George Karabatsos* and Stephen G. Walker

A Bayesian Nonparametric Causal Model for Regression Discontinuity Designs

June 23, 2014

Abstract: For non-randomized studies, the regression discontinuity design (RDD) can be used to identify and estimate causal effects from a "locally-randomized" subgroup of subjects, under relatively mild conditions. However, the accurate estimation of causal effects still relies on a correctly-specified statistical model. For RDDs, we propose a flexible Bayesian nonparametric regression model that can provide accurate estimates of causal effects, either in terms of the predictive mean, variance, quantile, probability density, distribution function, or any other chosen function of the outcome variable. We illustrate the model through the analysis of three real educational data sets, involving (resp.) a sharp RDD, a sharp multivariate RDD, and a fuzzy RDD.

Keywords: Bayesian Nonparametric Regression, Causal Inference, Sharp Regression Discontinuity, Fuzzy Regression Discontinuity, Multivariate Regression Discontinuity.

1 Introduction

A basic objective in scientific research is to infer causal effects from data. Randomized studies are the gold standard of causal inference. In an ideal randomized study, the investigator randomly assigns each subject into one of the treatment

---

1This research is supported by NSF grant SES-1156372. We thank 3 anonymous referees, and the Associate Editor, for their suggestions on a previous version of the manuscript. We also thank Peter Müller for helpful comments.

---
conditions, with equal probability, and each subject complies with her/his treatment assignment. Then, treatment subjects are the same as non-treatment subjects, in terms of the distribution of all observed and unobserved pre-treatment covariates, aside from sampling error (e.g., [45]); and the outcome variable is independent of the chosen treatment intervention, conditionally on the treatment variable [15]. Then the causal effect is given by a comparison of the outcome variable under the treatment intervention, against the outcome variable under the non-treatment intervention.

Often, it is necessary to estimate causal effects from a non-randomized, observational study, because a randomized study can be infeasible due to financial, ethical, or time constraints [45]. However, causal inferences from a non-randomized study is more challenging, because without randomization, treated and non-treated subjects differ almost-surely in terms of the pre-treatment covariates.

The regression discontinuity design (RDD) ([49], [11]) is a type of non-randomized design, where a continuous-valued assignment variable [36] assigns each subject to the treatment (non-treatment, resp.) condition, whenever her/his observed value of the assignment variable equals or exceeds (is less than, resp.) a fixed cutoff value. Under relatively mild conditions, notably when subjects have imperfect control of the assignment variable, the RDD provides a “locally-randomized experiment,” so that ”are as good as randomly assigned” for the subgroup of subjects with assignment variable values near the cutoff [35]. Then the causal effect is identifiable for that subgroup. As proven [24], the RDD can empirically produce causal effect estimates that are similar to those estimates of a standard randomized study ([1], [10], [8], [46], [7], [47]).

The RDD has existed over 50 years, with little initial interest [11]. However, since 1997, more than 74 RDD-based empirical studies have emerged from these fields ([36], [9], [52], [37]), for at least three reasons ([17], [36]). First, many non-randomized studies employ treatment assignment rules that can be easily conceptualized as RDDs. Second, the empirical results of RDDs are intuitive and can be easily conveyed graphically, say, by a plot of the outcomes against the assignment variable. Third, the identification of causal effects in a RDD requires weaker and hence more credible assumptions, compared to the stronger assumptions that are required by other popular causal models, mentioned below. This gives the researcher the flexibility to choose from a range of causal estimation methods.

The other popular causal models for non-randomized studies assume a ”potential outcomes” (counterfactual) framework of causal inference (e.g., [42], [43]). This is typically using notation simplified by the Stable Unit Treatment Value As-
sumption (SUTVA), which implies no interference between subjects and no versions of treatments [44]. The popular models make further assumptions of unconfoundedness (i.e., treatment and non-treatment outcomes are independent of treatment assignments, conditionally on all pre-treatment covariates) and overlap (i.e., there is a chance to receive either the treatment or the non-treatment, conditionally on any value of the pre-treatment covariates) [28]. These models are defined either by a regression of the outcome variable, on variables of treatment receipt and observed pre-treatment characteristics, and/or involves matching/weighting subjects on the observed pre-treatment variables and/or on propensity scores (e.g., [28]). The regression may also be on a hypothesized set of unobserved pre-treatment covariates, in order to study the sensitivity of causal effect estimates over varying degrees of hidden bias (e.g., [41]), i.e., over changes in the distribution of these covariates. However, it may be argued that for typical non-randomized studies, unconfoundedness and overlap are not credible assumptions (e.g., [28], [35]), and even SUTVA is questionable. We provide a more detailed discussion of these issues in Section 2.4.

For RDDs, the mainstream causal models are either linear, polynomial, or local-linear models that employ a regression of the outcome variable on the assignment variable. Such models aim to provide causal inferences in terms of mean comparisons of treatment outcomes and non-treatment outcomes, and to provide sufficiently-flexible modeling of the regression function ([28], [36]), in a neighborhood around the cutoff. However, outliers in a linear model may bias causal effect estimates. Furthermore, in many settings, it may also be of interest to base causal inferences on comparisons of additional features of the outcome variable, such as the variance, quantiles (percentiles), and/or the entire probability density function.

To address these modeling issues, we propose a Bayesian nonparametric regression model [33] for causal inference in RDDs. It is an infinite-mixture model that allows for the entire probability density of the outcome variable to change flexibly as a function of covariates. Our model can provide inferences of causal effects in terms of how the treatment variable impacts the mean, variance, a quantile, probability density function (p.d.f.), distribution function, and any other chosen function of the outcome variable. Finally, the accurate estimation of causal effects relies on an appropriate model for the data. Karabatsos and Walker [33] showed that their Bayesian nonparametric regression model tended to have better predictive performance than other parametric and flexible nonparametric regression models of common usage, over many real data sets.

Also, our model can be extended to handle causal inferences from a fuzzy
RDD [50]. In contrast to a standard ”sharp” RDD, a fuzzy RDD involves a study where not all subjects adhere to the treatment assignment rule. This is because, for example, some subjects do not comply with their respective treatment assignments, or because some subjects receive treatments for which they are not eligible. Moreover, the model can be easily extended to handle an assignment variable that has multiple cutoffs, or to handle a multivariate assignment variable that has a vector-valued cutoff.

In Section 2, we review the data assumptions that are required to identify and estimate causal effects from a RDD. Unlike all previous guides to performing causal inference from RDDs (e.g., [30], [36], [9], [52]), we avoid using the potential outcomes approach to causal inference. This approach has questionable foundations, and besides, causal inference can be handled entirely by standard probability theory ([14], [15]).

In Section 3, we review the current models, and then describe our Bayesian nonparametric model that can estimate causal effects from the various RDDs. In Section 4 we illustrate our model through the analysis of three real educational data sets. These data sets involve (resp.) a sharp RDD, a sharp multivariate RDD, and a fuzzy RDD. Section 5 ends with conclusions.

2 Identifying Causal Effects in a RDD

A non-randomized study from a RDD involves three variables that are observable from each of a sample of $n$ subjects, indexed by $i = 1, \ldots , n$. They are the outcome variable, $Y$; a binary treatment variable $T$, where $T = 1$ refers to treatment receipt and $T = 0$ refers to non-treatment receipt; and a continuous-valued assignment variable $R$. Each subject $i$ is assigned the treatment whenever $R_i \geq r_0$, and is assigned the non-treatment whenever $R_i < r_0$, given a known fixed cutoff $r_0$. The treatment assignment variable is denoted $\mathbf{1}_{R \geq r_0}$, with $\mathbf{1}_{(\cdot)}$ the indicator $(0,1)$ function. An RDD study gives rise to a sample data set, $\mathcal{D}_n = \{(r_i, t_i, y_i)\}_{i=1}^n$, including derived observations $\mathbf{1}_{R_i \geq r_0}$, and possibly observations of $p$ pre-treatment covariates $\mathbf{x}_i = (x_{1i}, \ldots , x_{pi})^\top$.

A characterizing assumption of the RDD is that the conditional probability of treatment receipt is discontinuous at $r_0$, i.e.,

$$\Pr(T = 1 \mid r_0^+) \neq \Pr(T = 1 \mid r_0^-),$$

where $r_0^+ = \lim_{r \uparrow r_0} r$ and $r_0^- = \lim_{r \downarrow r_0} r$. Equation (1) is named Assumption RD.
There are two types of RDDs. In the classical, sharp RDD ([49], [11]), the probability function \( \Pr(T = 1 \mid r) \) has a discontinuous jump of size 1 at \( R = r_0 \), with point mass probability density:

\[
f(t \mid r) = \Pr(T = t \mid R = r) = 1'_{r \geq r_0}(1 - 1'_{r \geq r_0})^{1-t}.
\]  

(2)

Then treatment receipt is identical to treatment assignment, with \( T = 1'_{R \geq r_0} \).

In the fuzzy RDD [50], \( \Pr(T = 1 \mid r) \) has a discontinuous jump that is smaller than 1 at \( r_0 \). Then \( f(t \mid r) \) is not a point mass density. The smaller jump results from imperfect treatment adherence (e.g., treatment non-compliance), where some of the subjects of the given study were either assigned \( 1'_{r \geq r_0} = 0 \) but received \( T = 1 \), or assigned \( 1'_{r \geq r_0} = 1 \) but received \( T = 0 \).

For either type of RDD, a typical measure of the causal effect is given by the difference of conditional means (expectations) of \( Y \) at \( R = r_0 \):

\[
\tau = \mathbb{E}(Y \mid r_0, t = 1) - \mathbb{E}(Y \mid r_0, t = 0) = \mathbb{E}(Y \mid r_0^+, t = 1) - \mathbb{E}(Y \mid r_0^-, t = 0)
\]  

(3a)

where \( \mathbb{E}(Y \mid r, t) = \int y dF(y \mid r, t) \), and \( r_0^+ = r_0^- = r_0 \) when \( R \) is continuous. Conditioning on \( R = r_0 \) is motivated by the fact that \( \Pr(T = 1 \mid r) \) jumps at that point.

In general, for any choice of function \( H \{ \cdot \} \) of \( Y \), the causal effect is:

\[
\tau_H = \mathbb{E}(H \{Y\} \mid r_0^+, t = 1) - \mathbb{E}(H \{Y\} \mid r_0^-, t = 0) = \mathbb{E}(H \{Y\} \mid r_0, t = 1) - \mathbb{E}(H \{Y\} \mid r_0, t = 0),
\]  

(4a)

Therefore, depending on the choice of function \( H \{ \cdot \} \), causal effects are not only interpretable in terms of the mean of \( Y \) (when \( H \{Y\} = Y \)), but also in terms of comparisons of the variance \( (H \{Y\} = \{Y - \mathbb{E}(Y \mid r, t)\}^2) \), cumulative distribution function (c.d.f.) \( F(y \mid r, t) (H \{Y\} = 1_Y \leq y) \), probability density function (p.d.f.) \( f(y \mid r, t) \), survival function \( 1 - F(y \mid r, t) \), and so on. Inverting the c.d.f. obtains \( F^{-1}(u \mid r, t) \), for \( u \in [0, 1] \), so then causal effects can also be interpreted in terms of the \( u \)th quantile of \( Y \).

The outcome \( Y \) (or \( H \{Y\} \)) is observable only under the treatment (non-treatment, resp.) when \( R \geq r_0 \) \((R < r_0, \) resp.), and there are no observations of \( R = r_0 \) due to the continuity of \( R \). Therefore, the estimation of a causal effect \( \tau_H \), from data \( \mathcal{D}_n \), requires some degree of extrapolation.

Next we describe how a causal effect \( \tau_H \) is identified from the sharp and fuzzy RDDs.
2.1 Identification in the Sharp RDD

Given sample data \( D_n \) from a sharp RDD, a causal effect \( \tau_H \) is inferred from \( D_n \) using a model for the conditional probability density function (p.d.f.), \( f(y \mid r, t) \), corresponding to cumulative distribution function (c.d.f.) \( F(y \mid r, t) = \Pr(Y \leq y \mid r, t) \).

Suppose that there is reason to believe that in the absence of treatment, subjects close to the threshold \( r_0 \) are similar. Then the causal effect \( \tau_H \) is identified by assumption RD (1) and that:

\[
f(y \mid r, t = 0) \text{ is continuous in } r \text{ at } r_0, \text{ for all } y. \tag{5}
\]

Equation (5) is a density version of the assumption in [27], who only look at mean shifts and hence assume that

\[
\mathbb{E}(Y \mid r, t = 0) \text{ is continuous in } r \text{ at } r_0. \tag{6}
\]

We believe it is important to model mean shifts not by having a mean shift model but rather by modeling the density of the observations and then picking out the mean from this. This is the correct approach and hence the usefulness of the nonparametric model.

Assuming (5) rather than only (6) allows the treatment effect to exhibit itself in more ways than a mean shift. For example, a variance shift would also be informative, even in the absence of a mean shift. The model we employ is quite general and allows many aspects of treatment effect to be explored by studying any differences between density estimate either side of the cut-off at \( r_0 \). However, we can obviously estimate the key mean shift, having modeled the density functions either side of the cut-off point \( r_0 \), simply by estimating the means of the two density functions.

The assumptions of RD (1) and continuity (5) are sufficient for estimating the treatment effect, when \( R \) is sufficient for the outcome, in the sense that no other variables effect \( Y \) at \( R = r_0 \). Authors such as [36] further elaborate on the continuity assumption, given the existence of observed and unobserved pre-treatment covariates of the subjects, collectively labeled as \( W \), which could also influence the outcome \( Y \). Their elaborations are as follows.

If subjects have imprecise control of \( R \) at \( r_0 \), i.e.,

\[
f(r \mid w) \text{ is continuous in } r \text{ at } r_0, \text{ for all } W = w; \tag{7}
\]

then: (a) local randomization at \( r_0 \) holds, i.e., \( f(w \mid r) \) is continuous in \( r \) at \( r_0 \) for all \( W = w \); (b) the identifying, continuity conditions (5, 6) hold ([27], [35], [36]).
[36]); and (c) the marginal density \( f(r) \) is continuous in \( r \) at \( r_0 \), to provide an empirically-testable consequence of the imprecise control assumption (see [39]). Local randomization is a consequence of imprecise control because the continuity of \( f(r \mid w) \) in \( r \) implies that the right hand side of the Bayes’ rule equation, \( f(w \mid r) = f(r \mid w)f(w) / f(r) \), is continuous in \( r \) [35]. Then, \( f(w \mid r_0^+) = f(w \mid r_0^-) \), i.e., the distribution of \( W \) is the same for treatment subjects and for non-treatment subjects, having \( R \) values located near \( r_0 \). Hence, treatments are “as good as randomly assigned” for this local subgroup of subjects (e.g., [35]).

Under local randomization, \((W,R)\) represent all the information which could influence the outcome \( Y \). So, if

\[
\mathbb{E}(Y \mid r, t = 0) = \int y f(y \mid w, r, t = 0) \, dy
\]  

(8)

is continuous in \( r \) at \( r = r_0 \), then any non-zero value for

\[
\int y \left[ f(y \mid w, r_0^+, t = 1) - f(y \mid w, r_0^-, t = 0) \right] \, dy
\]  

(9)

can only be caused by the difference in treatment. In particular, the continuity of \( f(w \mid r) \) at \( r_0 \) implies that (9) does not depend on \( w \). Therefore, in the remainder of the paper we will assume that \( R \) is sufficient for \( Y \), for the sharp RDD.

2.2 Conditional Independence Properties of the Sharp RDD

The sharp RDD can also be characterized by the extended conditional independence framework of causal inference, extending the ideas from [15], Sections 6.2, 7. Again, we consider the random variables \((R, T, Y)\), and make assumptions RD (1) and imprecise control at \( r_0 \) (7), implying that \( R \) is sufficient covariate and that \( W \) is ignorable. Also, for this framework, we introduce a non-random regime parameter \( \Psi_T \in \{0, 1\} \). An ”idle” or ”observational” regime is indicated by \( \Psi_T = 0 \), when the distribution of \((Y,T,R)\) “arises naturally.” An intervention regime is indicated by setting \( \Psi_T = 0 \) or 1, and such an intervention modifies the idle distribution, as will be described below.

Under a sharp RDD, the joint p.d.f. of \((R,T,Y)\) is:

\[
f(r, t, y) = f(r) f(t \mid r, \Psi_T) f(y \mid r, t).
\]  

(10)

This p.d.f. admits the property \( Y \perp \perp \Psi_T \mid R = r_0, T \), where \( \perp \perp \) denotes conditional independence [13]. This is a causal property, because it says that the distribution
of $Y$ is unaffected by the choice of interventional regime $\Psi_T$, conditionally on $(R = r_0, T)$ [15]. We therefore focus on the conditional p.d.f.:

$$f(t, y | r_0) = f(t | r_0, \Psi_T) f(y | r, t).$$

(11)

If the intervention parameter take on a null value $\Psi_T = \emptyset$, then the joint distribution of the random variables $(T, Y)$ arises naturally, and then (11) reduces to the joint p.d.f. $f(t, y | r_0) = f(t | r_0) f(y | r, t)$, where $f(t | r)$ is the point mass density of equation (2). In contrast, an intervention that sets $\Psi_T = t_0 \in \{0, 1\}$, modifies $f(t | r, \Psi_T)$ to $\mathbf{1}(t = t_0)$, and then changes the p.d.f. (11) to:

$$f(y | r_0, t_0) = f(y | r_0, t_0),$$

(12)

where $||$ denotes ”conditioning by intervention” [34]. The equality in (12) holds by virtue of the causal property. Then for a general choice of function $H\{\cdot\}$, the causal effect is given by a comparison of $E(H\{Y\} | r_0, t)$, for $t = 0, 1$, including the p.d.f.s $f(y | r_0, t)$.

### 2.3 Identification in the Fuzzy RDD

In the fuzzy RDD, the probability function $\Pr(T = 1 | r)$ has a discontinuous jump that is smaller than 1 at $r_0$, meaning that $R$ does not determine $T$. Then $\mathbf{1}_{R \geq r_0}$ and $T$ are distinct variables since the event $\mathbf{1}_{R \geq r_0} \neq T$ is possible; and then $T$ and $Y$ may both depend unobserved confounding variables, collectively labelled as $U$. Given these considerations, we may extend the joint p.d.f. (10) to:

$$f(r, \mathbf{1}_{r \geq r_0} = a, t, y, u) = f(r) \mathbf{1}_{r \geq r_0}^{a} \mathbf{1}_{r < r_0}^{1-a} f(t | r, \mathbf{1}_{r \geq r_0}, u, \Psi_T) f(y | r, t, u) f(u).$$

(13)

This p.d.f. admits the causal property $Y \perp \perp \Psi_T | R, T, U$. But for the purposes of making causal inferences from real data, we cannot condition on $(R, T, U)$, because $U$ is unobserved.

However, for the fuzzy RDD, the assumptions of RD (1) and continuity (5) at $r_0$ (incl. imprecise control) together imply that the causal effect can be represented by:

$$\tau_H = \frac{E(H\{Y\} | r_0^+) - E(H\{Y\} | r_0^-)}{E(T | r_0^+) - E(T | r_0^-)},$$

(14)

for a general choice of function $H\{\cdot\}$ of $Y$. This follows from the equality:

$$E(H\{Y\} | r_0^+) - E(H\{Y\} | r_0^-) = \tau_H \cdot [E(T | r_0^+) - E(T | r_0^-)]$$

(15a)

$$+ [E(H\{Y\} | r_0^+, t = 0) - E(H\{Y\} | r_0^-, t = 0)],$$

(15b)
and the fact that the term (15b) is zero under the continuity assumption (5) [27]. In the ratio (14), the numerator is the Intention to Treat (ITT) effect. The denominator is a measure of treatment adherence, which decreases as noncompliance increases. In the sharp RDD, where $1_{R \geq r_0} = T$, the denominator is 1 (i.e., perfect adherence), and then the ITT effect coincides with the causal effect $\tau_H$.

2.4 On Unconfoundedness and Overlap Assumptions

Arguably, the assumptions of RD (1) and imprecise control at $r_0$ (7) (implying the continuity condition at $r_0$ (5)), which together identify a causal effect $\tau_H$ in either a sharp or a fuzzy RDD, are weaker and hence more credible ([35], p. 679), compared to the unconfoundedness and overlap assumptions of some popular causal models that employ the potential outcomes framework of causal inference (e.g., [42], [28]).

The potential outcomes framework employs notation such as $(Y(1), Y(0))$, denoting the outcomes to treatment ($T = 1$) and to non-treatment ($T = 0$), at a given time point. The use of this specific notation requires SUTVA, which states that the $2^n$ potential outcomes $Y_i(t_n)$, $t_n \in \{0, 1\}^n$ are reducible to two potential outcomes, $(Y_i(1), Y_i(0))$ [44], for every subject $i$ of the $n$ subjects of a given study. Unconfoundedness requires $(Y(1), Y(0)) \perp \perp T \mid w$, and overlap requires $0 < \Pr(T = 1 \mid w) < 1$, for all $w$. If the assumptions of unconfoundedness and overlap hold, then the causal effect of $T$ on $Y$ is identified for all values of the observed and unobserved pre-treatment covariates, $w$ [28]. However, the assumption that unconfoundedness and overlap hold, for all $w$, is questionable [28], and moreover unconfoundedness is not empirically testable. Besides, it may be argued that it is virtually impossible for an investigator of a non-randomized study to have precise prior knowledge of, and then collect data on, all the important pre-treatment covariates.

Albeit, the RDD can identify causal effects only for the subgroup of subjects with $R$ values near $r_0$, pertaining to only a subset of all possible values of the pre-treatment covariates, $w$. Thus, the RDD trades off generalizability, for more credible assumptions in causal effect identification. Also the RDD requires a sufficiently-large number of observations of $R$ near $r_0$, to allow for statistically precise estimation of a causal effect $\tau_H$. Finally, because our review of the RDD avoids the potential outcomes framework, we do not address SUTVA. This assumption does not make sense outside of this framework [14].
3 Estimating Causal Effects In a RDD

Here, we review the current (published) models for estimating causal effects from sample data \( D_n \) of an RDD. Then we propose our Bayesian nonparametric model for causal inference for RDDs. Afterwards, we discuss how any of these models can be extended to handle RDDs with either multiple cutoffs, or with a multivariate assignment variable.

Throughout, we denote \( n(\cdot|\mu, \sigma^2) \) as the density of the normal \( N(\cdot|\mu, \sigma^2) \) distribution (c.d.f.) with mean and variance \( (\mu, \sigma^2) \); with \( \Phi(\cdot) = N(\cdot|0,1) \) the c.d.f. of the normal \( n(\cdot|0,1) \) p.d.f.; \( \text{ga}(\cdot|a,b) \) and \( \text{ig}(\cdot|a,b) \) (resp.) denotes the p.d.f.s of the gamma \( \text{Ga}(\cdot|a,b) \) distribution and inverse gamma \( \text{IG}(\cdot|a,b) \) distribution (c.d.f.), with shape parameter \( a \) and rate parameter \( b \); and \( U(\cdot|a,b) \) is the c.d.f. of a uniform distribution.

3.1 Review of Current Models

For the sharp RDD, the classical model is defined by a least-squares, normal linear polynomial regression of \( Y \) on \( (R, 1_{R \geq r_0}, x^\top) \), which permits a discontinuity at a cutoff \( r_0 \) [11]. The model admits the general form ([9]):

\[
\begin{align*}
  f(y_i| r, 1_{R \geq r_0}, x; \beta, \sigma^2) &= n(y_i| \eta_i, \sigma^2), \ i = 1, \ldots, n, \\
  \eta &= \beta_0 + \beta_1(r) + \tau 1_{r \geq r_0} + \beta_2(r) 1_{r \geq r_0} + \beta_x(x),
\end{align*}
\]

where \( t = 1_{r \geq r_0} \) in a sharp RDD, \( \tau \) gives the conditional average causal effect [3], and \( \beta_m(r) = \sum_{l=1}^{q_m} \beta_{ml} r^l \) is a polynomial of order \( q_m \geq 1 \), for \( m = 1, 2 \). Assuming \( \beta_2(r) = 0 \) (i.e., \( q_2 = 0 \)), a strictly linear model is specified by order \( q_1 = 1 \), and a strictly polynomial model is specified by a general order \( q_1 \geq 1 \). A general polynomial interaction model with separate polynomials \( \beta_1(r) \) and \( \beta_2(r) \) for either side of the cutoff \( r_0 \), is specified by choice of orders \( q_1, q_2 \geq 1 \) (resp.).

Similarly, \( \beta_x(x) = \sum_{k=1}^{\ell} \sum_{l=1}^{q_k} \beta_{xkl} x^l \) is a sum of polynomials for \( p \) observed pre-treatment covariates \( x \), a subset of the covariates that comprise \( W \). Often in practice, \( x \) is excluded from the model [16] (i.e., \( \beta_x(x) = 0 \)), because \( R \) is a sufficient covariate under assumptions RD [1] and continuity [5], making \( W \) (and hence \( x \)) ignorable. Then, excluding the covariates \( x \) from the model has little effect on the bias of the least-squares estimate \( \hat{\tau} \) of the causal effect [3]. However including these covariates in the model may improve estimation efficiency [30].

Under the classical linear model [16], outliers of \( y_i \) can bias estimates of the causal effect \( \tau \), even for observations \( r_i \) far from \( r_0 \) [11]. Therefore, local linear
models [18] were proposed for RDDs, because they can provide more outlier-robust estimation of $\tau$, by placing more emphasis on observations $(y_i, r_i)$ with values $r_i$ near $r_0$. A weighted local linear model (16) assigns smooth kernels $K(\{r_i - r_0\}/h) > 0$ to observations $(y_i, r_i)$ with chosen bandwidth $h > 0$. Kernels that have been considered in the RDD literature include the edge kernel $(K(u) = (1 - |u|)1_{|u| \leq 1})$ and the uniform kernel $(K(u) = 1_{|u| \leq 1/2})$ ([27], [30], [29]). The parameter of the local linear model can be estimated by generalized least-squares. However, the choice of bandwidth parameter can impact the bias of the causal estimate $\hat{\tau}$. For the edge and uniform kernels, [29] describe methods for selecting the optimal bandwidths for RDDs, having large-sample justifications. Another local approach is to specify a model (16) with terms $\beta_m(r)$ $(m = 1, 2)$ each constructed by a spline basis [16]. For example, the truncated linear spline basis, $\beta_m(r) = \beta_{m1}r + \sum_{l=1}^{L_m} \beta_{ml} \max(0, r - r_{ml})$, with $L_k$ knots $r_{ml}$ including $r_0$.

The weighted local linear model was extended to handle quantile regression [19]. This approach entails independent regressions of multiple indicator variables $H\{Y\} = 1_{Y < y_m}$ on $(R, 1_{R \geq r_0})$, for $y_{m-1} < y_m$, $m = 2, \ldots, M$, to construct a conditional quantile $F^{-1}(u | r, 1_{r \geq r_0})$ of $Y$ at $u \in (0, 1)$, via

$$F(y | r, 1_{r \geq r_0}) = \max_{m; y \leq y_m} \mathbb{E}(1_{Y < y_m} | r, 1_{r \geq r_0}).$$

However, this model suffers from the “quantile crossing” problem [6], and relies on an ad-hoc resorting correction that has only large sample justifications. The quantile crossing problem refers to the possibility of incoherent quantile estimates $F^{-1}(u | r, 1_{r \geq r_0}) > F^{-1}(u' | r, 1_{r \geq r_0})$ where $u < u'$, given some $(r, 1_{r \geq r_0})$.

Any of the linear models can be extended to provide an estimate $\hat{\tau}_H$ of a causal effect for a fuzzy RDD. One approach is to take the ratio estimate $\hat{\tau}_H = \hat{\tau}_H^{(F)} = \hat{\tau}_{HY}/\hat{\tau}_{HT}$, for (14), where $\hat{\tau}_{HY}$ is the coefficient estimate of $1_{R \geq r_0}$ in a linear regression of the outcome $Y$ (e.g., $H\{Y\}$) on $(R, 1_{R \geq r_0})$, and $\hat{\tau}_{HT}$ is the coefficient estimate of $1_{R \geq r_0}$ in a linear regression $T$ on $(R, 1_{R \geq r_0})$ ([30], [9], [19]). Again, estimation efficiency can be improved by also conditioning on covariates $x$. The standard error (SE) of the estimate $\hat{\tau}_H^{(F)}$ can be approximated by $\text{SE}(\hat{\tau}_{HY})/\hat{\tau}_{HT}$ [9], with the assumption that $T(r)$ is non-increasing in $r$ at $r_0$, under the potential outcomes framework [30].

Finally, for assessing the fit of a linear model to data, we may inspect standardized residuals $\hat{z}_i = (y_i - \hat{y}_i)/\hat{\sigma}$ of the observations $y_i$ (or $H\{y_i\}$), given least-squares estimates $\hat{\beta} = (1, r, 1_{r \geq r_0}, x_i)\hat{\beta}$). If $|\hat{z}_i| > 2$, then $y_i$ can be judged as an outlier.
3.2 Bayesian Nonparametric Model

For the sharp RDD, our Bayesian nonparametric model [33] is defined by:

\[
f(y_i | r_i, 1_{r_i \geq r_0}; \zeta) = \sum_{j=1}^{\infty} n(y_i | \mu_j, \sigma_j^2) \omega_j(\eta(r_i), \sigma(r_i)), \quad i = 1, \ldots, n \tag{17a}
\]

\[
\omega_j(\eta, \sigma) = \Phi\left(\frac{i-\eta}{\sigma}\right) - \Phi\left(\frac{i-1-\eta}{\sigma}\right) \tag{17b}
\]

\[
\eta(r) = \beta_0 + \beta_1 r + \beta_2 1_{r \geq r_0} \tag{17c}
\]

\[
\sigma(r) = \sqrt{\exp(\lambda_0 + \lambda_1 r + \lambda_2 1_{r \geq r_0})} \tag{17d}
\]

\[
\mu_j, \sigma_j^2 \sim N(\mu_j | \mu_0, \sigma_0^2) \tag{17e}
\]

\[
b_\sigma, \beta, \lambda \sim \text{Ga}(b_\sigma, a_0, b_0)N(\beta, \lambda | 0, vI) \tag{17f}
\]

where the mixture weights \( \omega_j(\eta(r), \sigma(r)) \) sum to 1 at each value of \( r \), and \( T = 1_{r \geq r_0} \) in the sharp RDD. Also, the terms (17a) and (17b) may be deconstructed via the generation of a latent indicator variable \( Z \sim N(\eta, \sigma^2) \), and then taking \( Y \sim N(\mu_j, \sigma_j^2) \) if \( j - 1 < Z \leq j \).

The model (17) allows the entire probability density of the outcome variable \( Y \) to change flexibly as a function of covariates. The parameter \( \sigma(r) \) measures the multimodality of \( f(y \mid r, 1_{r \geq r_0}) \) [33]. Specifically, as \( \sigma(r) \to \infty \), the density \( f(y \mid r, 1_{r \geq r_0}) \) becomes more multimodal, with weights \( \omega_j(\eta(r), \sigma(r)) \) converging to a discrete uniform distribution; and as \( \sigma(r) \to 0 \), the density \( f(y \mid r, 1_{r \geq r_0}) \) becomes more unimodal, and "local," with \( f(y \mid r, 1_{r \geq r_0}) \approx n(y_i | \mu_j, \sigma_j^2) \) and \( \omega_j(\eta(r), \sigma(r)) \approx 1 \) if \( j - 1 < \eta \leq j \). Furthermore, the model has a discontinuity at \( r_0 \) due to the presence of the term \( 1_{r \geq r_0} \) in both (17c) and (17d). The effect, controlled by the coefficients \( (\beta, \lambda) \), is to reallocating the weights either side of \( r_0 \), resulting in different densities either side of this value. Obviously, there is a discontinuity if and only if either of the coefficients \( (\lambda_2, \beta_2) \) is non-zero. The normal prior \( N(\beta, \lambda | 0, vI) \) consists of a prior variance parameter \( v \), which controls for both the prior support for the range of the mixture density component indices \( j = 0, \pm 1, \pm 2, \ldots \) (via the parameter \( \beta \)), and for the range of the level of multimodality in \( f(y \mid r, 1_{r \geq r_0}) \). As \( v \to \infty \), a wider range of component densities and multimodality is supported; and as \( v \to 0 \), \( f(y \mid r, 1_{r \geq r_0}) \) becomes a normal density.

When prior information is limited about the model parameters, we may attempt to specify non-informative priors, for example, by choosing \( \mu_0 = 0, \sigma_0^2 \to \infty \), \( a_0 \to 0, b_0 \to 0 \), and \( v = 10^5 \), and by choosing \( b_\sigma \) according to prior knowledge about range of the \( Y \) variance. For instance, if \( Y \) is known to have a variance...
of 1, then \( b_{\sigma \mu} = 5 \) provides a vague prior choice. For such choices of prior parameters, the Bayesian model (17), over 22 real data sets, demonstrated very good predictive accuracy, and better predictive accuracy compared to many other regression models, and compared to the Bayesian model under different choices of prior [33].

The model (17) has infinite-dimensional parameter, \( \zeta = (\mu_j, \sigma_j^2)_{j=-\infty}^{\infty}, \mu_\mu, \sigma_\mu^2, b_\sigma, \beta, \lambda), \) with prior density \( \pi(\zeta). \) A set of data \( \mathcal{D}_n = \{(y_i, r_i, t_i)\}_{i=1}^n \) updates the prior \( \pi(\zeta) \) to a posterior density, given by

\[
\pi(\zeta | \mathcal{D}_n) = \frac{\prod_{i=1}^n f(y_i | r_i, 1_{r_i \geq r_0}; \zeta) \pi(\zeta)}{\int \prod_{i=1}^n f(y_i | r_i, 1_{r_i \geq r_0}; \zeta) d\Pi(\zeta)},
\]

with \( \Pi(\zeta) \) and \( \Pi(\zeta | \mathcal{D}_n), \) resp. the c.d.f. of \( \pi(\zeta) \) (of \( \pi(\zeta | \mathcal{D}_n) \)). Also, let \( F(y | r, a; \zeta) \) be the c.d.f. of \( f(y | r, 1_{r \geq r_0}; \zeta). \) Then the posterior predictive density, \( f_n(y | r, a), \) and the conditional posterior predictive expectation (\( \mathbb{E}_n \)) and variance (\( \mathbb{V}_n \)) of the outcome \( H\{Y\} \) are given (resp.) by:

\[
f_n(y | r, a) = \int f(y | r, a; \zeta) d\Pi(\zeta | \mathcal{D}_n), \quad 1_{r \geq r_0} \in \{a = 0, 1\},
\]

\[
\mathbb{E}_n(H\{Y\} | r, a) = \int [H\{y\} dF(y | r, a; \zeta)] d\Pi(\zeta | \mathcal{D}_n),
\]

\[
\mathbb{V}_n(H\{Y\} | r, a) = \int \left[ \mathbb{E}_n(H\{Y\} | r, a)^2 dF(y | r, a; \zeta) \right] d\Pi(\zeta | \mathcal{D}_n).
\]

Depending on the choice of function \( H\{\cdot\}, \) the posterior mean \( \mathbb{E}_n \) and variance \( \mathbb{V}_n \) of the conditional expectation \( \mathbb{E}(Y | r, a), \) variance \( \mathbb{V}(Y | r, a), \) c.d.f. \( F(y | r, a) \) at a point \( y, \) are given (resp.) by \( \mathbb{E}_n(\mathbb{E}(Y | r, a)) \) and \( \mathbb{V}_n(\mathbb{E}(Y | r, a)); \mathbb{E}_n(\mathbb{V}(Y | r, a)) \) and \( \mathbb{V}_n(\mathbb{V}(Y | r, a)); \) and \( \mathbb{E}_n(F(y | r, a)) = f_n(y | r, a) \) and \( \mathbb{V}_n(F(y | r, a)). \) For assessing the fit of the Bayesian model to data, a standardized residual for each observation \( y_i \) may be computed by

\[
\zeta_i = \{y_i - \mathbb{E}_n(Y | r_i, 1_{r_i \geq r_0})\} / \{\mathbb{V}_n(Y | r_i, 1_{r_i \geq r_0})\}^{1/2}.
\]

If \( |\zeta_i| > 2 \), then \( y_i \) can be judged as an outlier.

Alternatively, we may consider a version of our model that performs covariate selection, via the specification of spike-and-slab priors [21] for \( (\beta, \lambda). \) Such priors may be defined by prior densities

\[
\pi(\beta_j, \lambda_j | \gamma) = n(\beta_j | v_1 \gamma_j^{(\beta_j)} + v_0(1 - \gamma_j^{(\beta_j)})) n(\lambda_j | v_1 \gamma_j^{(\lambda_j)} + v_0(1 - \gamma_j^{(\lambda_j)}))
\]

\[\gamma_j \sim \text{Bernoulli}(1/2), \quad j = 0, 1, 2.\]
say, for large \( v_1 = 1000 \) and small \( v_0 = .1 \) [33]. Then \( \gamma_j = 1 \) (\( \gamma_j = 0 \), resp.) means that the given covariate is included (excluded, resp.) from the model. In this version of the model, \( \gamma \) would be added to the full parameter vector \( \zeta \), and Bayesian inference of the model would proceed in the same way as described above. A covariate can then be judged to be a "significant predictor" when the posterior probability is \( \Pr[\gamma_j = 1 | \mathcal{D}_n] \geq .5 \) [5].

3.2.1 Estimating Causal Effects With the Bayesian Model

For either version of our Bayesian model, the posterior estimates of the causal effect of \( T \) on \( H\{Y\} \), conditionally on \( R = r_0 \), are given as follows, under the assumptions RD (1) and continuity at \( r_0 \) (5).

For the sharp RDD, the estimate \( \hat{\tau}_H \) of the causal effect is given by

\[
\hat{\tau}_H = \mathbb{E}_n(\tau_H^{(S)}) = \mathbb{E}_n(H\{Y\} | r_0, \mathbf{1}_{r \geq r_0} = 1) - \mathbb{E}_n(H\{Y\} | r_0, \mathbf{1}_{r \geq r_0} = 0),
\]

with posterior variance \( \mathbb{V}_n(\tau_H^{(S)}) = \mathbb{V}_n(H\{Y\} | r_0, 1) + \mathbb{V}_n(H\{Y\} | r_0, 0) \). Then \( \hat{\tau}_H^{(S)} \pm 2[\mathbb{V}_n(\tau_H^{(S)})]^{1/2} \) provides an approximate 95% posterior confidence band around \( \hat{\tau}_H \).

When inferring the causal effect in terms of the \( \alpha^{th} \) quantile, via

\[
\hat{\tau}_H^{(S)} = F_n^{-1}(u | r, 1) - F_n^{-1}(u | r, 0),
\]

we may judge whether \( \hat{\tau}_H \) is significantly different from zero, by using a P-P plot [51] to check for non-overlap of the 95% posterior credible intervals \( (F_n(0.025)(y | r, a), F_n(0.975)(y | r, a)) \) at \( u \), for \( a = 0, 1 \), and over a wide range of \( y \not\in \mathbb{Y} \). Here, \( F_n(0.025)(y | r, a) \) (\( F_n(0.975)(y | r, a) \), resp.) denotes the posterior 2.5%ile (97.5%ile, resp.) of \( F(y | r, a) \).

For the fuzzy RDD, the causal effect estimate \( \hat{\tau}_H \), in terms of the ratio estimator [14], may be obtained by two independent regressions. The first involves estimating the numerator using our regression model (17), and the second involves a regression of \( T \) on \( (R, \mathbf{1}_{R \geq r_0}) \) to estimate the denominator, via the posterior predictive expectations:

\[
\mathbb{E}_n(T | r, \mathbf{1}_{r \geq r_0} = a) = \int \Pr(T = 1 | r, a; \zeta_T) d\Pi(\zeta_T | \mathcal{D}_n), \ a = 0, 1.
\]

Our model (17) can be extended to binary regression, by replacing the model likelihood (17a) with

\[
\Pr(T_i = 1 | r, a; \zeta_T) = \int_0^\infty \left\{ \sum_{j=\pm\infty} \mathbb{N}(t_i^a | \mu_j, \sigma_j^2) \omega_j(\eta(r_i), \sigma_i) \right\} dt_i^a, \ i = 1, \ldots, n.
\]
This provides flexible modeling of the inverse link function by a covariate dependent, infinite mixture of normal c.d.f.s [33]. Since the posterior distributions of these two models are independent, the joint posterior of both models can either be estimated separately or jointly.

Then for the fuzzy RDD, the estimate $\hat{\tau}_H$ of the causal effect, in terms of the ratio (14), is given by:

$$
\hat{\tau}^{(F)}_H = \hat{\tau}^{(S)}_H / \left[ \mathbb{E}_n(T \mid r_0, 1_r \geq r_0 = 1) - \mathbb{E}_n(T \mid r_0, 1_r \geq r_0 = 0) \right] = \hat{\tau}^{(S)}_H / \mathbb{E}_n(D_T). \quad (19)
$$

For example, given the choice of functional $H\{Y\} = 1_{Y \leq y}$, we have the causal effect defined by a comparison of c.d.f.s at a point $y$, weighted by $\mathbb{E}_n(D_T)$, with

$$
\hat{\tau}^{(F)}_{1(Y \leq y)} = \left\{ F_n(y \mid r, 1) - F_n(y \mid r, 0) \right\} / \mathbb{E}_n(D_T).
$$

The general estimator (19) provides a computationally-fast, first-order Taylor approximation to the posterior expectation $\mathbb{E}_n(\hat{\tau}^{(F)}_H)$. The second-order approximation is

$$
\hat{\tau}^{(F[2])}_H = \{ \hat{\tau}^{(S)}_H / \mathbb{E}_n(D_T) \} + \{ (\hat{\tau}^{(S)}_H \mathbb{V}(D_T)) / (\mathbb{E}_n(D_T))^3 \},
$$

with $\mathbb{V}(D_T) = \mathbb{V}_n(T \mid r_0, 1) + \mathbb{V}_n(T \mid r_0, 0)$. The posterior variance $\mathbb{V}_n(\hat{\tau}^{(F)}_H)$ has first-order approximation:

$$
\mathbb{V}_n(\tau^{(F)}_H) \approx \left( \hat{\tau}^{(S)}_H / \mathbb{E}_n(D_T) \right)^2 \left\{ \mathbb{V}_n(\hat{\tau}^{(S)}_H) / \left( \hat{\tau}^{(S)}_H \right)^2 \right\} + \left\{ \mathbb{V}(D_T) / \mathbb{E}_n(D_T) \right\}^2 \right\}.
$$

These approximations are derived from standard results involving the distribution of the ratio of two random variables (e.g., [48], p. 351). Then $\hat{\tau}^{(F)}_H \pm 2\{\mathbb{V}_n(\hat{\tau}^{(F)}_H)\}^{1/2}$ gives a 95% posterior interval band around $\hat{\tau}^{(F)}_H$. Also, when inferring the causal effect $\hat{\tau}^{(F)}_H$ in terms of treatment and non-treatment differences at the $u$th quantile, we may judge whether $\hat{\tau}_H$ is significantly different from zero, by using a P-P plot of the 95% posterior interval bands $F_n(y \mid r, a) \pm 2\{\mathbb{V}_n(\hat{\tau}^{(F)}_H)_{1(y \leq y)}\}$, over $y \in \mathbb{R}$, and then checking for nonoverlap for these bands at point $u$.

Alternatively, it may be of interest to investigate the sensitivity of the causal effect estimate $\hat{\tau}^{(F)}_H$ to variations of treatment adherence (e.g., compliance). This can be achieved by estimating the ratio $\hat{\tau}^{(F)}_H$ for each of a set of fixed non-zero values (e.g., 1, 9, 8, ..., -1) for the denominator, with each estimate having posterior variance $\mathbb{V}_n(\hat{\tau}^{(F)}_H) \approx \left( \hat{\tau}^{(S)}_H / \mathbb{E}_n(D_T) \right)^2 \left\{ \mathbb{V}_n(\hat{\tau}^{(S)}_H) / \left( \hat{\tau}^{(S)}_H \right)^2 \right\}$.

15
Using Markov Chain Monte Carlo (MCMC), Gibbs sampling methods, along with a slice sampling step for $\sigma$, can be used to estimate all of the aforementioned posterior quantities [33]. We use Rao-Blackwell (RB) methods to estimate all the posterior linear functionals, such as $\mathbb{E}_{n}(H\{Y\} | r, a)$, $\mathbb{V}_{n}(H\{Y\} | r, a)$, $r_i$, $\mathbb{E}_{n}(T | r, a)$, $\mathbb{V}_{n}(T | r_0, a)$, $\mathbb{E}_{n}(\tau_{H}^{(S)})$, $\mathbb{V}_{n}(\tau_{H}^{(S)})$, $\mathbb{E}_{n}(\tau_{H}^{(F)})$, and $\mathbb{V}_{n}(\tau_{H}^{(F)})$ [20].

### 3.3 Extensions to Multiple Cutoffs

A RDD may employ an assignment variable $R$ that has multiple cutoff vector $r_0 = (r_{01}, \ldots, r_{0D})^\top$, assignment indicator vector $1_{R \geq r_0} = (1_{R \geq r_{01}}, \ldots, 1_{R \geq r_{0D}})^\top$, and treatment vector $T = (T_1, \ldots, T_D)^\top$ [40]. Then all the same approaches to identifying and estimating the causal effect $\tau_H$ are applicable (Sections 2, 3.1-3.2), after replacing $(r_0, 1_{R \geq r_0})$ with $(r_{0d}, 1_{R \geq r_{0d}})$ (resp.). This is because when there are multiple cutoffs, $R$ is still a sufficient covariate under the assumptions RD (1) and continuity (5) at $r_{0d}$ (i.e., imprecise control at $r_{0d}$).

Alternatively, it is possible to consider a model with covariates $1_{R \geq r_0}$ including $1_{R \geq r_{0d}}$. This approach adds $D - 1$ extra linear predictors (covariates) into the model, and thus this may improve the estimation efficiency of $\hat{\tau}_H$. For the Bayesian model, it is possible to assign spike-and-slab priors to the coefficients of $1_{R \geq r_0}$, in order to provide a type of change-point regression model.

### 3.4 Extensions to Multivariate RDDs

A RDD may involve a multidimensional continuous assignment variable, denoted by $R = (R_1, \ldots, R_D)^\top$, with vector cutoff $r_0 = (r_{01}, \ldots, r_{0D})^\top$ (for $d = 1, \ldots, D$). For such a design, it is possible to reduce $R$ to a scalar assignment variable, $R$, using one of at least two approaches. First, the centering approach [32] defines a new assignment variable by $R = \min_d (R_d - r_{0d})$. The second approach defines a new assignment variable $R$ by a thin-plate spline [25], with $R = ||R - r_0|| \log ||R - r_0||$, and with $|| \cdot ||$ the Euclidean norm. If either approach is used, then all the same approaches to identifying and estimating the causal effect $\tau_H$ are applicable (Sections 2, 3.1-3.2) to the new scalar assignment variable $R$ and cutoff $r_0 = 0$.

### 4 Illustrative Applications

The Bayesian nonparametric model was illustrated through the analysis of three data sets, using menu-driven software that was developed by the first author ([32],
The first two data sets were collected from four Chicago University schools of education, which established a new curriculum that aims to train and graduate teachers to improve Chicago public schools. Each of these data sets involved a sharp RDD, using (resp.) a scalar-valued and a multivariate assignment variable. The third data set, obtained from [4], involves a fuzzy RDD, from a study of the effect of class size on student achievement [3]. For each of the three data sets, it seems reasonable to make the assumptions of RD (1) and continuity (5) at \( r_0 \) (i.e., imprecise control at \( r_0 \)), in order to identify the causal effects of treatment on the outcome, conditionally on \( r_0 \) (see Section 2).

For all three data sets, the Bayesian nonparametric model assumed the same vague priors that were mentioned in Section 3.2. The model under these priors provided good predictive accuracy for each of these three data sets, and better predictive accuracy compared to the model under different choices of these priors. Also, the model under the vague priors yielded similar estimates of causal effects, compared to the model with the spike-and-slab priors. Hence, in this section, we will focus on presenting the results of the Bayesian model under the vague priors.

All posterior estimates of this model, reported in the next three subsections, are based on 40K MCMC samples that were obtained from every 5th iterate of a run of 200K MCMC sampling iterations, after discarding the first 2K burn-in samples. This provided accurate posterior estimates according to standard convergence assessments [22]. Specifically, univariate trace plots displayed good mixing of model parameters and posterior predictive samples, and all posterior predictive estimates obtained 95% MC confidence intervals with half-width sizes near .01. Also, for comparative purposes, we also analyzed each of the three data sets using different versions of the standard linear model (16), under least-squares estimation. They including the locally-weighted linear models with optimal bandwidth selection [29].

### 4.1 Learning Math Teaching: Time Series Data

For the first data set, the aim is to estimate the effect of the new teacher education curriculum on math teaching ability, among \( n = 347 \) undergraduate teacher education students attending one of four Chicago universities. This data set involves a sharp RDD, an interrupted time-series design [12] using an assignment variable of time, ranging from fall semester 2007 through spring semester 2013. The new curriculum (treatment) was instituted in Fall 2010 (the cutoff, \( r_0 \)), and the old teacher curriculum (non-treatment) was active before then. The outcome variable (\( Y \)) is the number-correct score on the 25-item Learning Math for Teaching (LMT)
test [38]. Each of the students completed the LMT test (89.9% female; 135 and 212 students under the old and new curriculum), after finishing a course on teaching algebra. Among them, the average LMT score was 12.9 (s.d. = 3.44), with Cronbach’s alpha reliability .63. The LMT scores were transformed to z-scores with mean 0 and variance 1.

Using our Bayesian model, we analyzed the data to estimate the effect of the new curriculum, versus the old curriculum, on student ability to teach math (LMT score), at the Fall semester 2010 cutoff. The model included the LMT test z-score as the dependent variable ($Y$), and included covariates of the assignment variable ($R$), given by $\text{TimeF10} = \text{Year} - 2010.6$, and of the treatment assignment variable $\text{CTPP} = 1_{\text{Year} \geq 2010.6}$. The cutoff 2010.6 is the time midpoint between Spring 2010 (2010.3) and Fall 2010 (2010.9).

For the model, R-squared was .99, and nearly all the standardized residuals ranged between $-1$ and 1 over the 347 observations, with one residual slightly exceeding 2. Figure 1 presents the model’s posterior predictive density estimate of the LMT outcome, for the new curriculum (treatment) and for the old curriculum (non-treatment), at Fall 2010. The new curriculum, compared to the old, increased the LMT scores, by shifting the density of LMT scores to the right. This shift corresponds to an increase in the mean (from $-1.17$ to $-1.13$), the 10%ile ($-2.01$ to $-1.97$), and the 25%ile ($-1.43$ to $-1.31$), but these increases were not statistically significant from zero according to 95% credible intervals of the predictive mean and of the posterior c.d.f. estimates. Also, each density presents two modes (clusters) of students.

---

Figure 1 here

---

For the data analysis, we also considered several linear models. We considered strictly linear and polynomial models, and linear/polynomial interaction models, up to order 5. The strictly linear, quadratic, and linear interaction models obtained the best values of the Akaike’s Information Criterion (AIC) [2]. They yielded (resp.) causal effect estimate of .36 (2-sided $p$-value = .06), .40 ($p = .04$), and .40 ($p = .04$). The local linear models, with edge and uniform kernels (resp.), yielded mean causal effect estimates of .38 ($p = .02$), and .40 ($p = .04$). However, all of these linear models had between 15 to 16 outliers.
4.2 Basic Skills and Teaching: Multivariate RDD Data

The second data set, from a sharp RDD, contains information on \( n = 205 \) undergraduate teacher education students who enrolled into one of the four Chicago education schools, at the Fall semester of 2010, 2011, or 2012 (90% female; mean age = 22.5, s.d. = 5.35, \( n = 203 \)); 47%, 21%, 10%, and 22% attended the 4 universities, resp.; 49%, 41% and 10% enrolled in 2010, 2011, and 2012, resp.).

Most U.S. schools of education base their undergraduate admissions decisions on the ability to pass basic skills tests. We investigate the causal effect of basic skills on teacher performance, as previously studied in the literature on teacher education (e.g., [23]).

Here, the assignment variable is a 4-variate random variable, defined by Illinois Test of Basic Skills subtest scores, in reading, language, math and writing. Each subtest has a minimum passing score of 240. The dependent variable (\( Y \)) is total score on the 50-item Haberman Teacher Pre-screener assessment. A score in the 40-50 range indicates a very effective teacher. This assessment has a test-retest reliability of .93, has a 95% accuracy rate in predicting which teachers will stay and succeed in the teaching profession, and is used by many schools to assess applicants of teaching positions [26]. Over the 205 students, the Haberman scores averaged 29.82 (s.d. = 4.3), and were later transformed to z-scores with mean zero and variance 1. The average basic skills score in reading (Read), language (Lang), math (Math), and writing (Write) was (resp.) 204.69 (s.d. = 33.7), 204.20 (s.d. = 35.8), 212.84 (s.d. = 42.2), and 238.40 (s.d. = 23.8). A scalar-valued assignment variable, \( B_{240d10} = \frac{\min(\text{Read}, \text{Lang}, \text{Math}, \text{Write}) - 240}{10} \), was defined using the centering method (Section 3.4).

Using the Bayesian model, we analyzed the data set to estimate the causal effect of passing the reading basic skills exam (treatment), versus not passing (non-treatment), on students’ ability to teach in urban schools. The model included the Haberman z-score as the outcome (dependent) variable (\( Y \)), and included as covariates the assignment variable (\( R \)), \( B_{240d10} \), and an indicator of passing all four basic skills subtests \( \text{BasicPass} = I_{B_{240d10} \geq 0} \).

For the data, the Bayesian model obtained an R-squared of .98, with standardized fit residuals ranging from −1.7 to 1.3 over the 205 observations. Thus the model had no outliers. Figure 2 presents the model’s posterior predictive density estimates, for the treatment versus the non-treatment (resp.). A more detailed inspection revealed that passing the basic skills reading test causally increased the Haberman z-score, in terms of the mean (from −.10 to .09), 25%ile (−1.11 to −1.06), median (−.19 to .05), 75%ile (.74 to 1.16), 90%ile (1.44 to 2.36), and
95%ile (2.18 to 2.82), and causally decreased in terms of the 5%ile (−1.69 to 
−3.66). Also, the treatment density and the non-treatment density each showed 2 
modes (clusters) of students.

We also analyzed the data using several linear regression models. We con-
sidered strictly linear and polynomial models, and linear/polynomial interaction 
models, up to order 5. Among these models, the linear interaction and strictly 
cubic models attained the best AIC scores, and (resp.) yielded estimated causal 
effects of 1.22 (2-sided $p$-value = .00) and 1.06 ($p = .02$). These models had be-
tween 6 and 7 outliers. The local linear models with edge and uniform kernels 
(resp.), yielded estimated causal effects of 1.11 ($p = .01$), and 1.21 ($p = .02$), and 
6 and 2 outliers.

4.3 Maimonides’ Data: Fuzzy RDD

The twelfth-century rabbinic scholar Maimonides proposed a rule that specifies 
a maximum class size of 40, under the belief that smaller class sizes promotes 
higher student achievement. Specifically, for a given class $c$ in school $s$, the rule 
assigns average class size ($Psize_{sc}$) as a function of beginning-of-the-year school 
enrollment ($e_s$), according to the prediction equation $Psize_{sc} = e_s/\text{floor}[(e_s - 
1)/40] + 1].$ The rule (equation) assigns students of a school into a single class-
room, when the school’s enrollment is less than 41, assigns students into two 
classrooms of average size 20.5 when school enrollment reaches 41; assigns stu-
dents into three classrooms of average size 27 when enrollment reaches 81, and 
so on. The cutoff number 20.5 distinguishes between small and large classes.

Here, we study the effect of class size on average class verbal achievement, 
through the analysis of data on 4th grade students who each attended one of 2,056 
classes in Israeli public schools during 1991. These schools used Maimonides’ 
rule to allocate students into classrooms. Demographic statistics are reported in 
[3] (three other classes were not analyzed because they had missing achievement 
data). For the Bayesian model, the dependent variable ($Y$) is average class verbal 
score (avgverb), which we transformed to z-scores with mean 0 and variance 1. 
The covariates include the assignment variable ($R$), defined by the rule-predicted 
class size centered at the cutoff 20.5 (i.e., $Psize_{205} = Psize - 20.5$), and include
the indicator of large (vs. small) class assignment, \( P_{\text{large}} = 1_{P_{\text{size}} \geq 20.5} \). Now, while Maimonides’ rule may assign a given class to be a large (small, resp.), the class could become small (large, resp.). For example, one school in the data set had an enrollment of 41, leading to some students receiving a large class of 21, and other students receiving a small class of 20. Therefore, the data arise from a fuzzy RDD, and for the data analysis, we also consider a variable defined by the indicator of large class receipt, \( \text{large} = 1_{\text{classize} \geq 20.5} \). We also fit the Bayesian model, with the treatment \((T)\) variable, \( \text{large} \), as the dependent variable, and with covariates \( \text{Psize205} \) and \( \text{Plarge} \).

For the avgverb \((Y)\) dependent variable, the Bayesian model obtained an \( R^2 \) squared of .88, with standardized fit residuals ranging from \(-1.1\) to \(1.3\) over the 2,056 observations. Thus the model had no outliers. For the treatment \((T)\) dependent variable, large, the Bayesian model had no outliers, and estimated .93 as the denominator of the causal effect estimator (14). Figure 3 presents the model’s posterior predictive density estimates, for the treatment versus the non-treatment. The left panel of the figure presents the density estimates (expectations) of avgverb, in terms of the numerator of the effect estimator (14), obtained from the Bayesian model with avgverb as the dependent variable \((Y)\). The right panel presents the density estimates divided by .93. It was found that passing the basic skills reading test causally increased class size, in terms of the 5\%ile \((-2.45\) to \(-2.19\)), 10\%ile \((-1.81\) to \(-1.66\)), 25\%ile \((-1.94\) to \(-1.88\)), and causally decreased in terms of the 75\%ile \((.90\) to \(.79\)), 90\%ile \((1.48\) to \(1.37\)), and 95\%ile \((1.80\) to \(1.66\)). Each of these estimates are based on taking the predictive quantile estimates of the Bayesian model for avgverb \((Y)\), and dividing them by .93.

---

Various linear models were also fit to the data. For the linear regression modeling of avgverb \((Y)\) on \((\text{Psize205, Plarge})\), and for the linear regression modeling of large \((T)\) on \((\text{Psize205, Plarge})\), we considered strictly linear and polynomial models, and linear and polynomial interaction models, up to order 9. For the avgverb dependent variable, the septic (order 7) interaction model yielded the best AIC value, however, it also yielded a large least-squares estimate \( \hat{\tau}_Y = 3.33 \) (\( SE = 1.59 \)) and 75 outliers. For the large \((T)\) dependent variable, the linear interaction model (order 1) attained the best AIC value, had no outliers, and yielded \( \hat{\tau}_T = .94 \) (\( SE = .02 \)). Then the estimate of the causal effect of large (versus small)
class is $\hat{\tau} = \hat{\tau}_Y / \hat{\tau}_T = 3.33 / .94 = 3.54$, with approximate SE = $1.59 / .94 = 1.69$. The local linear models with edge and uniform kernels (resp.), for the avgverb ($Y$) dependent variable, yielded $\hat{\tau}_Y$ estimates of $.82$ (SE = .23), and $.39$ (SE = .27), and 144 and 31 outliers (resp.); and for the large ($T$) dependent variable, yielded $\hat{\tau}_T$ estimates of $.80$ (SE = .05), and $.84$ (SE = .05), with no outliers. Then the local edge kernel model yielded an estimate of the causal effect of large (versus small) class as $\hat{\tau} = \hat{\tau}_Y / \hat{\tau}_T = .82 / .80 = 1.03$, and the local linear uniform model yielded a causal effect estimate of $\hat{\tau} = \hat{\tau}_Y / \hat{\tau}_T = .39 / .84 = .46$, with approximate standard errors (resp.) $23 / .80 = .29$ and $.27 / .84 = .32$. For all the linear models, the average causal effects varied widely, and there were many outliers for $Y$.

5 Conclusions

We proposed and illustrated a flexible Bayesian nonparametric regression model for causal inference in various RDDs. Such designs identify causal effects under relatively mild conditions. We demonstrated that our model improves upon the existing linear models for RDDs, both in terms of information yield, and in terms of predictive accuracy of the outcome data, as indicated by analyses of model fit residuals. While the linear models only focus on mean causal effects, the Bayesian model provides inferences of causal effects, in terms of the mean, variance, distribution function, quantile, probability density, or any other functional of the outcome variable.

References

[1] Aiken, L., S. West, D. Schwalm, J. Carroll, and S. Hsiung (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.

[2] Akaike, H. (1973): “Information theory and the an extension of the maximum likelihood principle,” in B. Petrov and F. Csaki, eds., Second International Symposium On Information Theory, Budapest: Academiai Kiado, 267–281.

[3] Angrist, J. and V. Lavy (1999): “Using Maimonides’ rule to estimate the effect of class size on scholastic achievement,” The Quarterly Journal of Economics, 114, 533–575.
[4] Angrist, J. and V. Lavy (2008): “Replication data for: Using Maimonides’ rule to estimate the effect of class size on student achievement,” [http://thedata.harvard.edu/dvn/dv/JAngrist/](http://thedata.harvard.edu/dvn/dv/JAngrist/) accessed: 2014-06-05.

[5] Barbieri, M. and J. Berger (2004): “Optimal predictive model selection,” *Annals of Statistics*, 32, 870–897.

[6] Bassett, G. and R. Koenker (1982): “An empirical quantile function for linear models with iid errors,” *Journal of the American Statistical Association*, 77, 407–415.

[7] Berk, R., G. Barnes, L. Ahlman, and E. Kurtz (2010): “When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment,” *Journal of Experimental Criminology*, 6, 191–208.

[8] Black, D., J. Galdo, and J. Smith (2005): “Evaluating the regression discontinuity design using experimental data,” Unpublished manuscript.

[9] Bloom, H. (2012): “Modern regression discontinuity analysis,” *Journal of Research on Educational Effectiveness*, 5, 43–82.

[10] Buddelmeyer, H. and E. Skoufias (2004): *An evaluation of the performance of regression discontinuity design on PROGRESA*, World Bank Publications.

[11] Cook, T. (2008): “Waiting for life to arrive: A history of the regression discontinuity design in psychology, statistics and economics,” *Journal of Econometrics*, 142, 636–654.

[12] Cook, T. and D. Campbell (1979): *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Chicago: Rand McNally.

[13] Dawid, A. (1979): “Conditional independence in statistical theory,” *Journal of the Royal Statistical Society, Series B*, 41, 1–31.

[14] Dawid, A. (2000): “Causal inference without counterfactuals,” *Journal of the American Statistical Association*, 95, 407–424.

[15] Dawid, A. (2002): “Influence diagrams for causal modelling and inference,” *International Statistical Review*, 70, 161–189.
[16] de Boor, C. (2001): *A Practical Guide to Splines (Revised Edition)*, New York: Springer.

[17] der Klaauw, W.V. (2008): “Regression-discontinuity analysis: A survey of recent developments in economics,” *Labour*, 22, 219–245.

[18] Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall/CRC.

[19] Frandsen, B., M. Frölich, and B. Melly (2012): “Quantile treatment effects in the regression discontinuity design,” *Journal of Econometrics*, 168, 382–395.

[20] Gelfand, A. and S. Mukhopadhyay (1995): “On nonparametric Bayesian inference for the distribution of a random sample,” *Canadian Journal of Statistics*, 23, 411–420.

[21] George, E. and R. McCulloch (1997): “Approaches for Bayesian variable selection,” *Statistica Sinica*, 7, 339–373.

[22] Geyer, C. (2011): “Introduction to MCMC,” in S. Brooks, A. Gelman, G. Jones, and X. Meng, eds., *Handbook of Markov Chain Monte Carlo*, Boca Raton, FL: CRC, 3–48.

[23] Gitomer, D., T. Brown, and J. Bonett (2011): “Useful signal or unnecessary obstacle? The role of basic skills tests in teacher preparation,” *Journal of Teacher Education*, 62, 431–445.

[24] Goldberger, A. (2008/1972): “Selection bias in evaluating treatment effects: Some formal illustrations,” in D. Millimet, J. Smith, and E. Vytlacil, eds., *Modelling and evaluating treatment effects in economics*, Amsterdam: JAI Press, 1–31.

[25] Green, P. and B. Silverman (1993): *Nonparametric regression and generalized linear models: A roughness penalty approach*, Los Altos, CA: CRC Press.

[26] Haberman, M. (2008): *The Haberman Star Teacher Pre-Screener*, Houston: The Haberman Educational Foundation.
[27] Hahn, J., P. Todd, and W.V. der Klaauw (2001): “Identification and estimation of treatment effects with a regression-discontinuity design,” *Econometrica*, 69, 201–209.

[28] Imbens, G. (2004): “Nonparametric estimation of average treatment effects under exogeneity: A review,” *The Review of Economics and Statistics*, 86, 4–29.

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

[30] Imbens, G.W. and T. Lemieux (2008): “Regression discontinuity designs: A guide to practice,” *Journal of Econometrics*, 142, 615–635.

[31] Karabatsos, G. (2014): “Bayesian Regression: Nonparametric and parametric models, version 2014b. Software users manual,” [http://www.uic.edu/georgek/HomePage/BayesSoftware.html](http://www.uic.edu/georgek/HomePage/BayesSoftware.html).

[32] Karabatsos, G. (2014): “Bayesian Regression: Nonparametric and parametric models, version 2014x,” [http://www.uic.edu/georgek/HomePage/BayesSoftware.html](http://www.uic.edu/georgek/HomePage/BayesSoftware.html).

[33] Karabatsos, G. and S.Walker (2012): “Adaptive-modal Bayesian nonparametric regression,” *Electronic Journal of Statistics*, 6, 2038–2068.

[34] Lauritzen, S. (2000): “Causal inference from graphical models,” in O. Barndorff-Nielsen, D. Cox, and C. Kluppelberg, eds., *Complex Stochastic Systems*, London: CRC Press, 63–107.

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

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

[37] Li, F., A. Mattei, and F. Mealli (2013): “Bayesian inference for regression discontinuity designs with application to the evaluation of Italian university grants,” Technical report, Department of Statistics, Duke University.

[38] LMT (2012): *Learning Mathematics for Teaching (LMT) Assessment*, Ann Arbor, MI: University of Michigan.
[39] McCrary, J. (2008): “Manipulation of the running variable in the regression discontinuity design: A density test,” *Journal of Econometrics*, 142, 698–714.

[40] Riecken, H., R. Boruch, D. Campbell, N. Caplan, T. Glennan, J. Pratt, A. Rees, and W. Williams (1974): *Social experimentation: A method for planning and evaluating social intervention*, New York: Academic Press.

[41] Rosenbaum, P. and D. Rubin (1983): “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome,” *Journal of the Royal Statistical Society, Series B*, 45, 212–218.

[42] Rubin, D. (1974): “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 66, 688–701.

[43] Rubin, D. (1978): “Bayesian inference for causal effects: The role of randomization,” *Annals of Statistics*, 6, 34–58.

[44] Rubin, D. (1990): “Neyman (1923) and causal inference in experiments and observational studies,” *Statistical Science*, 5, 472–480.

[45] Rubin, D. (2008): “For objective causal inference, design trumps analysis,” *The Annals of Applied Statistics*, 2, 808–840.

[46] Schochet, P. (2009): “Statistical power for regression discontinuity designs in education evaluations,” *Journal of Educational and Behavioral Statistics*, 34, 238–266.

[47] Shadish, W., R. Galindo, V. Wong, P. Steiner, and T. Cook (2011): “A randomized experiment comparing random and cutoff-based assignment.” *Psychological Methods*, 16, 179-191.

[48] Stuart, A. and K. Ord (1998): *Kendall’s Advanced Theory of Statistics: Volume 1: Distribution Theory*, New York: Wiley.

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

[50] Trochim, W. (1984): *Research design for program evaluation: The regression discontinuity approach*, Newbury Park, CA: Sage.
[51] Wilk, M. and R. Gnanadesikan (1968): “Probability plotting methods for the analysis of data,” *Biometrika, 55*, 1–17.

[52] Wong, V., P. Steiner, and T. Cook (2013): “Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods,” *Journal of Educational and Behavioral Statistics, 38*, 107–141.
Figure Captions

**Figure 1:** Posterior predictive density estimates of $Y$, under treatment ($T = 1$, red), and under non-treatment ($T = 0$, blue).

**Figure 2:** Posterior predictive density estimates of $Y$, under treatment ($T = 1$, red), and under non-treatment ($T = 0$, blue).

**Figure 3:** Figure 3: Posterior predictive density estimates of $Y$, under treatment ($T = 1$, red), and under non-treatment ($T = 0$, blue). Left panel: Densities of ITT effects. Right panel: ITT densities divided by .934.
CTPP = 0 (Blue) vs. 1 (Red)
TimeF10 = 0
Density p.d.f.

ITT Effect

Causal Effect

large = 0 (Blue) vs. 1 (Red)