presented to assess unconfoundedness (Athey and Imbens (2016); Imbens and Rubin (2015)).

5. Bandwidth Choice in Regression Discontinuity Designs

The currently preferred analysis in rdd settings, e.g., Hahn et al. (2001); Porter (2003) is to use local linear, or sometimes local quadratic methods (Calonico et al. (2014)) rather than simple kernel estimators or global high order polynomial methods. Simple kernel regression methods have poor properties at the boundary of the support, and that is precisely where we are interested in the estimates in this setting. Gelman and Imbens (2014) argue against the use of global high-order polynomials because of poor properties in terms of mean-squared error, coverage rates for confidence intervals and the difficulties in selecting the order of the polynomial. Given the use of local regression methods, the question is how to choose the degree of localness, that is, the bandwidth. Early on in the literature common practice was to use off-the-shelf bandwidth selection methods based on crossvalidation, e.g., Ludwig and Miller (2005). However, crossvalidation methods are not as attractive here as they are for bandwidth selection in nonparametric regression because in the current setting we are interested in the value of the regression only at a few points. More recently bandwidth selection methods have been developed that are specifically geared towards the goal of precisely estimating the magnitude of the discontinuity, at the threshold, in the conditional expectation of the outcome given the forcing variable, (Imbens and Kalyanaraman (2012); Calonico et al. (2014)).

These bandwidth selection methods are based on asymptotically balancing the square of the bias and the variance of the estimator for the limit of the value of the regression function at the threshold, from the right and the left.

6. External Validity in Regression Discontinuity Designs

One, and perhaps the, major limitation of rdd analyses is the lack of external validity. In many cases the methods lead to estimates with a high degree of internal validity, but the conclusions are limited in two aspects. First, they are restricted to the subpopulation of compliers, and second they are restricted to individuals with values for the forcing variable close to the threshold. Recently there has been some work examining the presence and credibility of any evidence that these estimates have wider external validity, be that for non-compliers, or for units with values of the forcing variable away from the threshold.

First consider only units close to the threshold. Battistin and Rettore (2008) and Bertanha and Imbens (2016) propose comparing compliers without the treatment ("control compliers") to never-takers at the threshold, and comparing compliers with the treatment ("treated compliers") to always-takers at the threshold. If one clearly rejects the null hypotheses that, say never-takers and control compliers, are comparable, than it appears less plausible that the average treatment effect for compliers (which is estimable) is useful as a predictor for the average effect of the treatment for never-takers (which we cannot estimate directly). If, on the other hand, we find that treated compliers are comparable to always-takers, and control compliers are comparable to never-takers, it appears more plausible that the average effect for compliers is indicative of the average effect for the other subpopulations close to the threshold. In that case the external validity of the rdd estimates is enhanced. Note that the same argument can be made in for other, non-rdd versions of instrumental variables, and it is related to the discussion on testing in Angrist (2004). Bertanha and Imbens (2016) point out that there is a convenient graphical interpretation of this null hypothesis, namely the continuity of the conditional expectation of the outcome as a function of the forcing variable, conditional on the treatment group, adjusted for other covariates.

Angrist and Fernandez-Val (2010) and Angrist and Rokkanen (2015) take different approaches to the extrapolation to other subpopulations. In the context of instrumental variables estimation, but in a way that can conceptually easily be extended to rdd settings, Angrist and Fernandez-Val (2010) focus on the difference between estimators based on unconfoundedness and estimators based on iv or rdd assumptions (in both sharp and fuzzy rdd settings). If exogenous covariates can eliminate the differences between the two, they argue that extrapolating the complier effects to the general population is more plausible. In the context of sharp rd designs Angrist and Rokkanen (2015) focus on the role of covariates to eliminate differences between units with different values of the forcing variable. If the other covariates can eliminate all or most of the association between the forcing variable and the outcomes away from the threshold, again it becomes more plausible to extrapolate the estimated effects at the threshold to other subpopulations.

Dong and Lewbel (2015) point out that under the rdd assumptions one can in principle identify not simply the level of the conditional expectation on both sides of the threshold, but also derivatives of this conditional expectation. They explore using estimates of these derivatives to extrapolate away from the threshold.

7. Multiple Thresholds and Multiple Assignment Variables

In many applications the assignment process is more complex than covered by the simple rdd setting. There may be multiple thresholds at which incentives to participate in a program change discontinuously, as in Bertanha (2015); Abdulkadiroğlu et al. (2014). In many such cases there is not sufficient information at a single threshold to obtain precise estimates of the causal effect at that threshold. In that case one may wish to combine the estimates at the different thresholds into a single average effect.

There may also be multiple measures that enter into the eligibility decision, as in Papay et al. (2011); Imbens and Zajonc (2011). For example, a student may be required to participate in a remedial program unless the student receives a passing grade in both mathematics and reading tests. In this case the researcher has several options. One can analyze the data using the minimum of the two grades in a sharp rdd. In that case one can also assess heterogeneity in the effects by comparing individuals close to the reading threshold among the subpopulation with mathematics test scores above the threshold, or the other way around. One can also analyze the data using either the reading or mathematics score as a forcing variable in a fuzzy rdd.

8. Regression Kink Designs

A very recent generalization of regression discontinuity designs is what has been labelled the regression kink design (rkd), Card et al. (2015) and Dong (2014). In this case the treatment of interest is a continuous one. At the threshold the conditional expectation of the outcome is not expected to change discontinuously. However, the derivative of the conditional expectation at that point may change discontinuously, leading to a kink in the conditional expectation, lending the approach its name. The discontinuity in the derivative of the conditional expectation of the outcome is attributed to the the discontinuity in the derivative of the conditional expectation of the treatment given the forcing variable changes at the threshold. For example, consider a case where the researcher is interested in the effect of unemployment benefits on subsequent earnings. The treatment of interest is the benefit level an individual receives. The forcing variable may be prior earnings, in a setting where the benefits decrease with earnings, with the rate of decrease changing discontinuously at the threshold. Card et al. (2015) and Dong (2014) extend rdd methods to such cases. Obviously estimating the change in the derivatives is a more challenging task than estimating the change in the level of a conditional expectation, and consequently regression kink analyses will require in practice more data than regression discontinuity analyses.

9. Conclusion

In this note I discuss some of the recent work in econometrics on regression discontinuity designs. Decades after these methods were first introduced by Thistlewaite and Campbell (1960), they are now among the most widely used methods for causal inference in economics and other social sciences. This has motivated more methodological advances in what is currently a very active research area.

References

Abdulkadiroğlu, A., Angrist, J., and Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, 82(1):137–196.

Angrist, J. and Fernandez-Val, I. (2010). Extrapolate-ing: External validity and overidentification in the late framework. Technical report, National Bureau of Economic Research.

Angrist, J. and Pischke, S. (2008). *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton University Press.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–472.

Angrist, J. D. and Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

Athey, S. and Imbens, G. (2016). The state of applied econometrics-causality and policy evaluation. *arXiv preprint arXiv:1607.00699*.

Battistin, E. and Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, 142(2):715–730.

Bertanha, M. (2015). Regression discontinuity design with many thresholds. *Available at SSRN*.

Bertanha, M. and Imbens, G. (2016). External validity in fuzzy regression discontinuity designs. CORE Discussion Paper 2016/25.

Black, S. (1999). Do better schools matter? parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599.

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

Calonico, S., Cattaneo, M., and Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.

Choi, J.-y. and Lee, M.-j. (2016). Regression discontinuity: review with extensions. *Statistical Papers*, pages 1–30.

Cook, T. (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.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Dong, Y. (2014). Jump or kink? identification of binary treatment regression discontinuity design without the discontinuity. *Unpublished manuscript.*

Dong, Y. and Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5):1081–1092.

Gelman, A. and Imbens, G. (2014). Why high-order polynomials should not be used in regression discontinuity designs. NBER Working Paper No. 20405.

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.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–970.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

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

Imbens, G. and Zajonc, T. (2011). Regression discontinuity design with multiple forcing variables. *Unpublished manuscript.*

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 61:467–476.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

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

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

Ludwig, J. and Miller, D. L. (2005). Does head start improve children's life chances? evidence from a regression discontinuity design. Technical report, National Bureau of Economic Research.

McCrary, J. (2008). Testing for manipulation of the running variable in the regression discontinuity design. *Journal of Econometrics*, 142(2):698–714.

Otsu, T., Xu, K.-L., and Matsushita, Y. (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics*, 186(1):94–112.

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161(2):203–207.

Porter, J. (2003). Estimation in the regression discontinuity model. Available on CiteSeer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company.

Skovron, C. and Titiunik, R. (2015). A practical guide to regression discontinuity designs in political science. Technical report, working paper, University of Michigan.

Thistlewaite, D. and Campbell, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(2):309–317.

Trochim, W. M. (1984). *Research design for program evaluation: The regression-discontinuity approach*, volume 6. SAGE Publications, Inc.

Van Der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43(2):1249–1287.

Van Der Klaauw, W. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour*, 22(2):219–245.