Journal of Research on Educational Effectiveness, 5: 43–82, 2012

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2011.578707

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

Modern Regression Discontinuity Analysis

Howard S. Bloom
MDRC, New York, New York, USA

Abstract: This article provides a detailed discussion of the theory and practice of modern regression discontinuity (RD) analysis for estimating the effects of interventions or treatments. Part 1 briefly chronicles the history of RD analysis and summarizes its past applications. Part 2 explains how in theory an RD analysis can identify an average effect of treatment for a population and how different types of RD analyses—"sharp" versus "fuzzy"—can identify average treatment effects for different conceptual subpopulations. Part 3 of the article introduces graphical methods, parametric statistical methods, and nonparametric statistical methods for estimating treatment effects in practice from regression discontinuity data plus validation tests and robustness tests for assessing these estimates. Section 4 considers generalizing RD findings and presents several different views on and approaches to the issue. Part 5 notes some important issues to pursue in future research about or applications of RD analysis.

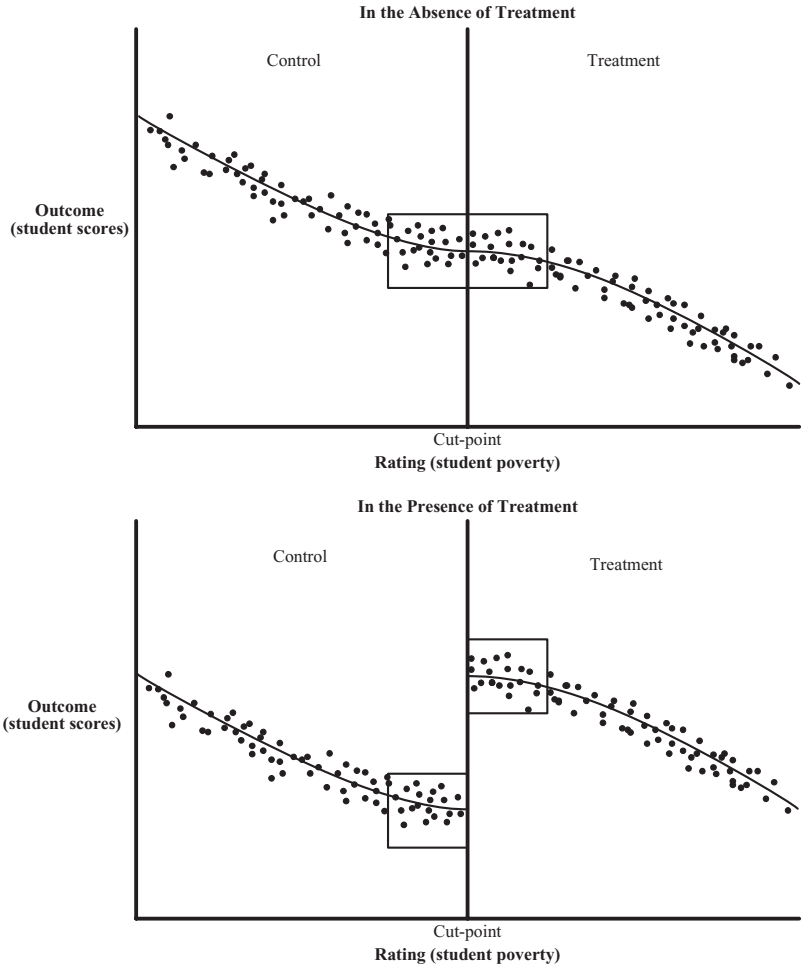
Keywords: regression discontinuity designs, program impacts, sample design

PART 1: HISTORY AND PAST APPLICATIONS

Regression discontinuity analysis applies to situations in which candidates are selected for treatment based on whether their value for a numeric rating exceeds a designated threshold or cut-point.[1] For example, students may be chosen for a scholarship based on a summary measure of their academic achievement. Because this type of process is used widely to allocate resources or impose sanctions, it provides many opportunities for using regression discontinuity analysis. For example, it is often the case that schools are selected for funding based on levels of student poverty, families are judged to be eligible for government assistance based on income, convicts are assigned to prison security levels based on prior criminal records, and government funding is awarded based on reviewer ratings of grant proposals. A valid regression discontinuity design exists in these and similar situations when decisions about where to set the cut-point are made independently of decisions about what ratings to assign to specific candidates.

Address correspondence to Howard S. Bloom, MDRC, 16 East 34th Street, New York, NY 10016, USA. E-mail: howard.bloom@mdrc.org

[1]The regression discontinuity literature uses various terms for the rating, including "forcing variable" and "assignment variable."



NOTE: Dots represent individual schools. The vertical line in the center of each graph designates a cut-point, above which candidates are assigned to the treatment and below which they are not assigned to the treatment. The boxes represent the portion of the distribution proximal enough to the cut-point to represent local randomization.

Figure 1. Two Ways to Characterize Regression Discontinuity Analysis.

Figure 1 illustrates two ways to characterize regression discontinuity analysis: as "discontinuity at a cut-point" (Hahn, Todd, & van der Klaauw, 1999) and as "local randomization" (Lee, 2008). The graphs in the figure portray a relationship that might exist between an outcome for candidates being considered for a prospective treatment and a rating used to prioritize candidates for that treatment. (In this case, mean student test scores for schools are an outcome for candidates being considered for the prospective treatment of school aid, whereas the percentage of students who live in poverty is a rating used to prioritize candidates for aid.) The vertical line in the center of each graph designates a cut-point, above which candidates are assigned to the treatment and below which they are not assigned to the treatment.

The top graph illustrates what one would expect in the absence of treatment. As can be seen, the relationship between outcomes and ratings is downward sloping to the right, which indicates that mean student test scores decrease as rates of student poverty increase.

This relationship passes continuously through the cut-point, which implies that there is no difference in outcomes for candidates that are just above and below the cut-point.

The bottom graph in the figure illustrates what would occur in the presence of treatment if it increased outcomes. In this case, there is a sharp upward jump at the cut-point in the relationship between outcomes and ratings. The first characterization of regression discontinuity analysis—discontinuity at a cut-point—focuses on this jump, the direction and magnitude of which is a direct measure of the causal effect of the treatment on the outcome for candidates near the cut-point.

The second characterization of regression discontinuity analysis—local randomization—is based on the premise that differences between candidates who just miss and just make a threshold are random. This could occur, for example, from random error in test scores used to rate candidates. Candidates who just miss the cut-point are thus identical, on average, to those who just make it, except for exposure to treatment. Any difference in subsequent mean outcomes must therefore be caused by treatment.

Regression discontinuity analysis was developed by Thistlethwaite and Campbell (1960) to study the effects of winning a national merit scholarship certificate.[2] Their work generated a flurry of related activity, which died out subsequently. After two decades of a regression discontinuity "dark ages," economists revived the approach (Angrist & Lavy, 1999; van der Klaauw, 1997, 2002), formalized it (Hahn, Todd, & van der Klaauw, 2001), strengthened its estimation methods (Imbens & Kalyanaraman, 2009), and began to apply it to many different research questions. This renaissance was culminated recently in a 2008 special issue on regression discontinuity analysis in the *Journal of Econometrics* (Imbens & Lemieux, 2008a).

Regression discontinuity analysis has been used to study, among other things, the effect of financial aid on college enrollment (van der Klaauw, 1997, 2002), incumbency on electoral success (Lee, 2001), mandatory summer school on student achievement (Jacob & Lefgren, 2004; Matsudaira, 2008), Head Start programs on children's mortality and educational attainment (Ludwig & Miller, 2007), class size on academic achievement (Angrist & Lavy, 1999), prison conditions on recidivism (Chen & Shapiro, 2005), unemployment insurance on recidivism (Berk & Rauma, 1983), preschool programs on children's preliteracy skills (Wong, Cook, Barnett, & Jung, 2007), Medicaid eligibility on health insurance coverage for low-income children (Card & Shore-Sheppard, 2004), federal Title I funding for low-income schools on student performance (van der Klaauw, 2008), financial incentives to improve school attendance and health care for children (Buddelmeyer & Skoufias, 2002), unemployment insurance on unemployment (Lalive, 2008), unionization on establishment closures (DiNardo & Lee, 2002), school quality on house values (S. Black, 1999), antidiscrimination legislation on minority hiring (Hahn et al., 1999), and the federal Reading First Program on student achievement (Gamse et al., 2008).

The next part of this article describes how, in principle, regression discontinuity analysis can identify a treatment effect for a population. In other words, it indicates how the basic logic of regression discontinuity analysis can demonstrate a causal effect. Part 3 discusses how, in practice, regression discontinuity analysis can estimate a treatment effect for a sample. It describes alternative statistical procedures for implementing regression discontinuity analyses and examines their strengths and limitations. Part 4 considers how to interpret and generalize results from regression discontinuity analyses. Part 5 considers some important frontiers for future regression discontinuity research.

[2]Cook (2008) chronicled the history of regression discontinuity analysis.

PART 2: HOW REGRESSION DISCONTINUITY ANALYSIS IDENTIFIES AVERAGE TREATMENT EFFECTS FOR POPULATIONS

The existing literature typically distinguishes two types of regression discontinuity design: the "sharp" design, in which all subjects receive their assigned treatment or control condition, and the "fuzzy" design, in which some subjects do not. Following the lead of Battistin and Retorre (2008), this chapter distinguishes three types of regression discontinuity designs:

1. Sharp designs, as defined conventionally.
2. Type I fuzzy designs, in which some treatment group members do not receive treatment. Such members are referred to as "no-shows" (Bloom, 1984).
3. Type II fuzzy designs, in which some treatment group members do not receive treatment, and some comparison group members do. (Members in the latter category are referred to as "crossovers"; Bloom et al., 1997).

This section describes how each regression discontinuity design identifies a treatment effect for a population. It first defines treatment effects using the potential outcomes framework from the statistics literature on causal inference.[3] Then, as a point of departure, the section describes how randomized trials identify average causal effects. Finally, it presents the conditions that are required for valid regression discontinuity designs and describes how such designs can identify average treatment effects (see the appendix for an elaboration of the findings presented).

Defining Treatment Effects

An individual, i, has two potential outcomes, Y_{1i} and Y_{0i}, which represent what would occur with and without treatment, respectively. These outcomes might refer to health status, academic achievement, economic success, criminal behavior, and the like. The causal effect of treatment on an outcome for individual i, designated as β_{i}, is the difference between his potential outcomes with and without treatment, $(Y_{1i} - Y_{0i})$. The average treatment effect or ATE for a population is the expected value (mean) of its individual treatment effects, or

$$\mathrm{ATE} \equiv E\{\beta\} \equiv E\{Y_1 - Y_0\} \tag{1}$$

This, in turn, equals the difference between expected population outcomes with and without treatment, or

$$\mathrm{ATE} = E\{Y_1\} - E\{Y_0\} \tag{2}$$

The main obstacle to identifying an individual treatment effect is the inability to observe outcomes with and without treatment for the same individual at the same time.

[3]This framework is often attributed to Rubin (1974) and was labeled the Rubin Causal Model by Holland (1986). However, its roots date back to Neyman (1923), Fisher (1935), Roy (1951), and Quandt (1972).

Identifying Treatment Effects With Randomization

Randomization produces treatment and control groups that are the same (in expectation) except for exposure to treatment. Any observed outcome differences therefore can be attributed to treatment. Stated formally, randomizing subjects to treatment status ($T = 1$) or control status ($T = 0$) creates two groups, whose expected outcomes in the absence of treatment ($E\{Y_0\}$) are the same, or

$$E\{Y_0 \mid T = 1\} = E\{Y_0 \mid T = 0\} \tag{3}$$

where $E\{Y_0 \mid T = 1\}$ = the expected outcome without treatment given randomization to treatment and $E\{Y_0 \mid T = 0\}$ = the expected outcome without treatment given randomization to control status.

The ATE for a randomized trial equals the difference between the expected treatment group outcome with and without treatment, or

$$\text{ATE} \equiv E\{Y_1 \mid T = 1\} - E\{Y_0 \mid T = 1\} \tag{4}$$

With full compliance to randomization (all candidates receive their assigned condition) the expected treatment group outcome in the absence of treatment (its counterfactual outcome) in Equation 4 can be replaced by the expected outcome for the control group, yielding

$$\begin{aligned} \text{ATE} &= E\{Y_1 \mid T = 1\} - E\{Y_0 \mid T = 0\} \\ &= E\{Y \mid T = 1\} - E\{Y \mid T = 0\} \end{aligned} \tag{5}$$

Equation 5 states that with full compliance, the difference between expected outcomes for the treatment group and control group identifies the ATE.

Defining a Valid Regression Discontinuity Design

A valid regression discontinuity design can identify a treatment effect in much the same way that a randomized trial does so. But for a regression discontinuity design to be valid, candidates' ratings and the cut-point must be determined independently of each other.[4] This condition can be ensured if the cut-point is determined without knowledge of candidates' ratings and if candidates' ratings are determined without knowledge of the cut-point.[5]

On the other hand, if the cut-point is to be chosen in the presence of knowledge about candidates' ratings, decision makers can locate the cut-point in a way that includes or excludes specific candidates. If these candidates differ, those on one side of the cut-point will not provide valid information about the counterfactual outcome for those on the other side. This situation could arise, for example, when a fixed sum of grant funding is allocated to a pool of candidates and average funding per recipient is determined in light of knowledge about candidates' ratings. With a fixed total budget, average funding per

[4]In the evaluation research literature, the form of validity referred to here is often called "internal validity" (Shadish, Cook, & Campbell, 2002).

[5]This is a *sufficient* condition.

recipient determines the number of candidates funded, which in turn determines the cut-point. Through this mechanism, the cut-point could be manipulated to include or exclude specific candidates.

Furthermore, if ratings are determined in the presence of knowledge about the corresponding cut-point, they can be manipulated to include or exclude specific candidates. For example, if a college's admissions director were the only person who rated students for admission, he could fully determine whom to accept and whom to reject by setting ratings accordingly. Consequently, students accepted could differ from those rejected in unobserved ways, and their counterfactual outcomes would differ accordingly. A second possible example is one in which students must pass a test to avoid mandatory summer school, and they know the minimum passing score. In this case, students who are at risk of failing but sufficiently motivated to work extra hard might be especially prevalent among just-passing scores and students with similar aptitude but less motivation might be especially prevalent among just-failing scores. The two groups therefore will not provide valid information about each other's counterfactual outcomes.

Lee (2008) and Lee and Lemieux (2009) provided an important insight into the likelihood of meeting the necessary condition for a valid regression discontinuity design. They did so by distinguishing between situations with precise control over ratings (which are rare) and situations with imprecise control over ratings (which are typical). Precise control means that candidates or decision makers can determine the exact value of each rating. This was assumed to be the case in the preceding two examples where a college admissions director could fully determine applicants' ratings or individual students could fully determine their test scores.

The situation is quite different, however, when control over ratings is imprecise, which would be the case in more realistic versions of the preceding examples. Most colleges have multiple members of an admissions committee rate each applicant, and thus no single individual can fully determine a student's rating. Consequently applicant ratings contain random variation due to differences in raters' opinions and variation in their opinions over time. Also, because of random testing error, students cannot fully determine their scores on a test.[6] Lee and Lemieux (2009) have demonstrated that such random variation is the sole factor determining which candidates fall just below and above a cut-point. They thereby have demonstrated that imprecise control over ratings is sufficient to produce random assignment at the cut-point, which yields a valid regression discontinuity design.

Identifying Treatment Effects with a Sharp Regression Discontinuity Design

Figure 2 illustrates how regression discontinuity analysis can identify a treatment effect. The top graph represents a sharp regression discontinuity design, the middle graph represents a Type I fuzzy regression discontinuity design, and the bottom graph represents a Type II fuzzy regression discontinuity design. To make the example concrete, assume that candidates are schools, the outcome for each school is average student test scores, and the rating for each school is a measure of its student poverty (e.g., the percentage of students eligible for subsidized meals). Also assume that the analysis represents a population, not just a sample.

[6]For example, students can misread questions or momentarily forget things they know.

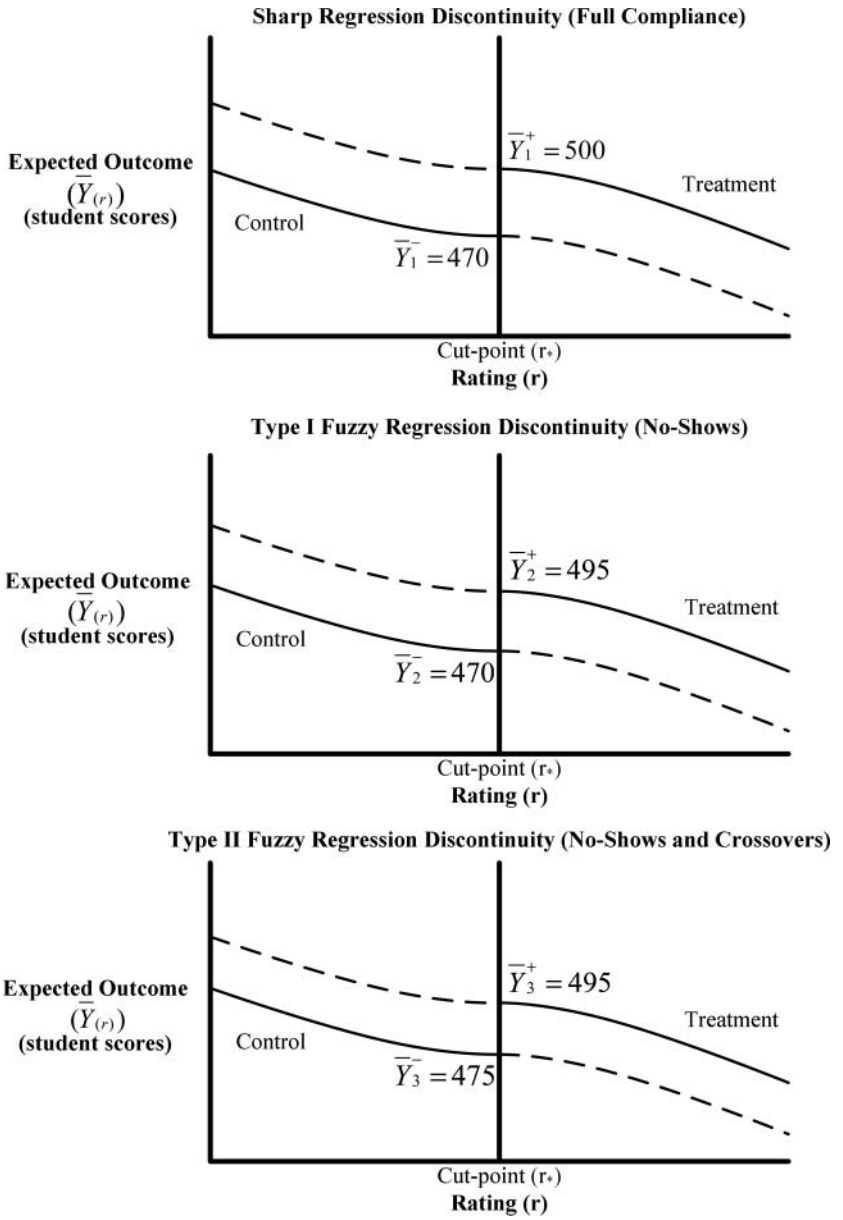


Figure 2. Illustrative Regression Discontinuity Analyses.

Curves in the graph are regression models of the relationship between expected outcomes ($\bar{Y}(r)$) and ratings (r).[7] These curves are downward-sloping to represent the negative relationship that typically exists between student performance and poverty. Schools with ratings at or above a cut-point ($r*$) are assigned to treatment (e.g., government assistance),

[7] A regression model represents the relationship between expected values of a dependent variable and specific values of an independent variable.

and others are assigned to a control group that is not eligible for the treatment. In the top graph, all schools assigned to treatment receive it, and no schools assigned to control status receive treatment. In the middle graph, some schools assigned to treatment do not receive it, but no schools assigned to control status do receive treatment. In the bottom graph, some schools assigned to treatment do not receive it, and some schools assigned to control status do receive treatment.

For each graph, the solid line segment to the left of the cut-point indicates that expected outcomes for the control group decline continuously as ratings approach the cut-point from below—that is, as ratings increase toward their cut-point value. The symbol $\bar{Y}^-$ represents the expected outcome at the cut-point approached by this line. The dashed extension of the control group line segment represents what expected outcomes would be without treatment for schools with ratings above the cut-point (their expected counterfactual outcomes). The two line segments for the control group form a continuous line through the cut-point; there is no discontinuity.

The solid line segment to the right of the cut-point indicates that expected outcomes for the treatment group rise continuously as ratings approach the cut-point from above—that is, as ratings decrease toward their cut-point value. The symbol $\bar{Y}^+$ represents the expected outcome at the cut-point approached by this line. The dashed extension of the treatment-group line segment represents what outcomes would be with treatment for subjects with ratings below the cut-point. The two line segments for the treatment group form a continuous line through the cut-point; again, there is no discontinuity.

When expected outcomes are a continuous function of ratings through the cut-point in the absence of treatment, the discontinuity, or gap, that exists between the solid line segment for the treatment group and the solid line segment for the control group, representing observable outcomes for each group, can be attributed to the availability of treatment for treatment group members. This discontinuity $\left(\bar{Y}^+ - \bar{Y}^-\right)$ equals the average effect of assignment to treatment, which is often called the average effect of intent-to-treat (ITT). For a regression discontinuity analysis, this is the average effect of intent-to-treat at the cut-point (ITTC).

Figure 3 indicates the key distinctions that exist among the three regression discontinuity analyses portrayed by Figure 2. The top graph in Figure 3 for a sharp regression discontinuity design indicates that the probability of receiving treatment equals a value of zero for schools with ratings below the cut-point and a value of one for schools with ratings above the cut-point. Hence the limiting value of the probability as the rating approaches the cut-point from below ($\bar{T}^-$) is zero, and its limiting value as the rating approaches the cut-point from above ($\bar{T}^+$) is one.[8] The discontinuity in the probability at the cut-point ($\bar{T}^+ - \bar{T}^-$) therefore equals a value of one for a sharp regression discontinuity.

Results in the top graphs of Figures 3 and 2 come together as follows. Moving from left to right, the probability of receiving treatment has a constant value of zero until the cut-point is reached and the probability shifts abruptly to a constant value of one. If expected potential outcomes vary continuously with ratings in the absence of treatment, then the only possible cause of a shift in observed outcomes at the cut-point (Figure 2) is the shift in the probability of receiving treatment (Figure 3).

Another way to explain this result is to note that as one approaches the cut-point, the resulting treatment group and control group become increasingly similar in all ways except for receipt of treatment. Hence, at the cut-point, assignment to treatment by ratings is

[8] $\bar{T}$ is used to represent the probability of receiving treatment because it equals the mean value of T.

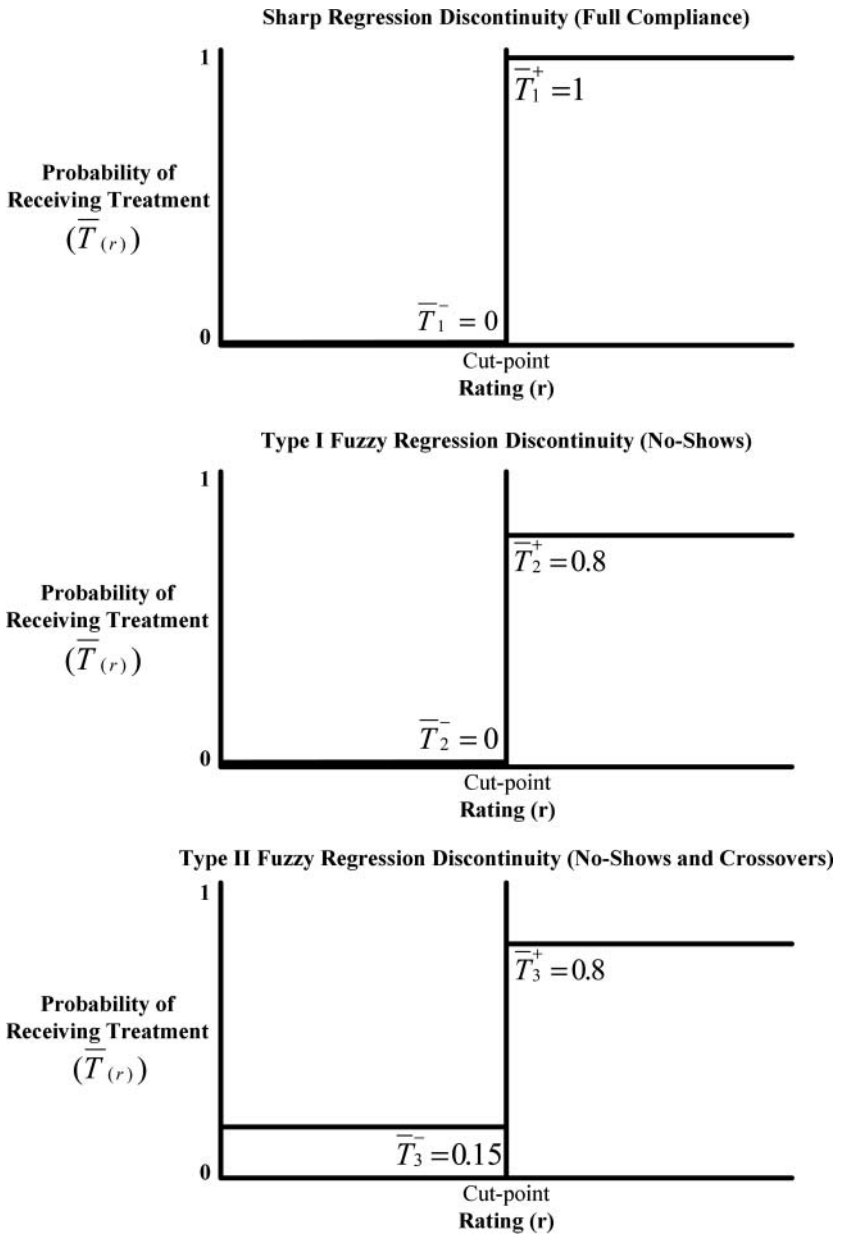


Figure 3. The Probability of Receiving Treatment As a Function of the Rating.

like random assignment to treatment as noted earlier. Differences at the cut-point between expected treatment group and control group outcomes therefore must be caused by the difference in treatment receipt.

The top graph in Figure 3 implies that for a sharp regression discontinuity design, assignment to treatment is the same as receipt of treatment. Hence, the average effect of assignment to treatment at the cut-point (ITTC) is the same as the average effect of treatment on the treated at the cut-point (TOTC), which in turn is the same as the average

treatment effect for the full population at the cut-point (ATEC). The fact that each of these parameters is defined at the cut-point has important implications for their generalizability (discussed later).

To complete the example for a sharp regression discontinuity design, assume that expected control-group outcomes converge as ratings approach the cut-point from below, to a limiting score ($\bar{Y}-$) of 470 points, and expected treatment-group outcomes converge as ratings approach the cut-point from above, to a limiting score ($\bar{Y}_1^+$) of 500 points. The resulting 30-point discontinuity ($\bar{Y}_1^+ - \bar{Y}_1^-$) is the average effect at the cut-point of assignment to treatment (ITT1C). For a sharp regression discontinuity this equals the corresponding average effect of receiving treatment (TOT1C) and the average treatment effect for all members of the cut-point population (ATE1C).

Identifying Treatment Effects With a Type I Fuzzy Regression Discontinuity Design

Many voluntary government programs base their eligibility on numeric ratings. For example, to qualify for food stamps, a family must have an income that falls below a threshold level. To qualify for compensatory education funding, a school's student poverty level must exceed a specified threshold. Candidates on one side of the threshold can participate in the program, whereas candidates on the other side cannot. However, not all eligible candidates necessarily participate; some may become "no-shows" (Bloom, 1984).[9] This situation represents a Type I fuzzy regression discontinuity.

The middle graphs in Figures 2 and 3 illustrate this case. In Figure 3 the probability of receiving treatment is zero for control group members (as with a sharp regression discontinuity) and is positive but less than one for treatment group members (unlike with a sharp regression discontinuity). In our example, this means that some schools that are eligible for treatment do not receive it and become no-shows. No-shows dilute the treatment contrast in a regression discontinuity analysis by reducing the difference in the proportion of treatment group members and control group members who receive the treatment being tested. Reducing the difference in the proportion of treatment group and control group members who receive treatment reduces the expected difference in outcomes for treatment group and control group members, which in turn reduces the magnitude of the discontinuity at the cut-point in expected outcomes.

In our example—which assumes a positive treatment effect—no-shows reduce expected outcomes for the treatment group in the presence of treatment. Hence, the treatment group line in the middle graph of Figure 2 is somewhat lower than its counterpart in the top graph. Consequently, $\bar{Y}_2^+$ is less than $\bar{Y}_1^+$. Because no-shows do not affect control group outcomes, the control-group line in the middle graph of Figure 2 is the same as its counterpart in the top graph and $\bar{Y}_2^-$ equals $\bar{Y}_1^-$.

In the Type I fuzzy regression discontinuity design example, the average effect of assignment to treatment or intent to treat (ITT2C) equals the difference between limiting values of expected outcomes for the treatment group and control group at the cut-point ($\bar{Y}_2^+ - \bar{Y}_2^-$)—as was the case for a sharp regression discontinuity. However this difference no longer represents the average effect at the cut-point of receiving treatment (TOT2C) or the average treatment effect for the cut-point population (ATE2C). Without further assumptions, these additional parameters cannot be identified.

[9]Bloom (1984) defined no-shows as treatment group members who do not receive treatment *for any reason*.

Fortunately, it is often reasonable to assume that no-shows experience approximately no effect of assignment to treatment because only receipt of treatment can produce the desired causal effect. If so, then the average effect of assignment to treatment at the cut-point (ITT2C) is a weighted mean of the average effect at the cut-point of receiving treatment (TOT2C) for treatment recipients and zero effect for no-shows, weighted by the proportion of treatment group members who receive treatment ($\bar{T}_2^+$) and do not receive treatment ($1 - \bar{Y}_2^+$).[10] In symbols:

$$\begin{aligned} \text{ITT2}_\text{C} &= \bar{T}_2^+ \cdot \text{TOT2}_\text{C} + (1 - \bar{T}_2^+) \cdot 0 \\ &= \bar{T}_2^+ \cdot \text{TOT2}_\text{C} \end{aligned} \tag{6}$$

Hence

$$\text{TOT2}_\text{C} = \frac{\text{ITT2}_\text{C}}{\bar{T}_2^+} = \frac{\bar{Y}_2^+ - \bar{Y}_2^-}{\bar{T}_2^+} \tag{7}$$

Intuitively, Equation 7 allocates all of the treatment and control group difference in expected outcomes to treatment recipients, which follows from the assumption that no-shows experience no effect of assignment to treatment.

In our numeric example, assume that 80% of treatment group members receive treatment ($\bar{T}^+ = 0.8$) and the limiting value of expected treatment-group outcomes at the cut-point is 495 points, instead of 500 for the sharp regression discontinuity, because no-shows experience no effect. The limiting value of expected control-group outcomes at the cut-point is 470 points, as it was for the sharp regression discontinuity. Then,

$$\text{TOT2}_\text{C} = \frac{495 - 470}{0.8} = 31.25$$

For the Type I fuzzy regression discontinuity design example, the average effect of treatment on the treated at the cut-point (TOT2C) is a gain of 31.25 points instead of 30 points for the sharp regression discontinuity design. (TOT1C). This difference reflects the fact that treatment recipients in the two designs comprise different populations. Specifically, the population for a Type I fuzzy regression discontinuity is only part of that for a sharp regression discontinuity.

Identifying Treatment Effects With a Type II Fuzzy Regression Discontinuity Design

Programs that choose participants based on numeric ratings often have exceptions to their assignment rules that result both in no-shows—candidates whose ratings should have them assigned to treatment but do not receive it—and crossovers—candidates whose ratings should have them assigned to a control group but do receive treatment. Hence, two factors can reduce the effective treatment contrast for a regression discontinuity analysis and thereby reduce the observable difference between expected outcomes for its treatment group and control group. This situation produces a Type II fuzzy regression discontinuity.

[10]TOT2C also can be identified if the average effect of assignment to treatment for no-shows is a specified fraction of that for recipients. By varying this fraction and recomputing the result, one can test the sensitivity of the result to the assumed relative effect for no-shows.

With a positive treatment effect, no-shows reduce expected outcomes for a treatment group. In our example, this phenomenon is represented by the fact that the treatment group line for a Type II fuzzy regression discontinuity in the bottom graph of Figure 2 is lower than its counterpart for a sharp regression discontinuity in the top graph of Figure 2. Hence, the limiting value of expected outcomes at the cut-point for a treatment group is lower for a Type II fuzzy regression discontinuity than for a sharp regression discontinuity ($\bar{Y}_3^+ < \bar{Y}_1^+$). With a positive treatment effect, crossovers increase expected outcomes for a control group. This is represented by the fact that the control group line for a Type II regression discontinuity in the bottom graph of Figure 2 is higher than its counterpart for a sharp regression discontinuity in the top graph of Figure 2. Hence, the limiting value of expected outcomes at the cut-point for control group members is higher for a Type II fuzzy regression discontinuity than for a sharp regression discontinuity ($\bar{Y}_3^- > \bar{Y}_1^-$).

In the presence of both crossovers and no-shows the discontinuity in expected outcomes at a cut-point $\bar{Y}_3^+ - \bar{Y}_3^-$ still represents the average effect of assignment to treatment (ITT3C). But it does not equal the average effect of treatment on the treated at the cut-point (TOT3C) or the average treatment effect for the cut-point population (ATE3C). In fact, neither of these parameters can be identified without fairly strong assumptions.

Fortunately, weaker assumptions can identify what is often referred to as a "local average treatment effect," or LATE (Angrist, Imbens, & Rubin, 1996); for regression discontinuity designs, this would be designated the LATE at a cut-point, or LATEC. This parameter is defined as the average effect of treatment on candidates at a cut-point who receive treatment because they are assigned to it. This subpopulation is often referred to as "compliers" (Angrist et al., 1996), because they comply with their treatment assignment; they receive treatment if assigned to a treatment group and do not receive treatment if assigned to a control group.

Conditions for identifying a LATE (the ATE for compliers) were derived by Angrist et al. (1996) based on a conceptual framework that specifies four mutually exclusive and collectively exhaustive subpopulations of a treatment group and its control group.

- *Compliers* receive treatment if and only if assigned to it. A regression discontinuity design (or randomized trial) can potentially identify an ATE for this subpopulation because it produces a treatment contrast for its members; they receive treatment if assigned to it and do not receive treatment if not assigned to it.
- *Always-takers* receive treatment regardless of their assignment. A regression discontinuity design (or randomized trial) cannot identify an ATE for this subpopulation because it does not produce a treatment contrast for its members; they receive treatment regardless of whether or not they are assigned to it.
- *Never-takers* do not receive treatment regardless of their assignment. A regression discontinuity design (or randomized trial) cannot identify an ATE for this subpopulation because it does not produce a treatment contrast for its members; they do not receive treatment regardless of whether or not they are assigned to it.
- *Defiers* receive treatment if and only if not assigned to it. Without further assumptions, a regression discontinuity design (or randomized trial) cannot identify an ATE for this subpopulation because its members cannot be distinguished from always-takers in a control group and never-takers in a treatment group.

Individual members of a given subpopulation cannot be identified as such, but in a randomized trial or a valid regression discontinuity design each subpopulation theoretically should comprise the same proportion of a treatment group and control group. Thus, if 10%

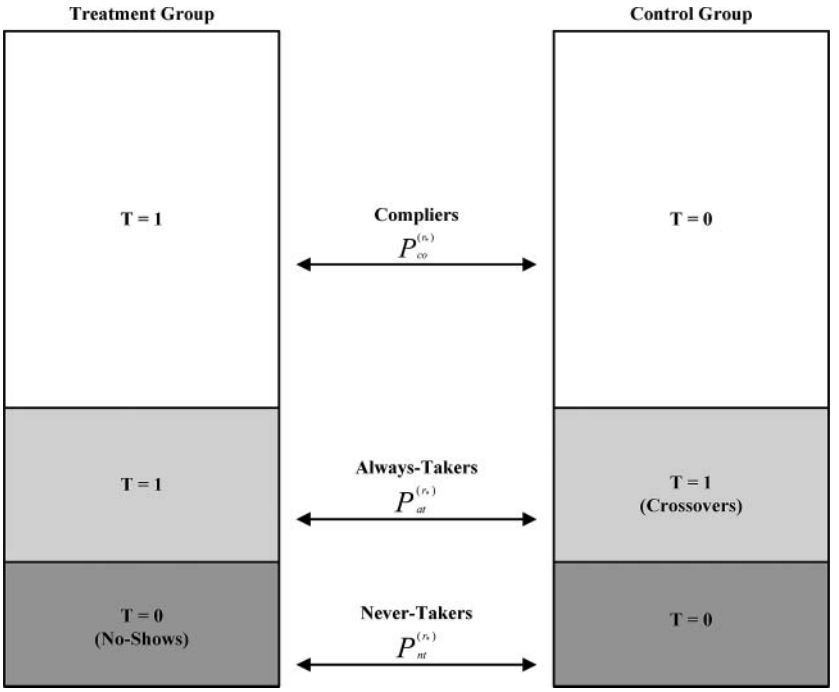


Figure 4. The Angrist, Imbens, and Rubin Causal Categories (Absent Defiers).

of a treatment group is composed of never-takers, then 10% of its control group also should be composed of never-takers. This result derives from the fact that randomization or a valid regression discontinuity design produces a treatment group and a control group that are theoretically the same in all ways, which includes their distribution of members across Angrist, Imbens, and Rubin's four subpopulations.

For many situations it is reasonable to assume that defiers do not exist, because making someone eligible for a treatment is unlikely to reduce their chances of receiving it.[11] Without defiers, the treatment population and control population for a regression discontinuity design consist of compliers, always-takers and never-takers in proportions $P_{co}^{(r)}$, $P_{at}^{(r)}$, and $P_{nt}^{(r)}$, respectively. The superscript (r) indicates that these proportions can vary with ratings. If they vary, it seems reasonable to assume that they do so continuously. Hence their limits from below and above the cut-point equal their values at the cut-point, $P_{co}^{(r*)}$, $P_{at}^{(r*)}$, and $P_{nt}^{(r*)}$.

Figure 4 illustrates this situation. The treatment-population bar in the figure represents compliers (T = 1), always-takers (T = 1), and never-takers (T = 0). The control-population bar represents compliers (T = 0), always-takers (T = 1), and never-takers (T = 0). (Never-takers in the treatment group are no-shows and always-takers in the control group are crossovers.)

[11]Angrist et al. (1996) called this assumption monotonicity, implying that the probability of receiving treatment is a monotone function of treatment assignment. Hahn et al. (2001) demonstrated how monotonicity enables fuzzy regression discontinuity designs to identify LATEs at a cut-point (LATEC).

Because compliers are the only subgroup whose treatment receipt is affected by assignment, they are the only subgroup that contributes to the treatment contrast. All of the observable difference between expected outcomes for the treatment and control populations ($\bar{Y}^+ - \bar{Y}^-$) is therefore due to compliers. Consequently ($\bar{Y}_3^+ - \bar{Y}_3^-$) is a weighted average of the mean effect of treatment on compliers at the cut-point (LATEC) and zero effect for always-takers and for never-takers, with weights equal to $P_{co}^{(r*)}$, $P_{at}^{(r*)}$ and $P_{nt}^{(r*)}$, respectively. In symbols:

$$\begin{aligned} \bar{Y}_3^+ - \bar{Y}_3^- &= P_{co}^{(r*)} \cdot \text{LATE}_\text{C} + P_{at}^{(r*)} \cdot 0 + P_{nt}^{(r*)} \cdot 0 \\ &= P_{co}^{(r*)} \cdot \text{LATE}_\text{C} \end{aligned} \tag{8}$$

Hence:

$$\text{LATE}_\text{C} = \frac{\bar{Y}_3^+ - \bar{Y}_3^-}{P_{co}^{(r*)}} \tag{9}$$

Because compliers and always-takers in the treatment population receive treatment, whereas only always-takers in the control population receive treatment:

$$\bar{T}_3^+ = P_{co}^{(r*)} + P_{at}^{(r*)} \tag{10}$$

$$\bar{T}_3^- = P_{at}^{(r*)} \tag{11}$$

Substituting Equations 10 and 11 into Equation 9 yields

$$\text{LATE}_\text{C} = \frac{\bar{Y}_3^+ - \bar{Y}_3^-}{\bar{T}_3^+ - \bar{T}_3^-} \tag{12}$$

To illustrate this point, add to our numeric example the fact that 15% of control group members receive treatment ($\bar{T}^- = 0.15$) and assume that this raises the expected outcome of the control group at the cut-point to 475 points. Now recall that 80% of the treatment group members at the cut-point received treatment ($\bar{T}^+ = 0.80$) and the expected outcome of the treatment group at the cut-point is 495 points. Substituting these facts into Equation 12 yields

$$\text{LATE}_\text{C} = \frac{495 - 475}{0.8 - 0.15} = 30.77$$

Treatment increases the outcomes of compliers by 30.77 points, on average. Because not all treatment recipients are compliers, the local average treatment effect at the cut-point (LATEC) does not necessarily equal the average effect of treatment on the treated at the cut-point (TOTC). And because compliers are only a portion of the target population, LATEC does not necessarily equal the average treatment effect at the cut-point (ATEC).

PART 3: HOW REGRESSION DISCONTINUITY ANALYSIS ESTIMATES TREATMENT EFFECTS FROM DATA FOR SAMPLES

This section considers how regression discontinuity analysis estimates treatment effects from sample data. It first describes graphical estimation approaches, then parametric statistical estimation approaches, and then nonparametric statistical estimation approaches. In addition the section provides a brief introduction to validation and robustness tests for regression discontinuity analysis.

Graphical Analysis

A major strength of regression discontinuity designs is their suitability for graphical analysis. This is because what you see is what you get. However a major limitation of graphical analysis is that much is in the eye of the beholder. This is because different analysts can interpret the same finding differently. Nevertheless, the first step in a regression discontinuity analysis should be to plot the data.

It is best to begin by plotting the probability of receiving treatment as a function of ratings (the top graph in Figure 5). Only if there is a discontinuity at the cut-point in this probability is there a treatment contrast to test. If there is a discontinuity (as shown in the top graph of Figure 5, where the horizontal line breaks at the cut-point), the next step is to examine the relationship between outcomes and ratings.

To do so, one could plot the value of the outcome for each data point on the vertical axis of a graph against the corresponding value of the rating on the horizontal axis. The second graph in Figure 5 illustrates such a plot for a downward-sloping outcome/rating relationship that has a pronounced upward shift in outcomes at the cut-point. The upward shift in outcomes, or the discontinuity, at the cut-point is the effect of the shift in the probability of receiving treatment at the cut-point. Note that individual data points in the graph bounce around a lot. In other words, a plot of individual data points is typically quite noisy.

The third graph in Figure 5 represents a simple first step toward summarizing the information contained in the individual data points. This approach divides the distribution of ratings into equal-size intervals, or bins, computes the mean outcome for each bin and plots mean outcomes at the center of each bin. Doing so smooths the data, making them less noisy.

The fourth graph in the figure smooths the data even further by using fewer bins with a larger bin width for each. Given a total sample size, larger bin widths imply more observations per bin, which produces less noise in the plotted points. However, larger bin widths provide less specificity about the likely functional form of the underlying relationship between outcomes and ratings (not shown in the graphs). Thus choosing a bin width for graphing regression discontinuity data involves making a trade-off between noise and specificity. Lee and Lemieux (2009) provided useful guidelines for making this trade-off, although it is ultimately a matter of judgment and taste.

Parametric Statistical Analysis

The next step in a regression discontinuity analysis is to fit a statistical model to the data. Parametric methods do so based on a specific functional form for the outcome/rating

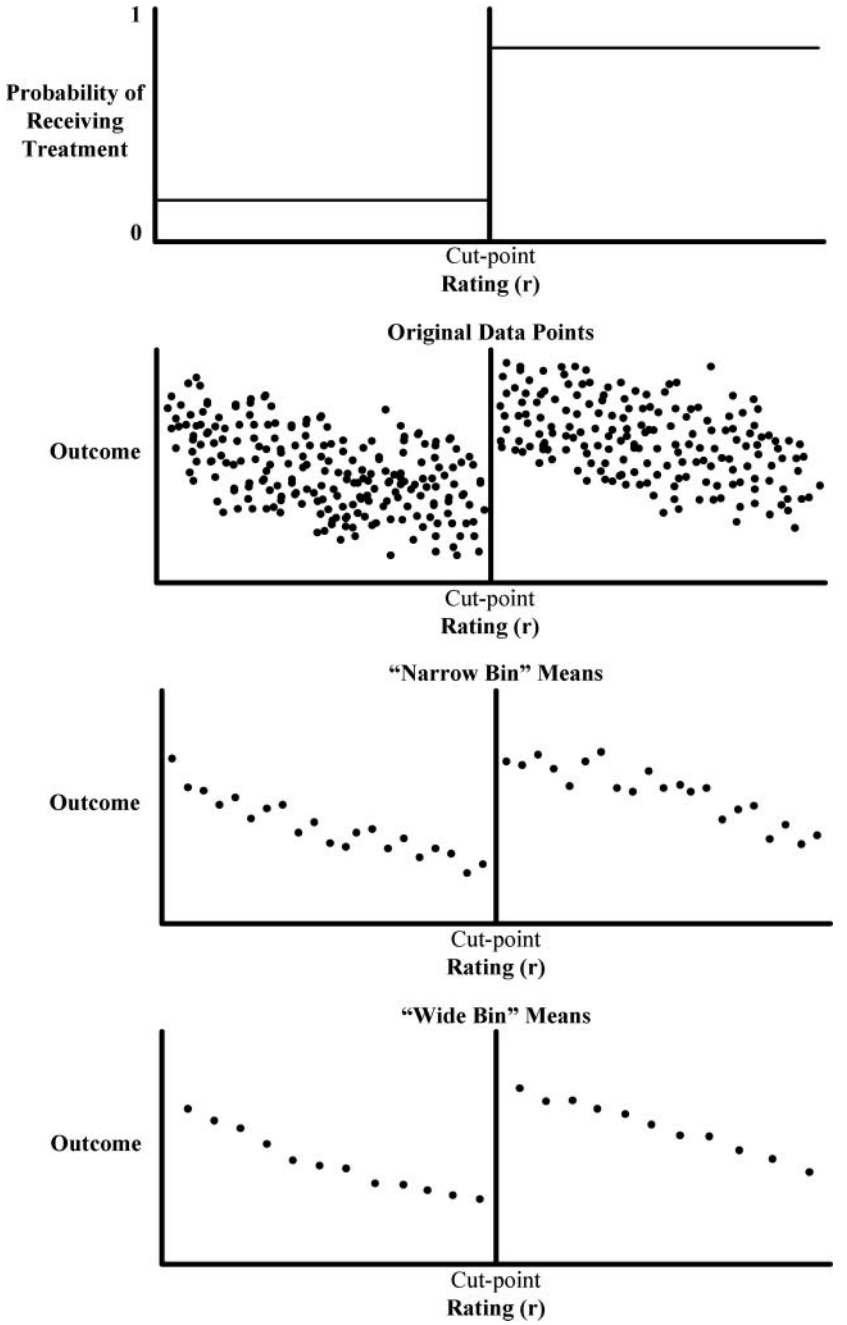


Figure 5. A Graphical Regression Discontinuity Analysis.

relationship. This functional form can range from a simple linear regression to complex nonlinear models.

A Simple Linear Regression. The top graph in Figure 6 illustrates the simplest parametric model for a regression discontinuity analysis: a linear regression with a constant slope and

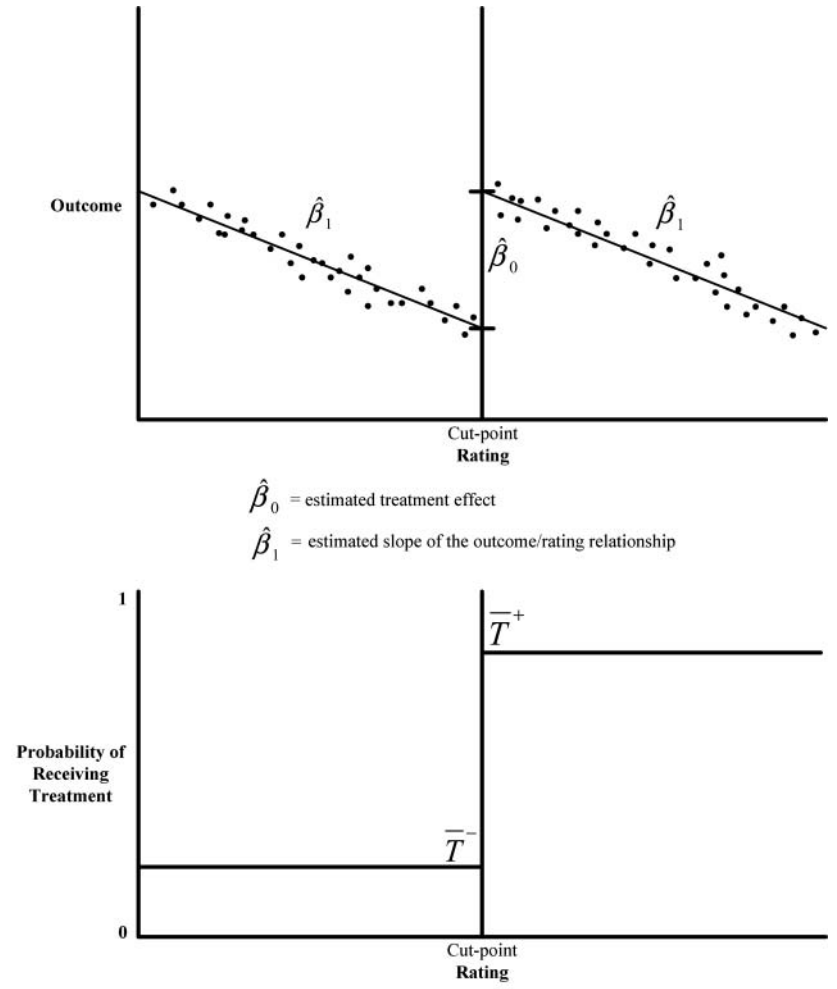


Figure 6. A Simple Linear Regression Discontinuity Analysis.

an intercept shift at the cut-point.

$$Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \eta_i \tag{13}$$

where:

Y_i = the outcome,
T_i = one for treatment group members and zero for control group members,
r_i = the rating,
η_i = a random error that is distributed independently and identically.

In Figure 6, $\hat{\beta}_1$ is the estimated slope of the outcome/rating relationship and $\hat{\beta}_0$ is the estimated intercept shift at the cut-point. Hence, $\hat{\beta}_0$ is the estimated effect of assignment to treatment or ITTC. The standard error and statistical significance (p value) of $\hat{\beta}_0$ are the standard error and statistical significance of the estimated ITTC.

The bottom graph in Figure 6 indicates that the probability of treatment group members receiving treatment ($\bar{T}^+$) is less than 1—that is, no-shows exist—and the probability of control group members receiving treatment ($\bar{T}^-$) is greater than zero—that is, crossovers exist. Assuming no defiers, the LATEC can be estimated as

$$\widehat{\text{LATE}}_{\text{C}} = \frac{\hat{\beta}_0}{\bar{T}^+ - \bar{T}^-} \tag{14}$$

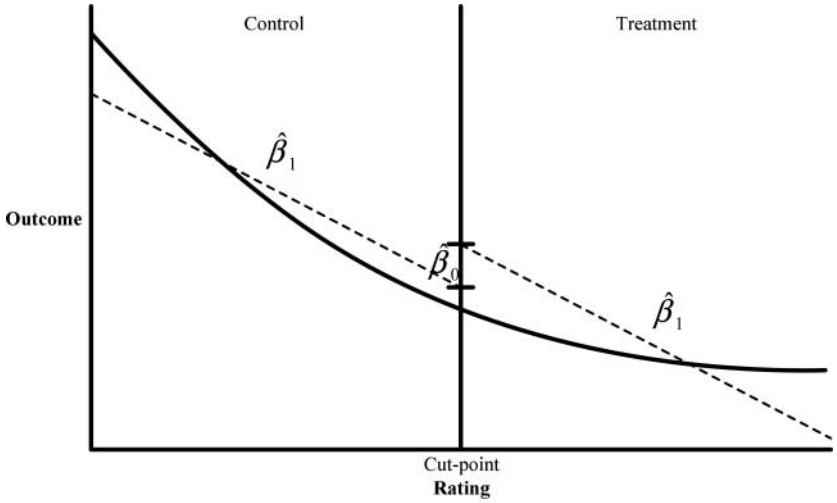
The standard error for $\widehat{\text{LATE}}_{\text{C}}$ approximately equals that for $\hat{\beta}_0$ divided by $(\bar{T}^+ - \bar{T}^-)$. The statistical significance (p value) of $\widehat{\text{LATE}}_{\text{C}}$ therefore is approximately equal to that for $\hat{\beta}_0$.

Without stronger assumptions, it is not possible to estimate the TOTC or the overall ATEC.

Functional Form. The most serious limitation of regression discontinuity designs is their possible sensitivity to misspecification of the functional form for the outcome/rating relationship. For example, if the true functional form is highly nonlinear, a simple linear model can produce misleading results. Figure 7 illustrates this situation. The solid curve in the figure denotes a true relationship that descends at a decreasing rate and passes continuously through the cut-point with no effect from the treatment.

Dashed lines in the figure represent a simple linear regression fit to data generated by the true curve. Imposing a constant slope ($\hat{\beta}_1$) for the treatment group and control group understates the average magnitude of the control-group slope and overstates the average magnitude of the treatment-group slope. This creates an apparent shift at the cut-point, which gives the mistaken impression of a discontinuity in the true function.

There are two theoretical reasons for a nonlinear relationship between outcomes and ratings. One is that the relationship between mean counterfactual outcomes and ratings is nonlinear, perhaps because of a ceiling effect or a floor effect given the nature of the



NOTE: The solid curve denotes a true relationship that descends at a decreasing rate.
The dashed lines represent a simple linear regression fit to data generated by the curve

Figure 7. Regression Discontinuity Estimation with an Incorrect Functional Form.

measure used; the other is that treatment effects vary systematically with ratings. For example, candidates with the highest ratings might experience the largest (or smallest) treatment effects. However, because regression discontinuity analyses are seldom if ever guided by theory that is powerful enough to accurately predict such nuances, choosing a functional form is typically an empirical task.

The following models are often used for parametric regression discontinuity analyses. When estimating them, in order to locate the intercept and its shift at the cut-point, ratings should be centered on the cut-point (i.e. measured as deviations).[12] Covariates can be added but usually are a secondary consideration.

1. Separate linear models for the treatment group and control group,

$$Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \beta_2 \cdot r_i \cdot T_i + \eta_i \tag{15}$$

2. A quadratic, cubic, or other polynomial,

$$Y_i = \alpha + \beta_0 \cdot T_i + \sum_{i=1}^{m} \beta_m \cdot r_i^m + \eta_i \tag{16}$$

Separate quadratic, cubic, or other polynomials for the treatment and control groups, such as

$$Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \beta_2 \cdot r_i^2 + \beta_3 \cdot r_i \cdot T_i + \beta_4 \cdot r_i^2 \cdot T_i + \eta_i \tag{17}$$

Decisions about functional form should take into account how tightly ratings are distributed around the cut-point. For example, Jacob and Lefgren (2004) used a regression discontinuity analysis to estimate the effect of mandatory summer school and grade retention on students in Chicago who do not pass an end-of-year examination. Given the tens of thousands of students tested, the authors were able to base their analysis on a large sample of students with test scores very near the cut-point. A simple linear model is likely to be adequate for situations like this, in which there is a very small interval of ratings, because even nonlinear functions approach linearity as the interval they span approaches zero width. If ratings vary widely, however, nonlinearities may be more pronounced and thus more important to model.

Nonparametric Statistical Analysis

With economists' revival of regression discontinuity analysis came the use of nonparametric statistical methods for such analyses. The two main nonparametric approaches used are kernel regression and local linear (or polynomial) regression. The flexibility of these methods enables them to accommodate many nonlinear relationships. Nevertheless, they have important limitations and should be viewed as "a complement to—rather than a substitute for—parametric estimation" (Lee & Lemieux, 2009, p. 4).

[12]In addition, centering values of the rating at the cut-point has important analytic implications for nonlinear regression discontinuity models, the standard errors of which can be badly distorted otherwise.

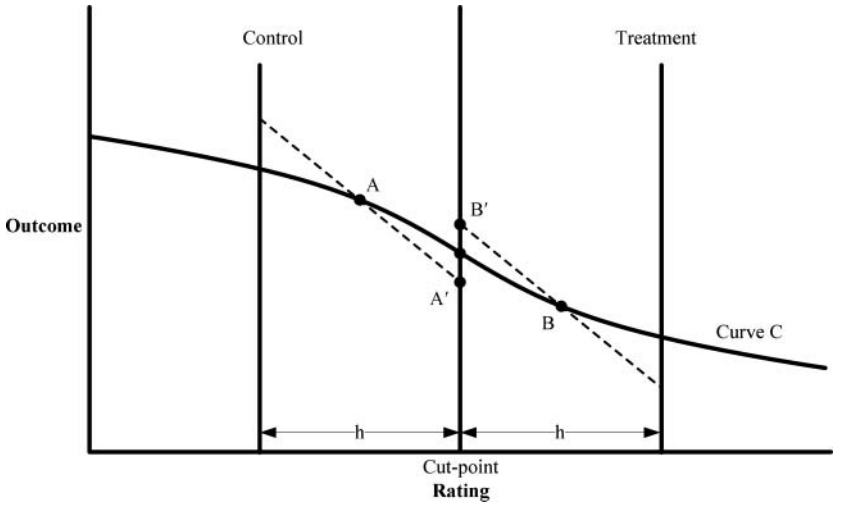


Figure 8. Boundary Bias from Kernel Regression Versus Local Linear Regression (Given Zero Treatment Effect).

The simplest form of nonparametric estimation is kernel regression. A kernel is a weighting function used to compute mean outcomes. These weights are nonzero within a given bin and zero outside of it, with a pattern within bins that depends on the type of kernel used. For example, a rectangular kernel weights all observations in a bin the same and an Epanechnikov kernel weights observations in a bin as an inverted U-shaped function of their distance from its center. Treatment effects are often estimated as the difference between mean outcomes for the treatment and control bins located immediately adjacent to the cut-point.[13]

Unfortunately kernel regression has poor boundary properties (Fan, 1992; Hahn et al. 2001; Hardle & Linton, 1994), which causes it to produce biased estimates of treatment effects. Figure 8 illustrates this problem for a downward sloping regression function with no treatment effect (the solid curve). The figure focuses on two bins of equal bandwidth (h) located immediately to the left and right of a cut-point.[14] Point A represents the mean outcome (in expectation) for the control bin, and Point B represents the mean outcome (in expectation) for the treatment bin. Therefore $(B - A)$ equals the expected value of the estimated treatment effect. This value is positive even though treatment has no effect. Hence a kernel regression with bandwidth h produces a biased estimator. As the bandwidth gets smaller, the bias gets smaller, but that bias still can be substantial.

To reduce this boundary bias it is recommended that local linear (or polynomial) regression be used (Hahn et al., 2001; Imbens & Lemieux, 2008; Lee & Lemieux, 2009). A local linear regression is estimated separately for each bin in a sample. The regression can be weighted (e.g., using a kernel) or unweighted. For many regression discontinuity analyses, treatment effects are estimated from local linear regressions for the two bins adjacent to the cut-point. Figure 8 illustrates this situation in terms of expected values for local linear regressions in the control and treatment bins. The intercept for the control regression (A') estimates the mean cut-point outcome without treatment, and the intercept for the

[13]This approach has many variations.

[14]The width of the two bins that are adjacent to the cut-point is typically referred to as a bandwidth.

treatment regression (B') estimates the mean cut-point outcome with treatment. ($B' - A'$) is therefore the expected value of the estimated treatment effect, which is nonzero and thus biased. However, its bias is much smaller than that for kernel regression.

In addition to the problems that exist for parametric regression discontinuity estimation, there are two further limitations of nonparametric regression discontinuity estimation. First is the need for very large samples to provide an adequate number of observations in the two bins adjacent to the cut-point. Second is the potential sensitivity of nonparametric regression discontinuity estimates to the choice of a bandwidth. Choosing a bandwidth for nonparametric estimation involves making a choice between introducing bias from a bandwidth that is too wide and losing precision from a bandwidth that is too narrow. The most widely used empirical approach for making such a trade-off is cross-validation (Imbens & Lemieux, 2008; Lee & Lemieux, 2009). This approach assesses the ability of a given nonparametric estimator with a given bandwidth to predict outcomes for each sample observation.[15] The bandwidth with the greatest predictive power is chosen for estimating the treatment effect.[16] The more robust regression discontinuity findings are in relation to differences in bandwidth, the more confidence one can place in them. However, there is no foolproof way to know when bias has been reduced to an acceptable level.

Validation and Robustness Tests. A regression discontinuity analysis, be it parametric or nonparametric, should include tests of the validity of the design used, the validity of the estimation model used, and the robustness of the findings obtained (Imbens & Lemieux, 2008; Lee & Lemieux, 2009). McCrary (2008) presented a simple test of the validity of a regression discontinuity design that assesses whether ratings were manipulated by examining the pattern of density of observations. If ratings were not manipulated, the density of observations should vary continuously with ratings at the cut-point. If ratings were manipulated, there could be a marked increase in their density on one side of the cut-point and a marked decrease on the other side. A graphical version of this test would subdivide ratings into bins of constant width and plot the number of candidates for each. McCrary (2008) presented formal tests based on this logic.

A discontinuity in the density of ratings at the cut-point indicates that they probably were manipulated. But the absence of a discontinuity does not necessarily indicate that ratings were not manipulated. This is because ratings could have been manipulated in a way that substitutes specific candidates in the control group for an equal number of specific candidates in the treatment group. Hence, there could be a discontinuity in candidate characteristics without a discontinuity in their density.

The most important test of the internal validity of a regression discontinuity estimation model is whether the model suggests a discontinuity at the cut-point in baseline characteristics for treatment-group and control-group members. This test is analogous to comparing treatment-group and control-group baseline characteristics for a randomized trial. If randomization is executed properly, there should be few, if any, large or statistically significant treatment/control group baseline differences. Likewise, if a regression discontinuity estimation model is internally valid, there should be few, if any, large or statistically significant treatment/control group baseline discontinuities. The mechanics of the test are

[15]Cross-validation typically estimates a given model for a given bandwidth omitting a sample observation and predicting the outcome for the missing observation from the estimated model. This process is repeated for each observation, and results are pooled across observations.

[16]The optimal bandwidth for nonparametric regression discontinuity estimation may not be the same as that for graphical regression discontinuity analysis.

straightforward. Simply estimate a parametric or nonparametric model of interest using each of a series of baseline characteristics as its dependent variable. The coefficient for the treatment indicator in the model measures the discontinuity for the baseline characteristic, and the statistical significance of the estimated coefficient indicates the significance of the discontinuity.

There are numerous tests of the robustness of regression discontinuity findings to variations in the estimation procedure used, the sample included, the functional form specified (for parametric approaches), and/or the bandwidth chosen (for nonparametric approaches). These tests compare findings produced by plausible alternative approaches to see how stable they are across approaches. The more stable findings are, the more confident one can be that the findings are not a methodological artifact.

For example, one might report parametric findings for alternative functional forms, nonparametric findings for alternative bandwidths, and/or both types of findings for samples that omit varying numbers of observations with the highest and lowest ratings (which is to say, for samples that trim outliers). The less these findings vary, the more confidence one can have in them.

Precision of the Estimates. Another important issue to consider when assessing the quality of a parametric or nonparametric regression discontinuity analysis is the precision of its estimates. The precision of estimated treatment effects typically is expressed as a minimum detectable effect (MDE) or minimum detectable effect size (MDES). An MDE is the smallest treatment effect that a research design has an acceptable chance of detecting, if it exists. MDEs are reported in natural units, such as scale-score points for tests, dollars for earnings, or percentage points for recidivism. An MDES is an MDE divided by the standard deviation of the outcome measure in the absence of treatment. It is reported in units of standard deviations.[17]

An MDE or MDES is typically defined as the smallest true treatment effect (or effect size) that has an 80% chance (80% power) of producing an estimated treatment effect that is statistically significant at the .05 level for a two-sided hypothesis test. This parameter is a multiple of the standard error of a treatment-effect estimator. The multiple depends on the number of degrees of freedom available (Bloom, 1995), but for more than about 20 degrees of freedom its value is roughly 2.8, assuming a two-tail hypothesis test with 80 percent power at the conventional 0.05 level of statistical significance.

Because most parametric regression discontinuity analyses have more than 20 degrees of freedom, their MDE or MDES can be approximated as follows:[18]

Minimum Detectable Effect

$$\text{MDE} \approx 2.8\sqrt{\frac{\left(1 - R_Y^2\right)\sigma_Y^2}{N \cdot P(1-P)\left(1 - R_T^2\right)}} \tag{18}$$

or

[17]Effect sizes are used to report treatment effects in education research, psychology, and other social sciences (see, e.g., Cohen, 1988; Grissom & Kim, 2005; Rosenthal, Rosnow, & Rubin, 2000).

[18]This expression is more complex for clustered regression discontinuity designs (Schochet, 2008). The degree of complexity is parallel to that for clustered randomized trials (see, e.g., Bloom, 2005; Bloom, Richburg-Hayes, & Black, 2007).

Minimum Detectable Effect Size

$$\text{MDES} \approx 2.8\sqrt{\frac{1 - R_Y^2}{N \cdot P(1-P)\left(1 - R_T^2\right)}} \tag{19}$$

where:

R_Y^2 = the proportion of variation in the outcome (Y) predicted by the rating and other covariates included in the regression discontinuity model,
R_T^2 = the proportion of variation in treatment status (T) predicted by the rating and other covariates included in the regression discontinuity model,
N = the total number of sample members,
P = the proportion of sample members assigned to the treatment group,
σ_Y^2 = the counterfactual variance of the outcome.

Choosing a target MDE or MDES requires considerable judgment and is beyond the scope of the present paper.[19] To gain some perspective on this issue it is useful to compare the precision of a standard parametric regression discontinuity design to that of a randomized trial. To make this comparison a fair one, assume that the two designs have the same total sample size (N), the same treatment/control group allocation (P vs. (1-P)), the same outcome measure (Y), and the same variance of the counterfactual outcome (σ_Y^2). In addition, assume that the rating is the only covariate for the regression discontinuity design and the randomized trial. (The rating might be a pretest used to increase a trial's precision). Hence, the ability of the covariate to reduce unexplained variation in the outcome (R_Y^2) is the same for both designs.

A randomized trial with the rating as a covariate would use the same regression model as a regression discontinuity design to estimate treatment effects (Equation 13). The MDE or MDES of the trial therefore can be expressed by Equations 18 and 19 for regression discontinuity designs. The only difference between the regression discontinuity design and an otherwise comparable randomized trial is the value of R_T^2, which is zero for a randomized trial and nonzero for a regression discontinuity analysis. This difference reflects the difference between the assignment processes of the two designs. The ratio of their MDEs or MDESs is therefore

$$\frac{\text{MDE}_{\text{RD}}}{\text{MDE}_{\text{randomized}}} = \frac{1}{\sqrt{1 - R_T^2}} = \frac{\text{MDES}_{\text{RD}}}{\text{MDES}_{\text{randomized}}} \tag{20}$$

R_T^2 represents the collinearity (or correlation squared) that exists between the treatment indicator and rating in a regression discontinuity design.[20] This collinearity depends on how ratings are distributed around the cut-point (Bloom, Kemple, Gamse, & Jacob, 2005; Goldberger, 1972; Schochet, 2008). Figure 9 illustrates two possibilities: a balanced uniform distribution and a balanced normal distribution. A uniform distribution would exist if ratings were expressed in rank-order without ties. A normal distribution might exist if ratings were scores on a test because test scores often follow a normal distribution. A balanced

[19]Bloom, Hill, Black, and Lipsey (2008) and Hill, Bloom, Black, and Lipsey (2008) presented an analytic approach and empirical benchmarks for choosing MDESs in education research.

[20]For a simple linear regression discontinuity model, this collinearity coefficient is the *R*-squared of a regression of the treatment indicator on the rating.

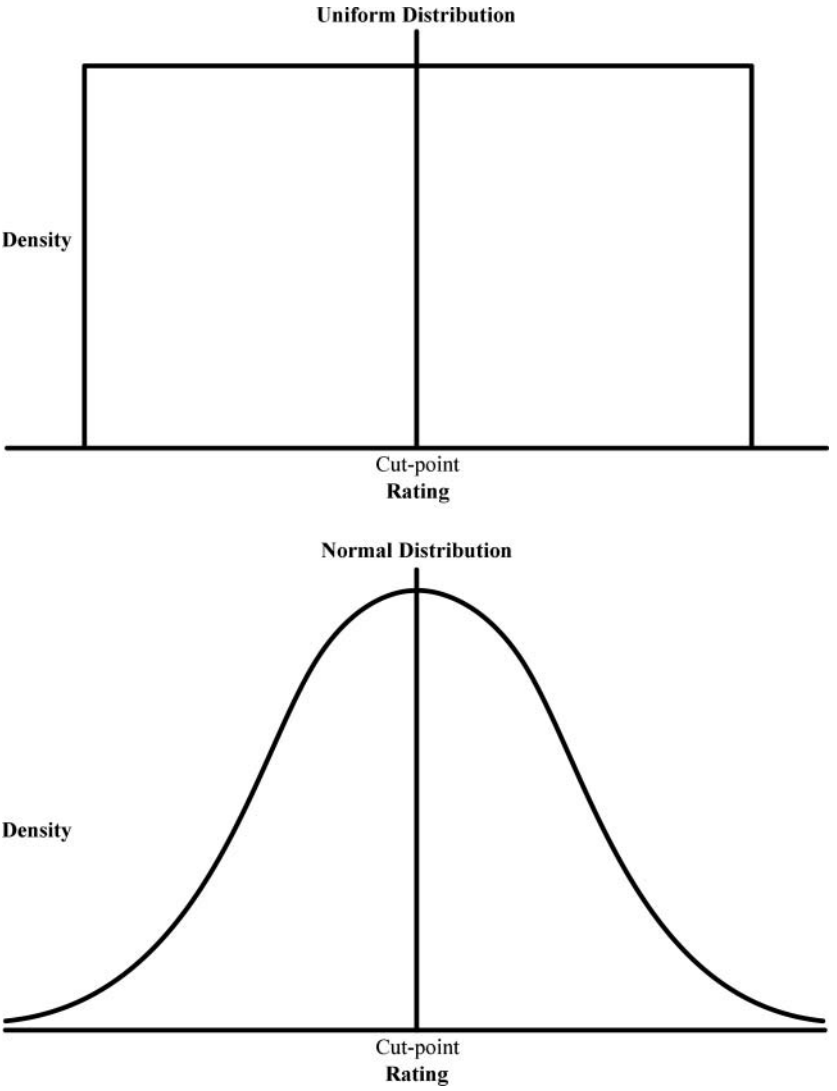


Figure 9. Alternative Distributions of Ratings.

distribution is one that is centered on the cut-point, so that half of the observations are on one side and half are on the other side. The degree of imbalance of a distribution reflects its mix of treatment and control candidates.

To compute R_T^2 for a given distribution of ratings one can generate ratings (r) from a distribution of interest, attach the appropriate value of the treatment indictor (T) to each rating and regress T on r. Doing so yields an R_T^2 of 0.750 for a balanced uniform distribution and 0.637 for a balanced normal distribution. Substituting these values into Equation 20 indicates that the MDE or MDES for a regression discontinuity design with a balanced uniform distribution of ratings is twice that for an otherwise comparable randomized trial. This multiple is 1.66 for a balanced normal distribution of ratings.

Table 1. Collinearity coefficient and sample size multiple for a regression discontinuity design relative to an otherwise comparable randomized trial

	Balanced Design[a]		Unbalanced Design[b]	
Regression Discontinuity Model	Uniform Rating Distribution	Normal Rating Distribution	Uniform Rating Distribution	Normal Rating Distribution
Collinearity coefficient (R_T^2)				
Simple linear	0.750	0.637	0.663	0.593
Simple quadratic	0.750	0.637	0.791	0.651
Simple cubic	0.859	0.744	0.808	0.716
Separate treatment/ control linear	0.750	0.637	0.750	0.632
Separate treatment/ control quadratic	0.889	0.802	0.828	0.743
Sample size multiple				
Simple linear	4.00	2.75	2.97	2.46
Simple quadratic	4.00	2.75	4.79	2.86
Simple cubic	7.11	3.91	5.22	3.52
Separate treatment/ control linear	4.00	2.75	4.00	2.72
Separate treatment/ control quadratic	9.00	5.04	5.80	3.89

Note. Regression discontinuity models:

Simple linear $Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \eta_i$

Simple quadratic $Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \beta_2 \cdot r_i^2 + \eta_i$

Simple cubic $Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \beta_2 \cdot r_i^2 + \beta_3 \cdot r_i^3 + \eta_i$

Separate treatment/ control linear $Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \beta_2 \cdot r_i \cdot T_i + \eta_i$

Separate treatment/ control quadratic $Y_i = \alpha + \beta_0 \cdot T_i + \beta_1 \cdot r_i + \beta_2 \cdot r_i^2 + \beta_3 \cdot r_i \cdot T_i + \beta_4 \cdot r_i^2 \cdot T_i + \eta_i$

[a] $p = .5$. [b] $p = .33$ or .67.

Equation 21 provides an expression for the "sample size multiple" required for a regression discontinuity design to produce the same MDE or MDES as an otherwise comparable randomized trial.

$$\frac{\mathrm{N_{RD}}}{\mathrm{N_{randomized}}} = \frac{1}{1 - R_T^2} \tag{21}$$

This expression indicates, for example, that a regression discontinuity sample with a balanced uniform distribution of ratings must be $(\frac{1}{1-0.75})$, or four times that for an otherwise comparable randomized trial. The multiple is $(\frac{1}{0.64})$, or 2.75, for a balanced normal distribution of ratings.[21]

Table 1 presents collinearity coefficients and sample size multiples for several regression discontinuity models and distributions of ratings. The first two columns in the table

[21] Goldberger (1972) proved this finding for a balanced normal distribution of ratings.

are for a balanced uniform and normal distribution of ratings, respectively. The second two columns are for an unbalanced uniform and normal distribution of ratings, respectively. The unbalanced distributions have a third of the sample on one side of the cut-point and two thirds on the other side. The top panel of the table reports a collinearity coefficient or R_T^2 for each situation, and the bottom panel reports the corresponding sample size multiple for a regression discontinuity design relative to an otherwise comparable randomized trial. Each row in the table represents a different parametric regression discontinuity model or functional form. Findings in the table indicate the following:

1. *The precision of a regression discontinuity design is much less than that of an otherwise comparable randomized trial*. At best, a regression discontinuity sample must be 2.72 times that of its randomized counterpart to achieve the same precision. At worst, this multiple could be appreciably larger.
2. *The precision of a regression discontinuity design erodes as the complexity of its estimation model increases*. Consequently it is essential to use the simplest model possible. Nevertheless, in some cases complex models may be needed. If so, precision is likely to be limited.

The precision of a regression discontinuity design depends on the distribution of ratings around the cut-point.[22] For example, a uniform distribution reduces precision to a greater degree than a normal distribution—especially for complex regression discontinuity models.

Because of the enormous flexibility and variety in implementation of nonparametric statistical methods for regression discontinuity analyses, it is not clear how to summarize the precision of such methods. What is clear, however, is that because they rely mainly, and often solely, on observations very near the cut-point (ignoring or greatly down-weighting all other observations), nonparametric methods are far less precise than parametric methods for a given study sample.

PART 4: GENERALIZING REGRESSION DISCONTINUITY FINDINGS

Having constructed a regression discontinuity design that identifies a treatment effect of interest and having developed an appropriate estimation strategy for this design, the next step is to consider the likely generalizability of the design's findings.

A Strict-Constructionist View

One often reads that regression discontinuity findings apply only to candidates at a cut-point. This idea represents a strict-constructionist view, which reflects the fact that regression discontinuity designs identify treatment effects through the limiting properties of continuous functions at a point. Without further assumptions, such designs can only identify effects for

[22] Schochet (2008) illustrated this point.

candidates at the cut-point margin. These findings are sometimes referred to as marginal average treatment effects (Bjorklund & Moffit, 1987; Black, Galdo, & Smith, 2005; Heckman, 1997). A marginal average treatment effect is the average effect of a program for candidates who would be added or dropped by marginally changing the program's eligibility criterion. This parameter is relevant for decisions about expanding or contracting programs but not necessarily for decisions about opening or closing programs.

A More Expansive View

Lee (2008) provided a more expansive—and more revealing—interpretation of the population to which regression discontinuity findings generalize (also see Lee & Lemieux, 2009). His interpretation focuses on that fact that control over ratings by decision makers and candidates is typically imprecise. Thus, observed ratings have a probability distribution around an expected value or true score.[23]

Figure 10 illustrates such distributions for a hypothetical population of three types of candidates: A, B, and C. Each candidate type has a distribution of potential ratings around an expected value. The top panel in the figure represents a situation in which control over ratings is highly imprecise. Highly imprecise ratings contain a lot of random error and thus vary widely around their expected values. To simplify the discussion, without loss of generality, assume that the shapes and variances of the three distributions are the same; only their expected values differ.

The expected value of ratings, $E\{r\}$, is 3 units below the regression discontinuity cut-point for Type A candidates, 5 units above the cut-point for Type B candidates, and 7 units above the cut-point for Type C candidates. Consequently, Type A candidates are the most likely to have observed ratings at the cut-point, Type B candidates are the next most likely, and Type C candidates are the least likely. Type A candidates therefore compose the largest segment of the cut-point population, Type B candidates compose the next largest segment, and Type C candidates compose the smallest segment.

Segment sizes at the cut-point are proportional to the height of each distribution (its density) at the cut-point. Assume that distribution heights at the cut-point are 0.7 for Type A candidates, 0.2 for Type B candidates, and 0.1 for Type C candidates. Type A candidates thus compose $\frac{0.7}{0.7+0.2+0.1}$, or 0.70, of the cut-point population, Type B students compose $\frac{0.2}{0.7+0.2+0.1}$, or 0.20, and Type C candidates compose $\frac{0.1}{0.7+0.2+0.1}$, or 0.10. The cut-point population is thus somewhat heterogeneous in terms of expected ratings ($E\{r^{(A)}\}$, $E\{r^{(B)}\}$ and $E\{r^{(C)}\}$). To the extent that expected ratings correlate with expected counterfactual outcomes ($E\{Y_0{}^{(A)}\}$, $E\{Y_0{}^{(B)}\}$, $E\{Y_0{}^{(C)}\}$) the cut-point population also is somewhat heterogeneous in terms of expected counterfactual outcomes.[24]

The bottom panel in Figure 10 illustrates a situation with more precise control over ratings, which implies narrower distributions of potential values. Type C candidates, whose

[23]Modeling ratings by a probability distribution of potential values with an expected value or true score is consistent with standard practice in measurement theory. Nunnally (1967) discussed such models from the perspective of classical measurement theory, and Brennan (2001) discussed them from the perspective of generalizability theory.

[24]The mean expected counterfactual outcome for the cut-point population is an average of the expected value for each type of candidate weighted by the proportion of the cut-point population each type composes.

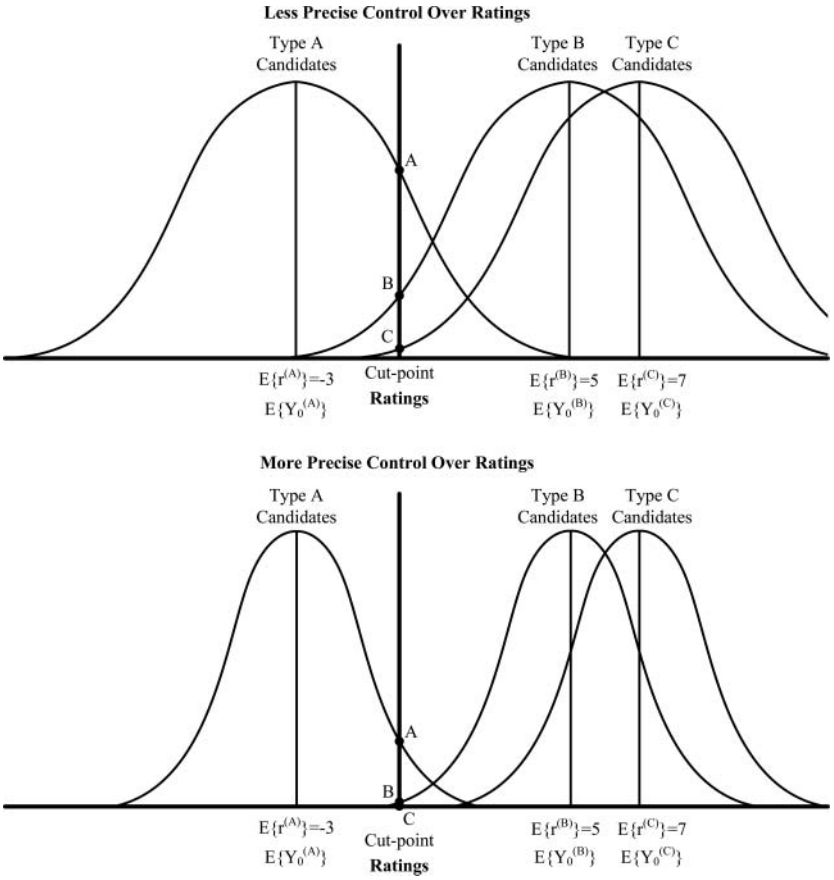


Figure 10. How Imprecise Control Over Ratings Affects the Distribution of Counterfactual Outcomes at Cut-Point of a Regression Discontinuity Design.

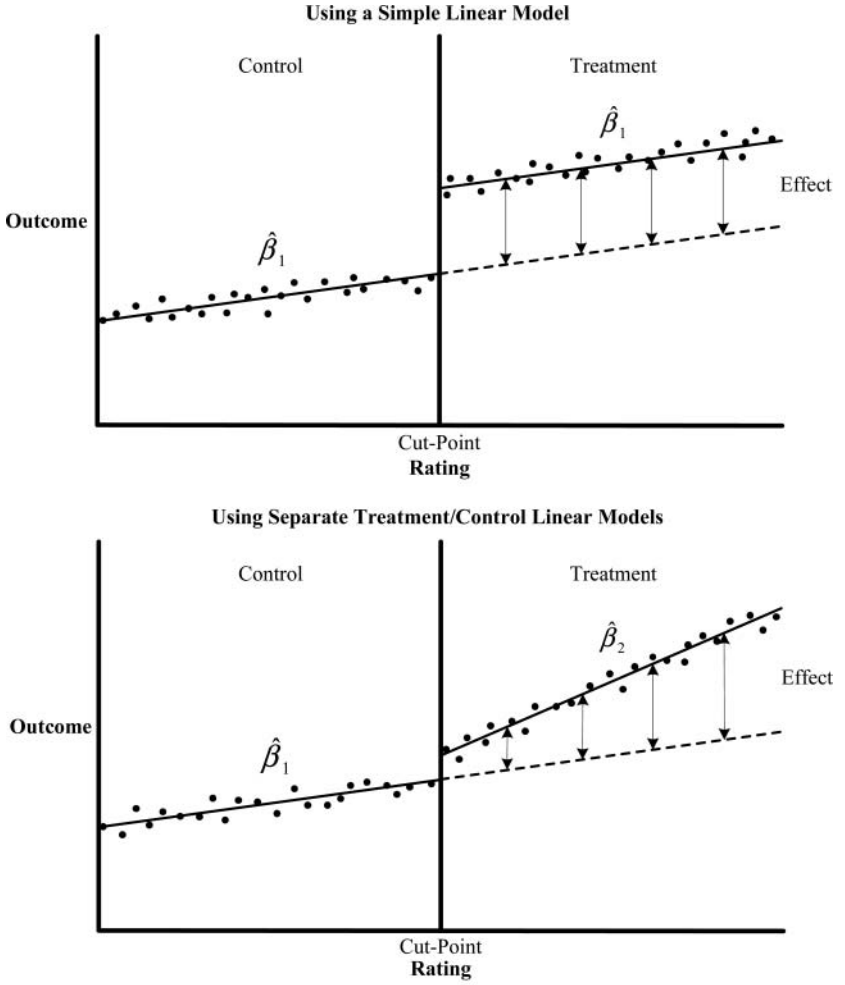
expected rating is furthest from the cut-point, are extremely unlikely to have observed ratings at the cut-point. Because of this, they represent a very small proportion of the cut-point population. Type B candidates also represent a very small proportion of the cut-point population, but one that is larger than that for Type C candidates. The cut-point population thus is comprised almost exclusively of Type A candidates, which makes it quite homogeneous.

Several important implications flow from Lee's insight about the generalizability of regression discontinuity results. First, when ratings contain random error (which is probably most of the time), the population of candidates at a cut-point is not necessarily homogenous. Second, other things being equal, the more random error that observed ratings contain, the more heterogeneous the cut-point population will be, and therefore the more broadly generalizable regression discontinuity findings will be. Third, in the extreme, if ratings are assigned randomly, then the full range of candidate types will be assigned randomly above and below the cut-point. This case is equivalent to a randomized trial and the resulting cut-point population will comprise the full target population.

An "Old-School" View: Extrapolation Beyond the Cut-Point

Much of the early work on regression discontinuity analysis reflects an even more expansive view of the generalizability of regression discontinuity findings. This view is based on a willingness to extrapolate findings beyond the cut-point using parametric statistical models.

For example, the upper panel of Figure 11 illustrates how a simple sharp linear regression discontinuity model can extrapolate (and thus generalize) estimated treatment effects. This model specifies a constant slope for the treatment group and control group plus an intercept shift between them at the cut-point. The average effect of assignment to treatment is the difference between expected outcomes for the treatment group (the solid line to the right of the cut-point) and an extrapolation of expected counterfactual outcomes for the treatment group (the dashed line to the right of the cut-point). The vertical distance between the two lines is the mean treatment effect for each value of the rating, which is constant across rating values. This is because a simple linear regression discontinuity model



NOTE: Dots represent hypothetical outcome/rating data points.

Figure 11. Extrapolating Regression Discontinuity Findings Beyond the Cut-Point.

implies that actual expected outcomes for treatment group members are parallel to their counterfactual outcomes.

The lower panel in the figure illustrates how a nonlinear parametric model can extrapolate or generalize treatment effects to candidates with ratings that do not lie at the cut-point. This figure represents a separate linear model for the treatment group and the control group. Hence, it specifies a different slope and intercept for each group. The dashed line to the right of the cut-point extrapolates expected counterfactual outcomes for the treatment group. The vertical distance between expected outcomes and expected counterfactual outcomes for a given rating is the mean treatment effect for candidates with that rating. As can be seen, the treatment effect in this example increases linearly with ratings.

To compare mean treatment effects from the two models algebraically, note that

Simple Linear Model

$$\bar{Y} = \alpha + \beta_0 \cdot T + \beta_1 \cdot r \tag{22}$$

$$\frac{d\bar{Y}}{dT} = \beta_0 \tag{23}$$

Separate T/C Linear Models

$$\bar{Y} = \alpha + \beta_0 \cdot T + \beta_1 \cdot r + \beta_2 \cdot T \cdot r \tag{24}$$

$$\frac{d\bar{Y}}{dT} = \beta_0 + \beta_2 \cdot r \tag{25}$$

Equations 22 and 24 are models of how expected outcomes ($\bar{Y}$) vary with observed ratings (r). Equations 23 and 25 present the first derivatives of these models with respect to treatment (T), which equal the effect of treatment on average outcomes. Equation 23 indicates that the average effect of treatment is the same (β_0) for all ratings. This is the constant vertical distance between actual and counterfactual expected outcomes in the top panel of Figure 11. Equation 25 indicates that the average effect of treatment is a linear function of ratings ($\beta_0 + \beta_2 \cdot r$). This is the varying vertical distance in the bottom panel of Figure 11.

As the degree of technical sophistication with respect to conducting regression discontinuity analyses has increased over time, the willingness of researchers to make the kinds of extrapolations as just illustrated has decreased. This in part reflects an appropriate degree of caution with respect to extrapolating findings beyond the center of one's data, given the uncertainty that exists when doing so. Furthermore, such extrapolations involve extra uncertainty for regression discontinuity analyses, because the analyses depend heavily on the functional form assumed for the outcome/rating relationship. Nevertheless, given the potential practical importance of such extrapolations, they should be considered for regression discontinuity analyses and reported if the pattern upon which they are based seems clear. However, any such findings should be qualified in order to make others aware of the assumptions upon which they are based.

PART 5: FUTURE REGRESSION DISCONTINUITY RESEARCH

This final section highlights some important frontiers for future regression discontinuity research. One frontier involves regression discontinuity analyses with multiple cut-points

and/or ratings. The simplest example of such a situation occurs when study sites assign candidates based on separate local ratings and cut-points. The issues here are (a) whether to pool regression discontinuity data or findings across sites; (b) if so, how to do so; and (c) how to interpret pooled findings. These issues were a major concern for the federal Reading First Impact Study, which was based on 17 regression discontinuity designs plus a cluster-randomized trial (Gamse et al., 2008). Each regression discontinuity site in that study established its own ratings and its own cut-points for choosing schools to participate in Reading First.

Some researchers consider a situation like this to be a curse; however, it is more likely to be a blessing. The purported curse concerns the issue of whether and how to pool across sites. Pooling the data across sites (especially graphically) may present some problems, but as long as the outcome measure is comparable across sites, there is no problem pooling their findings, much like in a standard meta-analysis.

The blessing represented by site-specific regression discontinuity designs is their ability to provide findings that are more broadly generalizable than those for a single regression discontinuity analysis. This is because the cut-point subpopulation for each site can vary. Thus, their pooled findings can represent a heterogeneous population. Researchers therefore should view multisite regression discontinuity designs as a promising basis for measuring treatment effects.

A more difficult situation arises when each candidate in a study is assigned to treatment based on more than one rating and cut-point (even for a single site). In other words, the candidate selection criterion is multidimensional instead of unidimensional. This process is used frequently in education (Cook et al., 2009; Gill, Lockwood, Martorell, Setodji, & Booker, 2008; Jacob & Lefgren, 2004; Robinson & Reardon, 2009; Weinbaum & Weiss, 2009). For example, the federal No Child Left Behind law imposes sanctions on schools that fail to meet one or more criteria for achieving "adequate yearly progress" (Cook et al., 2009; Weinbaum & Weiss, 2009). In addition, some school districts require students to pass more than one test in order to be promoted (see Jacob & Lefgren, 2004).

One way to address this problem is to focus on a single criterion and eliminate sample members who are assigned to treatment by any other criteria (Jacob & Lefgren, 2004). Another approach is to focus on a single criterion (or one at a time) and consider candidates assigned to treatment and control status through other criteria to be no-shows and crossovers (Robinson & Reardon, 2009). Other approaches pool all existing information into a single analysis. At this time there is no generally accepted practice for dealing with the problem. Thus, further research is needed to help resolve this issue.

Another emerging area of regression discontinuity research involves empirical attempts to cross-validate regression discontinuity findings with those from a linked, high-quality, randomized trial. The central question here is whether regression discontinuity estimates replicate those with unimpeachable validity. This question is important because the fact that regression discontinuity designs can be valid in theory does not guarantee that they are valid in practice. Cook, Shadish, and Wong (2008) summarized the three main studies that have addressed this question to date (Aiken, West, Schwalm, Carroll, & Hsuing, 1998; Black et al., 2005; Buddelmeyer & Skoufias, 2002), the results of which are mixed but encouraging. Gleason, Resch, and Berk (2009) described initial steps for a similar study currently under way. Clearly more such research is in order to help establish a firm empirical understanding of how the validity of regression discontinuity analysis varies with the conditions under which it is conducted.

Another emerging area of regression discontinuity research involves the attempt to combine regression discontinuity analysis with other strong quasi-experimental designs.

For example, Somers, Zhu, Jacob, and Bloom (2009) have been exploring possibilities to combine regression discontinuity analysis with comparative interrupted time-series analysis to estimate the effects of Reading First on student outcomes for schools in Kentucky. Given the limited statistical precision of regression discontinuity designs, it is important to explore the use of existing time-series data to increase precision. Furthermore, time-series data might reduce the potential sensitivity of regression discontinuity analyses to the functional form upon which they are based.

Yet another emerging area of future regression discontinuity research involves attempts to improve the quality of statistical inferences (confidence intervals and hypothesis tests) based on regression discontinuity analyses. This work is motivated by Lee and Card's (2008) approach to accounting for uncertainty about specification error in regression discontinuity analyses. Their approach groups observations into clusters within the distribution of ratings, and these clusters can have a separate error component if the difference between a true regression discontinuity functional form and that used for estimation varies with the rating—that is, if specification error varies with ratings. This approach uses a hierarchical model to distinguish variance components for individual observations and clusters of observations. For large numbers of clusters, the standard errors and statistical inferences of such models are well understood. But for a small number of clusters, these properties are more difficult to ascertain analytically, and they lend themselves to resampling methods such as bootstrapping.[25] Further work is needed to explore these and related issues.

The preceding examples represent only a fraction of the likely future advances in regression discontinuity methods and their applications. As the approach is applied to more settings, it surely will be adapted and developed further.

ACKNOWLEDGMENTS

This paper was supported by Grant Number R305D09008 from the Institute of Education Sciences, by the William T. Grant Foundation and by MDRC's Judith Gueron Fund. The author thanks Pei Zhu, Marie-Andree Somers, Robin Jacob, William Corrin, Alison Rebeck Black, Dick Murnane and John Willet for their valuable comments on drafts of the paper plus Collin F. Payne for producing its figures. All short-comings are the author's responsibility.

REFERENCES

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*, 207–244.

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

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*, 533–575.

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

Berk, R. A., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, *78*, 21–27.

[25] Wooldridge (2009) discussed bootstrapping methods for obtaining standard errors.

Bjorklund, A., &Moffit, R. (1987). Estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics*, *69*, 42–49.

Black, D., Galdo, J., & Smith, J. (2005). *Estimating the selection bias of the regression discontinuity design using a tie-breaking experiment* (Working Paper). Syracuse, NY: Department of Economics and Center for Policy Research, Syracuse University.

Black, S. (1999, May). Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics*, pp. 577–599.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*, 225–246.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*, 547–556.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York, NY: Russell Sage Foundation.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328.

Bloom, H. S., Kemple, J., Gamse, B., & Jacob, R. (2005, April 14). *Using regression discontinuity analysis to measure the impacts of Reading First*. Paper presented at the annual American Educational Research Association research conference, Montreal, Canada.

Bloom, H. S., Orr, L. L., Cave, G., Bell, S. H., Doolittle, F., & Lin, W. (1997). The benefits and costs of JTPA programs: Key findings from the national JTPA study. *The Journal of Human Resources*, *32*, 549–576.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30–59. doi:10.3102/0162373707299550

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Buddelmeyer, H., & Skoufias, E. (2002, January). *Can social programs be reliably evaluated with non-experimental methods? Evidence on the performance of a regression discontinuity design using PROGRESA data* (Working Paper). New York, NY: Inter-American Development Bank.

Card, D., & Shore-Sheppard, L. D. (2004). Using discontinuous eligibility rules to identify the effects of the federal Medicaid expansions on low-income children. *The Review of Economics and Statistics*, *86*, 752–766.

Chen, M. K., & Shapiro, J. M. (2005). *Does prison harden inmates? A discontinuity-based approach* (Working Paper). New Haven, CT: Yale School of Management.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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*, 636–654.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

Cook, T., Wong, V., Steiner, P., Taylor, J., Gandhi, A., Kendziora, K., . . . Choi, K. (2009, January). *Impacts of school improvement status on students with disabilities: Feasibility report*. Washington, DC: American Institutes for Research.

DiNardo, J., & Lee, D. (2002, June). *The impact of unionization on establishment closure: A regression-discontinuity analysis of representation elections* (Working Paper No. 8993). Cambridge, MA: National Bureau of Economic Research.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, *87*, 998–1004.

Fisher, R. A. (1935). *The design of experiments*. London, UK: Oliver and Boyd.

Gamse, B. C., Bloom, H. S., Kemple, J. J., Jacob, R. T., . . . Zhu, P. (2008). *Reading First impact study: Interim report* (NCES 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Educational Sciences, U.S. Department of Education.

Gill, B., Lockwood, J. R., Martorell, F., Setodji, C. M., & Booker, K. (2008). *State and local implementation of the No Child Left Behind Act*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Gleason, P., Resch, A., & Berk, J. (2009, March 6). *Replicating experimental impact estimates using a regression discontinuity design*. Mathematica Policy research presentation to the Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper No. 129-72). Madison: University of Wisconsin, Institute for Research on Poverty.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical perspective*. Mahwah, NJ: Erlbaum.

Hahn, J., Todd, P., & van der Klaauw, W. (1999, May). *Evaluating the effect of an antidiscrimination law using a regression-discontinuity design* (Working Paper No. 7131). Cambridge, MA: National Bureau of Economic Research.

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

Hardle, W., & Linton, O. (1994). Applied nonparametric methods. In R. F. Ingle & D. F. MacFadden (Eds.), *Handbook of econometrics 4* (pp. 2295–2339). Amsterdam, the Netherlands: North Holland.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*, *32*, 441–462.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

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

Imbens, G., & Kalyanaraman, K. (2009, February). *Optimal bandwidth choice for the regression discontinuity estimator* (Working Paper No. 14, 726). Cambridge, MA: National Bureau of Economic Research.

Imbens, G. W., & Lemieux, T. (Eds.). (2008a). The regression discontinuity design: Theory and applications [Special issue]. *Journal of Econometrics*, *142*(2).

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

Jacob, B. A., & Lefgren, L. (2004, February). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*, 226–244.

Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression-discontinuity approach. *Journal of Econometrics*, *142*, 785–806.

Lee, D. S. (2001). *The electoral advantage to incumbency and voters' valuation of politicians' experience: A regression-discontinuity analysis of close elections* (Working Paper No. 8441). Cambridge, MA: National Bureau of Economic Research.

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

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*, 655–674.

Lee, D. S. & Lemieux, T. (2009). "Regression Discontinuity Designs in Economics" Working Paper 14723, National Bureau of Economic Research (February). Cambridge: MA.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*, 159–208.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, *142*, 829–850.

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

Neyman, J. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Association*, *2*, 107–180.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.
Quandt, R. (1972). Methods of estimating switching regressions—Unexplored reservoirs of scientific knowledge? *Journal of the American Statistical Association*, *67*, 306–310.
Robinson, J. P., & Reardon, S. F. (2009). *Multiple regression discontinuity design: Implementation issues and empirical examples from education*. Paper presented to the annual research conference of the Society for Research on Educational Effectiveness.
Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.
Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*, 135–146.
Rubin, D. B. (1974). Estimating the causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688–710.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
Schochet, P. Z. (2008, July). *Statistical power for regression discontinuity designs in education evaluations* (Technical Methods Report). Washington, DC: Institute of Education Sciences, U.S. Department of Education.
Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2009). *Combining regression discontinuity analysis and interrupted time-series analysis* (Grant R305D090009). Washington, DC: Institute of Education Sciences, U. S. Department of Education.
Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *51*, 309–317.
van der Klaauw, W. (1997). *A regression-discontinuity evaluation of the effect of financial aid offers on college enrollment* (Working Paper No. 97-10). New York, NY: C.V. Starr Center for Applied Economics, New York University.
van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A Regression-discontinuity approach. *International Economic Review*, *43*, 1249–1287.
van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I? *Journal of Econometrics*, *142*, 731–756.
Weinbaum, E. H., & Weiss, M. (2009, March). *School response to NCLB labels*. Paper presented to the annual Research Conference of the Society for Research on Educational Effectiveness, Washington, DC.
Wong, V. C., Cook, T., Barnett, W. S., & Jung, K. (2007, June). *An effectiveness-based evaluation of five state pre-kindergarten programs using regression-discontinuity* (Working Paper). Evanston, IL: Northwestern University.
Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach*. Florence, KY: South-WesternCENGAGELearning.

APPENDIX
How Regression Discontinuity Designs Identify Treatment Effects

This appendix demonstrates how the three regression discontinuity designs described in this article identify treatment effects for a population or subpopulation. The discussion assumes that control group members have ratings below a cut-point and treatment group members have ratings above the cut-point. All findings apply to the reverse situation as well.

Identifying Treatment Effects With Full Compliance to Assignment: Sharp Regression Discontinuity Designs

Assume that expected counterfactual outcomes and expected treatment effects are continuous functions of ratings. Denote conditional expected counterfactual outcomes as $\bar{Y}_0(r)$ and conditional expected treatment effects as $\bar{\beta}(r)$. The counterfactual outcome and treatment

effect for candidate i is thus:

$$Y_{0i} = \bar{Y}_0(r_i) + \varepsilon_i \tag{A1}$$

and

$$\beta_i = \bar{\beta}(r_i) + \nu_i \tag{A2}$$

where ε_i and ν_i are random errors that have mean zero and are independently and identically distributed. The outcome with treatment for candidate i is thus:

$$Y_{1i} = \bar{Y}_0(r_i) + \bar{\beta}(r_i) + \nu_i + \varepsilon_i \tag{A3}$$

Using an indicator variable T which equals one for candidates who receive treatment and zero for candidates who do not receive treatment, Equations A1 and A2 can be combined as:

$$Y_i = \bar{Y}_0(r_i) + \bar{\beta}(r_i) \cdot T_i + \nu_i \cdot T_i + \varepsilon_i \tag{A4}$$

The expected value of the outcome for a given rating is thus:

$$\bar{Y}(r) = \bar{Y}_0(r) + E\{\beta(r) \cdot T(r)\} \tag{A5}$$

For a sharp regression discontinuity the value of T(r) is zero for all ratings below the cut-point (for control group members) and one for all ratings at or above the cut-point (for treatment group members).

Now consider the limiting values of $\bar{Y}_0(r)$ and $\bar{\beta}(r)$ as they approach a cut-point $(r*)$ from below (for control group members) and from above (for treatment group members). Note first that for any function X(r) that is continuous at a point $(r*)$ its limiting values (limits) from below and above the point equal its value at the point, or:

$$\lim_{r \uparrow r*} X(r) = \lim_{r \downarrow r*} X(r) = X(r*) \tag{A6}$$

Limits of the expected outcome regression (Equation A5) from below the cut-point $(\bar{Y}^-)$ and above the cut-point $(\bar{Y}^+)$ are thus:

Limit from below:

$$\begin{aligned} \lim_{r \uparrow r*} \bar{Y}(r) &\equiv \bar{Y}^- \\ &= \lim_{r \uparrow r*} \bar{Y}_0(r) + \lim_{r \uparrow r*} E\{\beta(r) \cdot T(r)\} \\ &= \bar{Y}_0(r*) \end{aligned} \tag{A7}$$

Limit from above:

$$\begin{aligned} \lim_{r \downarrow r*} \bar{Y}(r) &\equiv \bar{Y}^+ \\ &= \lim_{r \downarrow r*} \bar{Y}_0(r) + \lim_{r \downarrow r*} E\{\beta(r) \cdot T(r)\} \end{aligned}$$

$$= \bar{Y}_0(r*) + \lim_{r \downarrow r*} \bar{\beta}(r)$$
$$= \bar{Y}_0(r*) + \bar{\beta}(r*) \tag{A8}$$

The difference between these two limits equals the average effect of intent-to-treat at the cut-point (ITTC), or:

$$\text{ITT}_C = \bar{Y}^+ - \bar{Y}^- = \bar{\beta}(r*) \tag{A9}$$

Because all subjects assigned to treatment receive treatment, the average effect of ITTC equals the average effect of treatment on the treated at the cut-point (TOTC), which in turn, equals the average treatment effect at the cut-point (ATEC). Thus:

$$\text{ATE}_C = \text{TOT}_C = \text{ITT}_C = Y^+ - Y^- = \bar{\beta}^+ \tag{A10}$$

Identifying Treatment Effects With No-Shows: Type I Fuzzy Regression Discontinuity Designs

Consider what happens to a regression discontinuity design if some subjects assigned to treatment do not receive it, and therefore become no-shows (Bloom, 1984). With no-shows ($Y^+ - Y^-$) still equals the average effect of ITTC, but it no longer equals the TOTC or the ATEC. Specifically:

$$\begin{aligned} \text{ITT}_\text{C} &\equiv Y^+ - Y^- \\ &= \lim_{r \downarrow r*} E\{\beta(r) \cdot T(r)\} - \lim_{r \uparrow r*} E\{\beta(r) \cdot 0\} \\ &= \lim_{r \downarrow r*} \text{E}\{\beta(r) \cdot \text{T}(r)\} \end{aligned} \tag{A11}$$

If no-shows experience no treatment effect (because they are not exposed to treatment), the effect of intent-to-treat is a weighted average of zero treatment effect for no-shows and the average effect of treatment on participants $\bar{\beta}_{pa}^{(r)}$ weighted respectively by the nonparticipation rate ($1 - \bar{T}(r)$) and the participation rate ($\bar{T}(r)$) for treatment group members. (The superscript r in $\bar{\beta}_{pa}^{(r)}$ denotes that it is a function of r). In symbols:

$$\begin{aligned} E\{\beta(r) \cdot T(r)\} &= (1 - \bar{T}(r)) \cdot 0 + \bar{T}(r) \cdot \bar{\beta}_{pa}^{(r)} \\ &= \bar{T}(r) \cdot \bar{\beta}_{pa}^{(r)} \end{aligned} \tag{A12}$$

Substituting Equation A12 into Equation A11 yields:

$$\begin{aligned} \text{ITT}_C &= \lim_{r \downarrow r*} E\{\beta(r) \cdot T(r)\} \\ &= \lim_{r \downarrow r*} (\bar{T}(r) \cdot \bar{\beta}_{pa}^{(r)}) \\ &= \lim_{r \downarrow r*} \bar{T} \cdot \lim_{r \downarrow r*} \bar{\beta}_{pa}^{(r)} \\ &= \bar{T}^+ \cdot \bar{\beta}_{pa}^{(r*)} \end{aligned} \tag{A13}$$

where $\bar{\beta}_{pa}^{(r*)}$ is the average effect of treatment on participants in the treatment group at the cut-point. Equation A13 implies that:

$$\bar{\beta}_{pa}^{(r*)} = \frac{\text{ITT}_\text{C}}{\bar{T}^+} = \frac{\bar{Y}^+ - \bar{Y}^-}{\bar{T}^+} \tag{A14}$$

Note that $\bar{\beta}_{pa}^{(r*)}$ is the average effect of treatment on the treated at the cut-point, or the TOTC. Thus:

$$\text{TOT}_\text{C} = \frac{\bar{Y}^+ - \bar{Y}^-}{\bar{T}^+} \tag{A15}$$

However, because not all treatment group members receive treatment, TOTC does not represent all members of the population at the cut-point and therefore does not equal the ATEC. To identify this latter effect requires stronger assumptions.

Identifying Treatment Effects With No-Shows and Crossovers: Type II Fuzzy Regression Discontinuity Designs

When some treatment group members do not receive treatment, indicating the presence of no-shows and some comparison group members do receive treatment, indicating the presence of crossovers, the probability of treatment is less than one for treatment group members and greater than zero for control group members. This further dilutes the treatment contrast.

With no-shows and crossovers ($\bar{Y}^+ - \bar{Y}^-$) still equals the average effect of intent-to-treat at the cut-point (ITTC). But now:

$$\bar{Y}^+ - \bar{Y}^- = \lim_{r \downarrow r*} E\{\beta(r) \cdot T(r)\} - \lim_{r \uparrow r*} E\{\beta(r) \cdot T(r)\} \tag{A16}$$

This article introduced the causal framework of Angrist et al. (1996), which specifies four conceptual subgroups that are distributed in the same proportions (in expectation) within a treatment group and control group: compliers, always-takers, never-takers, and defiers. If defiers do not exist (a plausible possibility for many situations), the treatment group and control group for a regression discontinuity design consists of compliers, always-takers, and never-takers in proportions $P_{co}^{(r)}$, $P_{at}^{(r)}$, and $P_{nt}^{(r)}$, respectively.

The average effect of treatment on treatment group members therefore equals the weighted mean of its expected effect on compliers ($\bar{\beta}_{co}^{(r)}$), its expected effect on always-takers ($\bar{\beta}_{at}^{(r)}$), and zero effect for never-takers, with weights equal to $P_{co}^{(r)}$, $P_{at}^{(r)}$ and $P_{nt}^{(r)}$, respectively. The average effect of treatment on control group members equals a weighted mean of zero effect for compliers, the expected effect for always takers and zero effect for never-takers, with weights equal to $P_{co}^{(r)}$, $P_{at}^{(r)}$ and $P_{nt}^{(r)}$, respectively. In symbols:

For the treatment group:

$$\begin{aligned}
\bar{Y}^+ &= \lim_{r \downarrow r*} E\{\beta(r) \cdot T(r)\} = \lim_{r \downarrow r*} \left[P_{co}^{(r)} \cdot \bar{\beta}_{co}^{(r)} + P_{at}^{(r)} \cdot \bar{\beta}_{at}^{(r)} + P_{nt}^{(r)} \cdot 0 \right] \\
&= \lim_{r \downarrow r*} \left[P_{co}^{(r)} \cdot \bar{\beta}_{co}^{(r)} + P_{at}^{(r)} \cdot \bar{\beta}_{at}^{(r)} \right] \\
&= \lim_{r \downarrow r*} \left[P_{co}^{(r)} \cdot \bar{\beta}_{co}^{(r)} \right] + \lim_{r \downarrow r*} \left[P_{at}^{(r)} \cdot \bar{\beta}_{at}^{(r)} \right]
\end{aligned} \tag{A17}$$

For the control group:

$$\begin{aligned}\bar{Y}^{-} &= \lim_{r\uparrow r*} E\{\beta(r)\cdot T(r)\} = \lim_{r\uparrow r*}\left[P_{co}^{(r)}\cdot 0 + P_{at}^{(r)}\cdot\bar{\beta}_{at}^{(r)} + P_{nt}^{(r)}\cdot 0\right] \\ &= \lim_{r\uparrow r*}\left[P_{at}^{(r)}\cdot\bar{\beta}_{at}^{(r)}\right] \end{aligned} \tag{A18}$$

Consequently:

$$\begin{aligned}\bar{Y}^{+} - \bar{Y}^{-} &= \lim_{r\downarrow r*}\left[P_{co}^{(r)}\cdot\bar{\beta}_{co}^{(r)}\right] + \lim_{r\downarrow r*}\left[P_{at}^{(r)}\cdot\bar{\beta}_{at}^{(r)}\right] - \lim_{r\uparrow r*}\left[P_{at}^{(r)}\cdot\bar{\beta}_{at}^{(r)}\right] \\ &= \left[\lim_{r\downarrow r*} P_{co}^{(r)}\right]\left[\lim_{r\downarrow r*}\bar{\beta}_{co}^{(r)}\right] + \left[\lim_{r\downarrow r*} P_{at}^{(r)}\right]\left[\lim_{r\downarrow r*}\bar{\beta}_{at}^{(r)}\right] - \left[\lim_{r\uparrow r*} P_{at}^{(r)}\right]\left[\lim_{r\uparrow r*}\bar{\beta}_{at}^{(r)}\right] \end{aligned} \tag{A19}$$

Assume that $P_{co}^{(r)}$, $P_{at}^{(r)}$, $\bar{\beta}_{co}^{(r)}$ and $\bar{\beta}_{at}^{(r)}$ are continuous functions of r at the cut-point. Hence they converge from above and below the cut-point to their value at the cut-point. Consequently the last two limits in Equation A19 cancel each other, leaving:

$$\begin{aligned}\bar{Y}^{+} - \bar{Y}^{-} &= \lim_{r\downarrow r*}\left[P_{co}^{(r)}\cdot\bar{\beta}_{co}^{(r)}\right] \\ &= \left[\lim_{r\downarrow r*} P_{co}^{(r)}\right]\cdot\left[\lim_{r\downarrow r*}\bar{\beta}_{co}^{(r)}\right] \\ &= P_{co}^{(r*)}\cdot\bar{\beta}_{co}^{(r*)} \end{aligned} \tag{A20}$$

The proportion of treatment group members at the cut-point receiving treatment ($\bar{T}^{+}$) equals the proportion of compliers plus the proportion of always-takers in this group. The proportion of control group members at the cut-point receiving treatment ($\bar{T}^{-}$) equals the proportion of always-takers in this group. In symbols:

$$\bar{T}^{+} = P_{co}^{(r*)} + P_{at}^{(r*)} \tag{A21}$$

$$\bar{T}^{-} = P_{at}^{(r*)} \tag{A22}$$

Hence:

$$\bar{T}^{+} - \bar{T}^{-} = P_{co}^{(r*)} \tag{A23}$$

Substituting Equation A23 into Equation A20 yields:

$$\bar{Y}^{+} - \bar{Y}^{-} = [\bar{T}^{+} - \bar{T}^{-}]\bar{\beta}_{co}^{(r*)} \tag{A24}$$

which implies that:

$$\bar{\beta}_{co}^{(r*)} = \frac{\bar{Y}^{+} - \bar{Y}^{-}}{\bar{T}^{+} - \bar{T}^{-}} \tag{A25}$$

Consequently, the average effect of treatment on compliers at the cut-point, or local average treatment effect at the cut-point (LATEC), is:

$$\text{LATE}_{\text{C}} = \frac{\bar{Y}^{+} - \bar{Y}^{-}}{\bar{T}^{+} - \bar{T}^{-}} \tag{A26}$$