Recent Perspectives on the Regression Discontinuity Design*

Richard Berk
Department of Criminology
Department of Statistics
University of Pennsylvania

March 17, 2008

Abstract

The regression discontinuity design was originally proposed in 1960 as a powerful alternative to randomized experiments. It has been little used since. Over the past decade, however, the design has been increasingly and successfully employed by economists in a variety of studies. In this paper, the fundamentals of the regression-discontinuity are discussed. Recent advances are emphasized

1 Introduction

The regression discontinuity design has a long and complex history. It was originally proposed by psychologists Thistlewaite and Campell in 1960 (Thistlewaite and Campbell, 1960). The design became widely known with the publication of the justly famous *Experimental and Quasi-Experimental Designs for Research* by Campbell and Stanley (Campbell and Stanley, 1963) and was later elaborated by Trochim (1984, 2001). In 1972, the design was independently rediscovered by econometrician Arthur Goldberger (Goldberger, 1972).

*Work on this paper was funded by a grant from the National Science Foundation: SES-0437169, "Ensemble methods for Data Analysis in the Behavioral, Social and Economic Sciences."

Even in these early formulations, the design was simple and powerful. But there were few applications and apparently only two published studies with significant crime and justice content (Berk and Rauma, 1983; Berk and de Leeuw, 1991). Over the past 15 years, a number of economists have extended the design (Imbens and Lemieux, 2008b) and applied it in a wide variety of settings (Imbens and Lemieux, 2008a). There is one analysis that criminologists should find especially instructive (Chen and Shapiro, 2007), but most of the recent applications are well-worth a close reading. An account of how and why interest in the regression discontinuity design has varied over the years can be found in recent paper by Thomas Cook (Cook, 2008).

In this chapter, the fundamentals of the regression discontinuity design are considered. Some recent advances are highlighted. The discussion begins with brief introduction to the ways in which statisticians think about causal inference. Then, the classic regression discontinuity design is examined. Newer material follows.

2 The Basic Regression Discontinuity Design

The regression discontinuity (RD) design has many of the assets of a randomized experiment, but can be used when random assignment is not feasible. Typically, the goal is to estimate the causal effect of an intervention such gate money for prison inmates, anger management for troubled teens, or changes in police patrolling practices. In the simplest case, there is an experimental and comparison group with assignment fully determined by a explicit and observable rule. For example, whether community policing is introduced in a neighborhood depends on whether the crime rate is above a specific threshold. Neighborhoods falling above that threshold get community policing. Other neighborhoods get business as usual. A relatively simple and compelling analysis can follow. In some circles, the RD design is called a "quasi-experiment."

One appeal of the RD design is "political." If the assignment rule represents need, either empirically or morally, one can can sometimes more easily garner the support of stakeholders. Cooperation from study subjects can also more easily follow. For example, it might seem to just make good sense to assign a policing innovation to high crime neighborhoods. The technical complication is that neighborhoods assigned to the treatment condition will not be comparable on the average to neighborhoods assigned to the alterna-

tive. Indeed, systematic selection is explicitly build into the design. How the potential biases can be overcome in practice is a key theme in the material that follows.

To appreciate the underlying machinery of the RD design requires some familiarity with the way causal inference has come to be formulated by most statisticians. The framework was proposed by Neyman in 1923 and later extended by Rubin (1974) and Holland (1986). These days, it is sometimes called the "Rubin Causal Model."

In its most simple form, there are observational units: people, neighborhood, police departments, prisons or other entities. There is a binary intervention. Some of the units are exposed to one "arm" of the intervention, and the other units are exposed to the alternative "arm" of the intervention. There is interest in the intervention's causal effect. For example, one might want to learn how intensive parole supervision, compared to conventional parole supervision, affects recidivism.

Each *unit* is assumed to have two *potential* outcomes, one if exposed to the treatment and one if exposed to the alternative. These outcomes are hypothetical and can vary across units. Thus, a given parolee would have one response if placed under intensive supervision and another response if placed under the usual supervision. Another parolees would also have two potential responses, which could differ from those of the first parolee.

Following Imbens and Lemieux (2008), let $Y_i(1)$ denote the outcome if unit i is exposed to the treatment, and $Y_i(0)$ denote the outcome if unit i is exposed to the alternative condition. Interest centers on a comparison between $Y_i(1)$ and $Y_i(0)$, often their difference. An example is the number of failed urine tests should a given parolee be placed under intensive supervision minus the number of failed unit tests should that same parolee be placed under supervision as usual.

In practice, however, one can never observe both $Y_i(1)$ and $Y_i(0)$. A given unit will only experience the treatment or the alternative. Let $W_i = 1$ if unit i is actually exposed to the treatment, and $W_i = 0$ if unit i is actually exposed to the alternative. The *observed* outcome is then

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1). \tag{1}$$

Both Y_i and W_i can be observed. For the basic RD design, one also has a single observed covariate X_i that fully determines whether unit i is exposed to the treatment or the alternative. In the simplest case, X_i is a continuous

covariate with an imposed threshold. A unit that falls on one side of the threshold is assigned to the treatment condition. A unit that falls on the other side of the threshold is assigned to the control condition.

X can be associated with the potential outcomes. For example, X may be a measure of parole risk. A parolee who scores above some threshold is assigned to intensive supervision. If that parolee scores below the threshold, assignment is to conventional supervision. The risk measure is by design associated with the potential outcome of re-offending. This may seem curious, but it will soon be clear that possible biases can be addressed.

There can be other observed covariates Z. In some cases, they are used to define a multivariate assignment rule. For example, parolees who are younger than 21 and who have more than two prior convictions for a violent crime may be assigned to intensive supervision. In other cases, the Z play no role in the assignment, but are related to the potential outcomes. For now, there is no need to consider either complication.

Because the definition of a causal effect, $[Y_i(1) - Y_i(0)]$, is unobservable in practice,[1] there is a shift to the group level. Interest centers on the observed *average* response of the units exposed to the treatment compared to the observed *average* response of the units exposed to the alternative. We are seeking the average treatment effect. Given the assignment rule, however, there is the possibility that the units exposed to the treatment will differ systematically from the units exposed to the alternative. And those systematic differences can be related to the potential outcomes. There is the risk of building in selection bias.

Under certain assumptions, there is a solution. Consider Figure 1. On the horizontal axis is the assignment variable X. On the vertical axis is the response (outcome) variable Y. Units scoring at or above 5 on X are exposed to the intervention. Units scoring below 5 on X are exposed to the alternative. The higher line (dotted below the threshold of 5 and solid on or above the threshold of 5) represents the potential response for different values of the assignment variable when a unit is exposed to the intervention. The lower line (solid below the threshold of 5, and dotted on or above the threshold of 5) represents the potential response for different values of the assignment variable when a unit is exposed to the alternative. The solid parts of each response function represent the potential responses actually observed under the assignment rule. The vertical solid line at $X = 5$ is the causal

[1]Sometimes the ratio rather than the difference is used, but that too is unobservable.

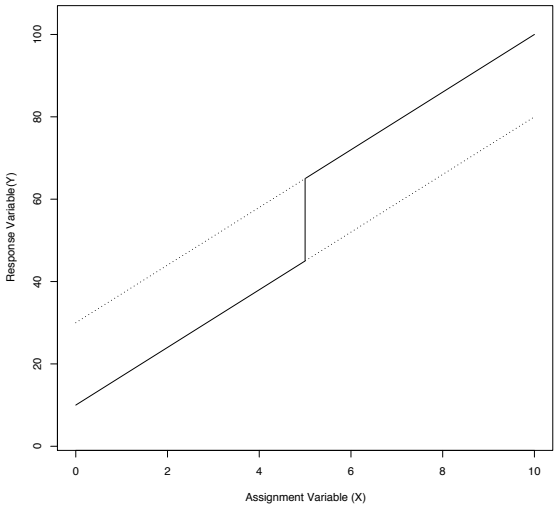

Figure 1: Conventional Linear Case

effect of the intervention.

Figure 1 shows the theory required for the basic RD design, and there is a lot going on. For both groups, the relationship between X and Y is positive. The larger the value of X, the larger the value of Y tends to be. If one just compared the observed mean of the group receiving the intervention (i.e., with $X \geq 5$) with the observed mean of the group receiving the alternative (i.e., with $X < 5$), as one might do in a randomized experiment, an estimate of the average treatment effect could be substantially biased. By design, the group exposed to the treatment has larger values for X, which imply larger values of Y regardless of the treatment. Analogous risks follow if the relationship between X and Y is negative.

However, note that the relationship between the assignment variable and the response variable is for both groups assumed to be linear. This is true for potential responses that are observed (i.e., the solid lines) and potential responses that are not observed (i.e., the dotted lines). The linearity is a very restrictive but very convenient. Note also that the linear relationships are assumed to be parallel. This too is very restrictive but very convenient. One has a strong two-part theory about how the assignment variable is related to the response: the two response functions can differ only by a constant. That constant is the treatment effect.

If the relationships shown in Figure 1 are a good approximation of reality, a very simple and effective analysis can follow. The difference between the vertical placement of the linear relationship for the units exposed to the intervention and the vertical placement of the linear relationship for the units exposed to the alternative can provide an estimate the average treatment effect.

More specifically, let

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \varepsilon_i. \tag{2}$$

Equation 2 is a conventional regression expression representing how the values of the response variable are generated. All of the observables are defined as before, and ε_i is the usual regression disturbance. When $W_i = 0$, the conditional expectation of Equation 2 gives the linear relationship between X and Y for the group exposed to the alternative. When $W_i = 1$, the conditional expectation of Equation 2 gives the linear relationship between X and Y for the group exposed to the treatment. β_2 is the common slope of the two response functions. β_0 is the intercept of the response function for the group exposed to the alternative. β_1 is the vertical distance between the two estimated response function, which quantifies how much the response function for the treatment group is shifted up (i.e., $\beta_1 > 0$) or down (i.e., $\beta_1 < 0$) because of the intervention. In Figure 1, the length of the vertical line at $X = 5$ is equal to the value of β_1. If the intervention has no effect, $\beta_1 = 0$, and the two lines collapse into one.

In practice, one can use the data on hand and ordinary least squares to estimate the parameters of Equation 2. If the two linear relationships are at least nearly linear and parallel, $\hat{\beta}_1$ is effectively an unbiased estimate of the average treatment effect. A key is that the true selection mechanism is known and can be accurately represented by a threshold on X. The usual hypothesis test is that $\hat{\beta}_1 = 0$; there is no treatment effect.

An obvious question is how one would determine in practice whether the two relationships are linear and parallel. A good way to start is to construct a scatterplot of Y against X. On each side of the threshold, one should see a truncated version of the idealized elliptical scatterplot. If there is a treatment effect, the plotted points should be shifted up or down in the region to the right of the threshold (where the units are exposed to the intervention). An example is discussed later.

A useful second step is to examine a scatterplot constructed from the residuals and fitted values of Equation 2. When the residuals are plotted

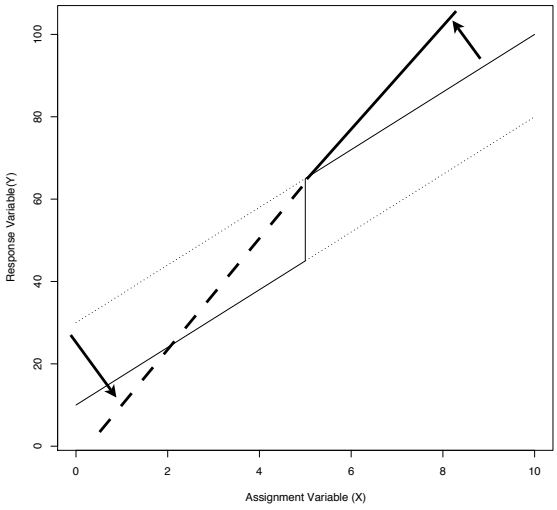

Figure 2: The Conventional Linear Case Gone Wrong

against the fitted values, one should see with no dramatic patterns. In particular, there should be no evidence that the assumed linear and parallel regression response functions are something else. It can also be useful to consider whether the usual assumption of a constant residual variance is credible. In short, one should examine the results from Equation 2 as one would the results from any regression analysis. Graphical methods can be especially instructive and are less constrained by the untestable assumptions that test-based diagnostic methods require (Freedman, 2008). An excellent discussion of a number of useful diagnostic methods can be found in the text by Cook and Weisberg (1999).

What can go wrong if proper diagnostic methods are not employed? Figure 2 provides a simple illustration. The two response functions for the potential outcomes are not parallel. The response function for the treatment group is much steeper. (The solid part of the line is observed and the dashed part of the line is unobserved.) In this particularly perverse example, the gap at the threshold between the observed outcomes for the group exposed

to the treatment and the observed outcomes for the group exposed to the alternative just happens to be the same as the treatment effect when the assumption of parallel response functions is met. But this gap is not caused by a discontinuity. There is no evidence of a treatment effect.

It is certainly possible to alter Equation 2 allowing for linear response functions that are not parallel. Looking again at Figure 2, however, one would have to know about the non-parallel response functions in advance. One cannot determine from the data alone whether the response functions are not parallel or whether the intervention altered the slope and intercept of the treatment group.

This can lead to an identification problem. Suppose one were prepared to bet that the two response function were linear but not parallel. Then, β_1 combines the intercept of the response function for the treatment group in the absence of the intervention with a new intercept for the treatment group should the intervention shift the treatment group response function. Because there are no observations for the treatment group in the absence of the intervention, the confounding cannot be disentangled, and the basic regression discontinuity design fails.

Suppose one were prepared to assume that the two response functions are linear and parallel before the intervention, but that after the intervention the slope and intercept are altered. By including $(W_i \times X_i)$ as new regressor in the Equation 2, the intervention-altered linear relationship for those exposed to the treatment can be identified, and the linear relationship for those exposed to the alternative can be identified. But there is an important complication. The size of the gap between the response function of the treatment group and the response function of the comparison group depends on the value of X. Indeed, the treatment effect may be positive for some values of X and negative for other values of X. Although this is statistically acceptable, constructing a coherent substantive explanation can be challenging. One simplification is to focus only on the estimated average treatment effect in the immediate neighborhood of the threshold. This is an approach to which we return shortly.

Some of lessons from the linear case carry over to the nonlinear case. If the two response functions are nonlinear but parallel, one can in principle estimate a shift resulting from an intervention. However, it is not clear what a change in slope means if the response functions are not linear. The nonlinear case is discussed a more later.

3 The Generalized Regression Discontinuity Design

One problem with the basic RD design is that the response is assumed to be quantitative. In many applications, the response is categorical or a count. An example of the former is whether a parolee reoffends. An example of the latter is the number of crimes the parolee commits.

The basic RD design is easily extended to include formulations from the generalized linear model. Logistic regression and Poisson regression are two common illustrations. Consider, for instance, a binary response.

We can proceed initially as before. Equation 3 is the same as Equation 2 except that the quantitative response Y_i^* is now completely unobserved regardless of whether a unit falls above or below the threshold. For example, Y_i^* may be the proclivity of a prison inmate to engage in some form of serious misconduct, and no measures of this proclivity are available. As before, we let

$$Y_i^* = \beta_0 + \beta_1 W_i + \beta_2 X_i + \varepsilon_i. \tag{3}$$

We now hypothesize a second kind of threshold. If an inmate's proclivity toward misconduct exceeds a certain value, an act of misconduct is observed. If that threshold is not exceeded, no act of misconduct is observed. Suppose this outcome is coded so that "1" denotes an observed act of misconduct, and "0" denotes no observed act of misconduct.

Drawing on a common motivation for logistic regression (Cameron and Trivedi (2005: Section 14.4), the probability P_i of observing a "1" depends on the probability that Y_i^* will exceed the misconduct proclivity threshold. If one then assumes that ε_i has a logistic distribution, Equation 4 can follow directly:

$$\log \frac{P_i}{1 - P_i} = \gamma_0 + \gamma_1 W_i + \gamma_2 X_i, \tag{4}$$

where γ_0 through γ_2 are new regression coefficients. In other words, the systematic part of the formulation is the same as in Equation 2, but the response is now in units of log-odds (also called "logits"). Figure 3, therefore, is analagous to Figure 1.

If Equation 4 is solved for P_i, the result is shown in Figure 4. Note that in Figure 3 when the response is in logit units, the two response functions are linear and parallel. But when in Figure 4 the response is in probability units, the two response functions are neither linear nor parallel. This is not

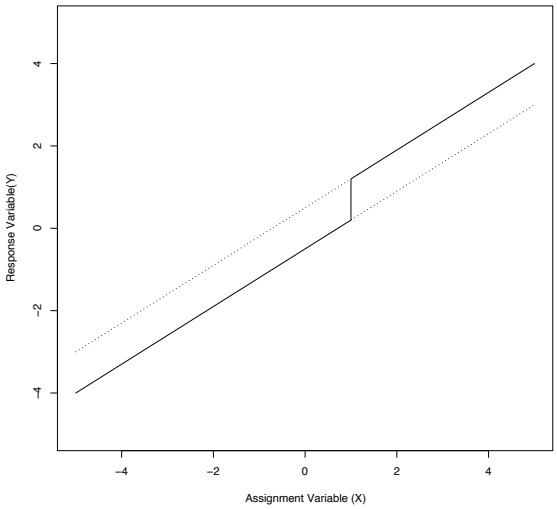

Figure 3: The Log Odds Representation of the Logistic Case

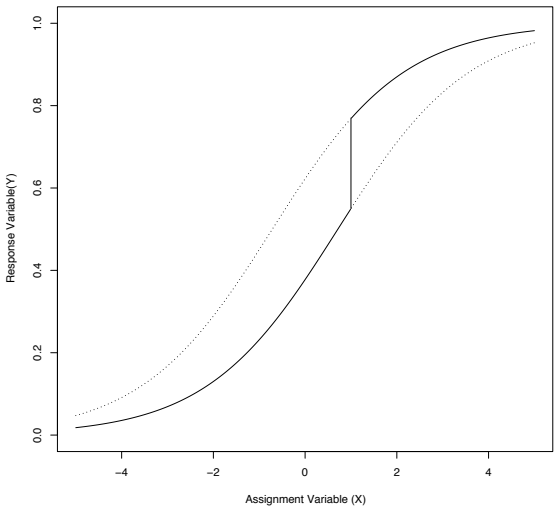

Figure 4: The Generalized Linear Case

a problem as long as, consistent with Equation 4, the estimated treatment effect is reported in logit units or as an odds multiplier. In addition, any regression diagnostics should be applied to the log-odds form of the model as well. Happily, many graphical diagnostic procedures for logistic regression are designed to work in this fashion. Again, a good reference is Cook and Weisberg (1999). Berk and de Leeuw (1999) provide an instructive application.

Analogous issue arise if the response variable is a count. The canonical mean function is not the log of the odds, but the natural logarithm. The generalized RD design still applies and can easily be employed.

4 Matching Strategies

From the issues just discussed, it should be clear that the credibility of the basic and generalized RD design depends on the assumptions one makes about the two response functions linking X to Y. Both designs respond to the same fundamental problem: the observed responses of the group exposed to the treatment and the observed responses of the group exposed to the alternative reside in disjoint regions of the assignment variable. There are no values of X at which one can observe responses for both groups. Therefore, all comparisons necessarily depend on extrapolations, which in turn, depend on the functional forms assumed.

An alternative strategy is to focus on the observed responses on either side of the threshold and very close to it. If this region, often called a "window," is sufficiently narrow, the values of X within it are likely to be very similar. Because Y is a function of X, this suggests a matching strategy comparing the average observed response of the units just to the right of the threshold to the average observed response of the units just to the left of the threshold. The difference between the two averages can serve as an estimate of the average treatment effect.

An important assumption is that the small piece of the response function just to the right of the threshold and the small piece of the response function just to the left of the threshold are linear and parallel to each other. To the degree that this assumption is violated, the size of the possible bias increases. For example, if both are linear and parallel but with a positive or negative slope, the absolute value of the difference between the two averages will be

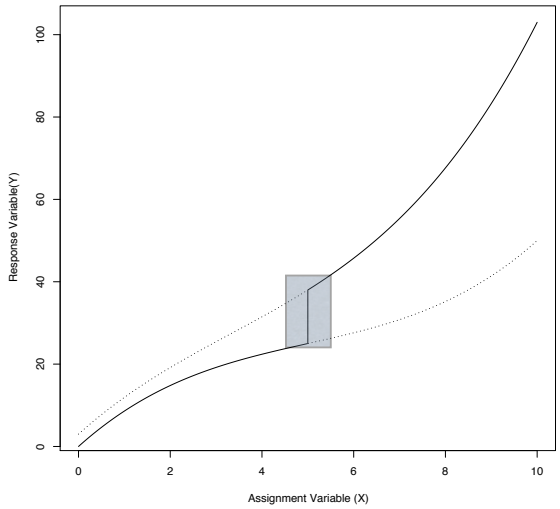

Figure 5: Matching within the RD Design

larger than the absolute value of the gap size at the threshold.[2] The hope is that if the window can be made sufficiently narrow, this bias will be negligible. Ideally, therefore, the matching strategy can provide useful estimates of average treatment effects with far less reliance on assumed functional forms over the entire length of the response functions.

Figure 5 illustrates the general idea. The two response response functions are neither linear nor parallel. And these would likely be unknown in any case. If within the shaded region the two functions were linear and flat, the difference between the two averages would be the same as the treatment effect at the threshold. In this illustration, the estimated treatment effect would be a bit too large. The gap is larger then had the two functions been linear and flat.

It follows that a critical feature of the window approach is the width of

[2]One can construct special cases in which the assumptions are violated, and there is no bias. But in practice, such scenarios would be extremely rare.

the shaded area. How wide should the shaded area be? If the shaded area is more narrow, the true causal effect often can be better approximated. But there will be fewer observations with which to estimate the average treatment effect. With less information, there will be more uncertainty in the estimate. The tension is an example of the well-known bias-variance tradeoff (Hastie and Tibshirani, 1990: 40-42). To consider the implications of this tradeoff, we to turn to how one can estimate the average treatment effect. In this context, the terms "bandwidth" or "span" are sometimes used instead of the term "window."

Perhaps the simplest estimator of the average treatment effect within the window is the difference between the mean of the responses to the right of the threshold and mean of the responses to the left of the threshold (Imbens and Lemieux, 2008: 623-624). Sometimes, a better approach is to use an explicit kernal function (Hastie and Tibshirani, 1990: 18-19), the impact of which can be to weight observations so that those closer to the threshold are given more weight when the means are computed. Values of Y closer to the threshold will often provide more accurate information about the value of Y at the threshold.

Different weighting schemes can be used. For example, when a linear kernel is used, weights decline linearly with distance from the threshold. When a Gaussian kernel is used, the weights decline as a function of the normal distribution centered on the threshold. In practice, all of the common weighting functions that decline with distance from the threshold usually give very similar results. In fact, there is often not much difference between the estimated treatment effect using any of the common weighting functions and the estimated treatment effect using no weighting at all (i.e., all cases get the same weight).

In contrast, the size of the window can have a dramatic effect. Therefore, a criterion is needed that strikes a useful balance between the bias and the variance. One popular criterion is the mean squared error: the mean of the squared disparities between the fitted values $\hat{Y}_i$ and the observed values of Y_i. If a window size can be chosen to minimize the mean squared error, the sum of the variance and the bias squared will be minimized as well (Hastie et al., 2001: 196-200). In this sense, minimizing the mean squared error provides a good way to define an appropriate balance between the variance and the bias.

Several procedures can be used as an operational stand-in for the theoretical mean squared error. Cross-validiation is a useful option and with

modern computers, the calculations can be done very quickly. Other options include the generalized cross-validation statistic, AIC, BIC, and Mallows Cp. In practice, each of these approaches usually will lead to about the same results. An excellent discussion of the issues can be found in the book by Hastie and his collegues (2001: Chapter 7).

The comparison between means within the window can be generalized. Suppose that within the window the two unknown response functions are linear, but not necessarily flat or even and parallel. Then, one can often obtain a more accurate estimate of the treatment effect by employing two linear regressions within the window, one on each side of the threshold (Imbens and Lemiux, 2008: 624-625). Each is a form of local linear regression. The sole regressor is the distance from the threshold. More specifically, we assume that

$$Y_{iL} = \delta_{0,L} + \delta_{1,L} \cdot (X_{iL} - T) + \varepsilon_{iL}, \tag{5}$$

and

$$Y_{iR} = \delta_{0,R} + \delta_{1,R} \cdot (X_{iR} - T) + \varepsilon_{iR}, \tag{6}$$

where within the window, L denotes the left side of the threshold, R denotes the right side of the threshold, T is the value of X at the threshold, X_i is the value of X for given observations, and ε_{iL} and ε_{iR} are conventional regression disturbances. Because $\delta_{1,L} \neq \delta_{1,R}$, the linear response functions do not have the same slope.

At the threshold, $(X_i - T) = 0$. Therefore, the average treatment effect at the threshold is $\delta_{0,R} - \delta_{0,L}$. With data for Y and X, one can estimate the values of the parameters from both equations and obtain an estimate of the treatment effect at the threshold by computing the difference between the two intercepts. That is, the estimate of the average treatment effect is $\hat{\delta}_{1,L} - \hat{\delta}_{1,R}$.

Although Equations 5 and 6 have much the same structure as Equation 2, their purpose is rather different. The goal is to estimate the difference between the response functions at the threshold. Therefore, despite the use of regression, the matching logic still prevails. If the regression model is approximately correct, there should be gains by the mean squared error criterion. In addition, the usual statistical inference undertaken with linear regression can apply although sometimes robust standard errors are desirable (Cameron and Trivedi, 2005: Section 4.4.5)

In principle, the regression estimate can be improved upon in two ways. First, one can employ a kernel weighting scheme so that when the regression coefficients are estimated, observations closer to the threshold are given greater weight. In practice, the means squared error gains are usually modest at best. Second, one can assume that $\hat{\delta}_{1,L} = \hat{\delta}_{1,R}$ and re-estimate the values of $\hat{\delta}_{0,L}$ and $\hat{\delta}_{0,R}$. In other words, one proceed as if the two linear response functions are parallel. There can be small but noticeable gains by the mean squared error criterion if the equivalence is approximately true. Once again, regression diagnostics can be very instructive.

There can be substantial bias in the estimated average treatment effect at the threshold if regression diagnostics indicate that the response functions are not linear. One possible remedy is to replace the local linear regression with local polynomial regression (Fan and Gijbels, 1996) or an even more flexible smoother such as found in the generalized additive model (Hastie and Tibshirani, 1990). But, the moment one opens the door to nonparametric regression, there may be no longer a need to stay within the window, and a wide variety of tools are in play. Regression splines and regression smoothers, for instance, can be very effective (Berk, 2008: Chapter 2; Bowman et al., 2004). The task at hand can be reformulated as function estimation problem, with the amount of smoothing replacing the size of the window as a key matter for tuning. The result can be

$$Y_i = \beta_0 + \beta_1 W_i + f(X_i - T) + \varepsilon_i, \tag{7}$$

where $f(X_i - T)$ is determined empirically. Equation 7 is easily extended to the generalized linear model so that binary and count response variables can be analyzed.

One interesting feature of Equation 7 is that if the fitting procedure for $f(X_i - T)$ is made sufficiently flexible, there may be no need to include W_i as a regressor. Should there be an important change in the response function in the neighborhood of the threshold, the nonparametric regression procedure is likely to find it. The size of the neighborhood will be determined as well. At the very least, this suggests first using a very flexible nonparametric regression procedure as an exploratory technique to help in the specification of Equation 7.

5 Some Extensions Not Addressed

The RD design may be extended further, but space limitation preclude more than a very brief discussion. One easy and direct extension is to have a deterministic assignment rule constructed from more than one covariate. If there are two such covariates, for example, the threshold is a plane not a line. The various estimation procedures can be altered accordingly. Likewise it is relatively easy to have a proper RD design with more than one intervention, much in the spirit of factorial designs in true experiments.

A more complicated extension addresses the problem of compliance. Just as in randomized experiments, study subjects do not always comply with the treatment or alternative condition assigned. Then, one might be interested in trying to estimate the impact of the intervention assigned (i.e., an "intention-to-treat" analysis) or the impact of the intervention received. The former can be estimated with the procedures we have described. The latter requires more complicated and fragile procedures. A possible approach is using the treatment assigned, conditional on the assignment variable, as an instrumental variable. These and other alternatives are discussed by Imbens and Lemioux (2008b).

6 External Validity

The results from an RD design raise the same external validity issues as those from a randomized experiment and more. If the study subjects are probability sample from a well-defined population, generalizations to that population can be appropriate. If the study subjects are not selected by probability sampling, generalizations beyond the study subjects must rely on theory or replications. However, any estimates using the matching approach can restrict generalizations further. The relevant units are now only those that fall in the window or for some procedures, only at the threshold. For data that are a proper probability sample, generalizations can now only be to elements in the population that would fall within the window or at the threshold, respectively. When the data are not a proper probability sample, generalizations based on theory or replications must also take these restrictions into account.

In short, the basic and generalized RD design has the same kinds of external validity constraints as randomized experiments. The external validity

constraints can be far more binding if a matching estimation approach is used. The tradeoff is that the matching approach's internal validity may be stronger.

7 An Illustration

We turn now to an illustration. The intent is to provide an overview of how some of the procedures described above can be used in practice. Space limitation preclude an in-depth discussion using several different data sets with varying properties.

A good way to begin the analysis of data from any RD design is to examine a scatterplot of the response variable against the assignmen t variable. Figure 6 is just such a plot for some fictitious data. The relationship between X and Y is by construction a fourth degree polynomial. This implies that the response functions on either side of the threshold are the same, save for a possible offset. Also by construction, the threshold is at 0.0, and the treatment effect equals 10.0. It is readily apparent that there is a discontinuity at the threshold. For real data, evidence of a treatment effect is usually not so easily discovered.

One would normally not know that the relationship between X and Y was a fourth degree polynomial. But it would likely be empirically apparent that the two response functions were approximately linear and flat over much the range of X, at least where there are data. Except for a few observations at the tails, two linear parallel response functions might seem to be a sufficiently good approximation for the RD analysis.

Table 1 shows some results. Its four columns contain in order the kind of estimator, the average treatment effect estimate, the standard error and the proportion of the deviance that is accounted for by the estimator.[3] In the first row are the results if the correct relationship between X and Y were known. This is the estimation gold standard. The estimated treatment effect is 10.4, the standard error is .37 and about 83% of the deviance is accounted for. The estimate is about one standard error from the truth of 10.0. One would not reject the null hypothesis that the treatment effect is 10.0.[4]

[3]Fit quality was evaluated using the AIC, but the adjusted proportion of deviance accounted for by the model is reported for ease of exposition.

[4]The treatment effect estimate is not exactly 10.0 because of random sampling error.

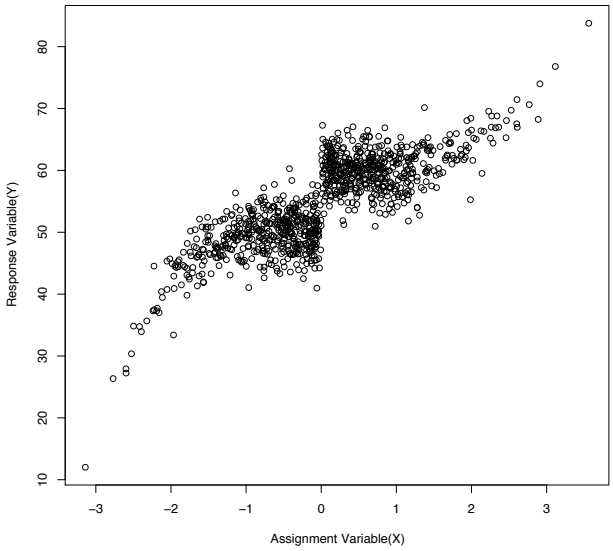

Figure 6: A Scatter Plot for Some Fictitious Data (N=1000)

Model	Estimate	SE	Fit
Correct Model	10.4	.37	.83
Linear Model	6.3	.37	.74
Difference in Means (N=400)	9.8	.31	.72
Regression Splines (k=3)	6.16	.37	.75
Regression Splines (k=5)	11.8	.43	.82
Regression Splines (k=7)	9.97	.51	.83

Table 1: Estimates of the RD Average Treatment Effect for Fictitious Data — True Treatment Effect = 10.0

The second row contains the results if the basic RD design is assumed and the linear model applied. The estimate of the average treatment effect is now 6.3, a reduction of about a third. The standard error has not changed (up to two decimal places), but the fit drops from .83 to .74. In this case, if after looking at the scatterplot a researcher decided that the linear model was good enough, the results would not be grievously wrong. The linear fit is not terribly far from the ideal fit: .83 versus .74 (although the ideal fit would not be known) because where most of the data are, the linear model is about right. Whether the size of the treatment effect underestimate is large enough to matter would depend on context. In many policy settings, underestimating an average treatment effect by a third can transform a cost-effective intervention into one that is not cost-effective.

The third row shows the results for the difference between means in the region immediately on either side of threshold. The region was defined by the X between -1 and 1. Four hundred observations are contained between these two boundaries. The estimated average treatment effect is 9.8, the standard error is .31 and about 72% of the deviance is accounted for. The reduction in the standard error despite the smaller sample size results from a substantial decrease in the variance of the response variable once the more extreme values of X are trimmed. Overall, these are all very good results, which suggest that within the window, the response functions are close to linear and and flat. A close look at Figure 6, confirms these inferences. Enlarging or shrinking the region by as much as 25% does not materially change the results. Had the relationship not been approximately flat, a very different causal effect estimate could have resulted.

Rows four through six contain the results for a particular form of non-parametric regression: regression splines. In effect, the procedure seeks to fit the response functions inductively (Berk, 2008, Chapter 2). A single functional form with a possible offset is assumed, just as in Equation 7. And the entire dataset is used.

A key tuning parameter determines the amount of smoothing. Here, that tuning parameter is the number of "knots" (k). The larger the number of knots, the more flexible the fitting function, and the less smooth the fitted values will be. Results when $k = 3$, suggest that the fitted values are not flexible enough. The story is about the same as for the linear model. When $k = 5$, the fit improves dramatically, but the estimated average treatment effect is a bit too large. When $k = 7$, the fit of the fourth degree polynomial is improved a bit more, and the estimated average treatment effect is almost

perfect. Also, the quality of the fit is virtually the same as for the gold standard in the first row.

From the sequence of results, there are several lessons. To begin, the key to obtaining useful estimates of the average treatment effect is to first arrive at good approximation of the relationship between X and Y. To this end, no single method dominates over the variety of scatterplot patterns one is likely to find in practice. In this illustration, all of the estimates obtained were positive, some very close to the truth. In each case, one would easily reject the null hypothesis that the average treatment effect is zero.

With real data, it will often make sense to proceed in the following steps.

1. Construct and examine a scatterplot of Y against X. The goal is to obtain some initial hunches about how Y is related to X and whether there may be a discontinuity at the threshold.

2. Apply the difference in means estimator with several different window sizes. A key factor will be the number of observations within the window. If there are too few, estimated average treatment effect will be very unstable. But enlarging the window may introduce additional bias. One can use a goodness-of-fit measure such as the cross-validation statistic to help determine the best window size. The goal is to obtain a instructive initial sense of what impact the intervention may be having.

3. Apply a smoother such as lowess or regression splines to the data ignoring the W. Set the tuning parameters so that a very flexible fitting function is applied and then try several different values of these tuning parameters. An overlay of the fitted values on the scatter plot will help to reveal how Y is related to X and what sharp shifts up or down may be apparent for different values of X. Ideally there will be only one, and it will located at the threshold. If there are large discontinuities elsewhere, it may lead to suspicions about the one found at the threshold.

4. Apply a form of non-parametric regression with X and W as predictors. Vary the tuning parameters and use a measure like the cross-validation statistic to pick the best model. If the number of observations is relatively large, any of the common fit measures will likely lead you to the same models. In addition, examine the usual regression diagnostics for to provide additional information about model quality. It will

often turn out that two or three models are about equally good. Then, the results for all three should be reported. For example, the last two estimates in Table 1 are reasonable.

Given the many steps and need for some tuning, it can be a very good idea to randomly partition the data into a training dataset and a test dataset. All of the model building is done with the training dataset. At the end, the model is evaluated with the test dataset. A useful discussion of how to use training data and test data can be found in Haste et al. (2001: 193-196).

8 Conclusions

When the RD design can be implemented properly, it has the same capacity as randomized experiments to obtain unbiased estimates of the average treatment effect. Why then has it not been more widely use? Perhaps the most important reason is that researcher have too often failed to appreciate that a wide variety of social interventions are delivered conditional on some explicit and deterministic rule. Another reason may be that because of the correlation between X and W, the RD design deliver less precise estimates than randomized experiments with the same number of subjects. In practice, statistical power will be less. However, when an RD design is built into the way one or more social interventions are delivered, it is often easy to obtain a very large sample at little additional expense. The bulk of the costs associated with data collection are born by the organization that is responsible for the intervention. In short, the RD design can be a very useful tool that should be far more widely exploited in crime and justice settings.

References

Berk, R.A. (2008) *Statistical Learning from a Regression Perspective*. New York: Springer.

Berk, R.A., and 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(381): 21-27, 1983.

Berk, R.A., and de Leeuw, J. (1999) "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design." *Journal of the American Statistical Association* 94(448): 1045-1052.

Bowman, A.W., Pope, R, and Ismail, B. (2004) "Detecting Discontinuities in Nonparametric Regression Curves and surfaces." *Statistical Computing* 16: 377-390.

Cameron, A.C., and Trivedi, P.K. (2005) *Microeconometrics: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Campbell, D.T., and Stanley, J.C. (1963) *Experimental and Quasi-Experimental Designs for Resaerch*. Boston: Houghton Miffin.

Chen, M.K. and Shapiro, J.M. (2007) "Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-based Approach." *American Law and Economics Review* 9(1): 1-29.

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, R.D., and Weisberg, S. (1999) *Applied Regression Analysis Including Computing and Graphics*. New York: John Wiley and Sons. Fan, J., and Gijbels, I. (1996) *Local Polynomial Regression Modeling and its Applications*. London: Chapman & Hall.

Freedman, D.A. (2008) "Diagnostics Cannot Have Much Power Against General Alternatives." www.stat.berkeley.edu/ freedman/

Goldberger, A.S. (1972) "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Madison, WI. Unpublished manuscript.

Hastie, T.J., and Tibshirani (1990) *Generalized Additive Models*. New York: Chapman and Hall.

Hastie, T.J., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning*. New York: Springer.

Holland,P. (1986) "Statistics and Causal Inference." *Journal of the American Statistical Association* 8:945-60.

Imbens, G., and Lemieux, T. (2008a) "Special Issue Ediutors' Introduction: The Regression Discontinuity Design — Theory and Applications." *Journal of Econometrics* 142: 615-635.

Imbens, G., and Lemieux, T. (2008b) "Regression Discontinuity Designs:A Guide to Practice."' *Journal of Econometrics* 142: 611-614.

Neyman, J. (1923) "Sur Les Applications de la Thorie des Probabilits aux Experiences Agricoles: Essai des Principes." *Roczniki Nauk Rolniczych*10: 151. In Polish.

Rubin, D. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66: 688701.

Thistlewaite, D.L., and Campbell, D.T. (1960) "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Design." *Journal of Educational Psychology* 51: 309-317.

Trochim, W.M.K. (1984) *Research Design for Program Evaluation*. Beverly Hills: Sage Publications.

Trochim, W.M.K. (2001) "Regression Discontinuity Design," in N.J. Smelser and P.B Bates (Eds.) *International Encyclopedia of the Social and Behavioral Sciences*, volume 19: 12940-12945.