Over-Identified Regression Discontinuity Design*

Carolina Caetano† *University of Rochester*

Gregorio Caetano‡ *University of Rochester*

Juan Carlos Escanciano§ *Indiana University*

October, 2017.

Abstract

We propose a new identification and estimation strategy for the Regression Discontinuity Design (RDD). Our approach explores the heterogeneity in the "first stage" discontinuities for different values of a covariate to generate over-identifying restrictions. This allows us to identify quantities which cannot be identified with the standard RDD method, including the effects of multiple endogenous variables, multiple marginal effects of a multivalued endogenous variable, and heterogeneous effects conditional on covariates. Additionally, when this method is applied in the standard RDD setting (linear model with a single endogenous variable), identification relies on a weaker relevance condition and has robustness advantages to variations in the bandwidth and heterogeneous treatment effects. We propose a simple estimator, which can be readily applied using packaged software, and show its asymptotic properties. Then we implement our approach to the problem of identifying the effects of different types of insurance coverage on health care utilization, as in Card, Dobkin and Maestas (2008). Our results show that Medicare eligibility affects health care utilization both through the extensive margin (i.e., one insurance vs. no insurance) and the intensive margin (i.e., more generous insurance vs. less generous insurance). While the extensive margin affects the utilization of recurrent, lower cost, care (e.g., doctor visits), the intensive margin affects the utilization of sporadic, higher cost, care (e.g., hospital visits).

Keywords: Regression Discontinuity Design; Multiple Treatments; Varying Coefficients; Nonparametric; Medicare; Health Insurance.

JEL classification: C13; C14; C21; I13

*We would like to thank seminar participants at Boston College, Boston University, Harvard-MIT, Northwestern, University of Michigan, University of Colorado-Boulder and the 2015 World Congress of the Econometric Society for useful comments.

†Department of Economics, University of Rochester, 238 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627-0156, USA. E-mail: carol.caetano@rochester.edu. Web Page: http://www.carolinacaetano.net/.

‡Department of Economics, University of Rochester, 238 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627-0156, USA. E-mail: gregorio.caetano@gmail.com. Web Page: http://www.gregoriocaetano.net/.

§Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/. Research funded by the Spanish Plan Nacional de I+D+I, reference number ECO2014-55858-P.

1 Introduction

Regression Discontinuity Design (RDD) has emerged as one of the most credible identification strategies in the social sciences; see Imbens and Lemieux (2008) and Lee and Lemieux (2010) for early surveys of the literature and Cattaneo and Escanciano (2017) for a more recent overview. We propose a new method of identification and estimation in the RDD setting which uses the variations on the first stage discontinuities of the endogenous variable(s) at the threshold for different values of a covariate. Since the covariate can be multivalued, this generates over-identifying restrictions, which can be leveraged in several directions unexplored in the standard RDD setting.

Our paper relates to a number of papers discussing the inclusion of covariates in the standard RDD setting; see, e.g., Imbens and Lemieux (2008), Frolich (2007) and Calonico, Cattaneo, Farrell and Titiunik (2016). The specific way in which we include the covariates is new. We show that covariates can be used to improve identification in the standard RDD setting as well as to achieve identification in settings in which the standard RDD cannot be used at all, as we discuss next.

Although the RDD literature has been mainly focused on the case of a binary treatment variable, there are many applications of the RDD methodology where the treatment variable of interest can take multiple values; see, for example, empirical applications in Angrist and Lavy (1999), Chay and Greenstone (2005), Ludwig and Miller (2007), Card, Dobkin and Maestas (2008), Carpenter and Dobkin (2009) and Brollo, Nannicini, Perotti and Tabellini (2013), among others. In such cases the standard RDD identifies the compound effect of all the changes, but not the specific effects. Thus, for example, if there are two treatment levels, the RDD identifies a weighted average of the effects of the changes in both treatment levels, but not the effect of each treatment level separately. To the best of our knowledge, this article proposes the first identification and estimation strategy for RDD with multiple treatment variables.[1] This includes cases in which there are multiple endogenous variables, one multivalued endogenous variable with varying marginal effects, and models with varying coefficients (heterogeneous effects) conditional on covariates.

In the standard RDD setting (linear model with a single endogenous variable), our approach can achieve identification when the expected value of treatment conditional on a covariate varies with the covariate, even if on average the expected value of treatment is the same across the threshold and thus a standard RDD approach would not be able to identify the treatment effect. When both approaches can be applied, our estimation strategy has efficiency advantages due to its over-identifying restrictions, leading to smaller mean squared errors in our simulations that are less sensitive to the bandwidth choice.

We apply our approach to the problem of estimating the effects of insurance coverage on health care utilization with a regression discontinuity design, as in Card, Dobkin and Maestas (2008). They exploit the fact that Medicare eligibility varies discontinuously at age 65. Medicare eligibility may affect health care utilization via two channels: (1) the extensive margin, because Medicare eligibility provides coverage to people who were previously uninsured, and (2) the intensive margin, because

[1] Not to be confused with having multiple running variables or multiple thresholds, for which several proposals are available; see the review in Cattaneo and Escanciano (2017).

it provides more generous coverage to people who were previously insured by other, less generous, insurance policies. We find that minorities and people with less education are more likely to be affected by Medicare eligibility in the extensive margin (i.e., one insurance vs. no insurance), while Whites and people with higher education are more likely to be affected by Medicare eligibility in the intensive margin, (i.e., more generous vs. less generous insurance). Using our approach to exploit this heterogeneity in the first stage allows us to identify the partial effects at both these margins. While the extensive margin seems to matter for recurrent, lower cost, care utilization (e.g., doctor visits), the intensive margin seems to matter for sporadic, higher cost, care utilization (e.g., hospital visits).

The paper is organized as follows. Section 2 develops our approach as well as the identification conditions. Section 3 presents the estimator and its asymptotic behavior, as well as a simple implementation strategy. Section 4 reports the results of Monte Carlo experiments, including comparisons to the standard RDD in cases where both approaches could be applied. Section 5 presents the application of our method to the problem of identifying the effects of different types of insurance coverage on health care utilization. Finally, we conclude in Section 6. An Appendix contains general results on identification and proofs of the identification and asymptotic results.

2 Identification

We begin with a simple model with homogeneous (in covariates) marginal effects for exposition purposes. See the Appendix Section A for a more general model, and Section 2.3 for a very applicable model with variable marginal effects. Consider the following model for the outcome variable Y_i,

$$Y_i = g(T_i, W_i) + h(Z_i, W_i, \varepsilon_i), \tag{1}$$

where T_i is a treatment variable, whose effects summarized in g are of interest, Z_i is a vector of covariates, W_i is the so-called running variable, and ε_i is an unobservable error term. We assume that T_i is multivalued, with discrete support $\mathcal{T} = \{t_0, ..., t_d\}$, $d < \infty$, and where $t_0 < t_1 < \cdots < t_d$. Define the d-dimensional vector $X_i = (1(T_i \geq t_1), ..., 1(T_i \geq t_d))'$, where $1(A)$ denotes the indicator function of the event A and B' denotes the transpose of B. Define $\alpha(w) = g(t_0, w)$, and define the vector of marginal effects $\beta(w) = (g(t_1, w) - g(t_0, w), ..., g(t_d, w) - g(t_{d-1}, w))'$, where w is in the support of W_i. So, the $j-$th component of $\beta(w)$ represents the marginal effect of moving from treatment level t_{j-1} to t_j at $W_i = w$ (e.g., in our application β_1 represents the marginal effect of going from $t_1 = 1$ (one or more health insurance policies) to $t_2 = 2$ (two or more health insurance policies) at the age of 65, i.e. $w = 65$. Then, noting that

$$g(T_i, W_i) = g(t_0, W_i) + [g(t_1, W_i) - g(t_0, W_i)]\, 1(T_i \geq t_1) + \cdots + [g(t_d, W_i) - g(t_{d-1}, W_i)]\, 1(T_i \geq t_d),$$

and defining $H_i = \alpha(W_i) + h(Z_i, W_i, \varepsilon_i)$, the model can be succinctly written as

$$Y_i = \beta(W_i) X_i + H_i, \tag{2}$$

where $\beta(\cdot)$ is the parameter of interest, and the distribution of H_i given X_i and W_i are nuisance parameters. We further assume that the running variable W_i is univariate and continuously distributed.

Note that the results that follow are all based on equation (2), and thus they are also valid for an arbitrary vector X (not necessarily a vector of binary indicators), provided that equation (2) holds. Thus, for example, if one is interested in estimating the effects of having an additional alcoholic drink on mortality rates in a linear model, as in Carpenter and Dobkin (2009), all our results apply directly and, importantly, allow for heavy drinking to have a different marginal effect on mortality rates than mild drinking.[2]

A key assumption in the RDD literature is the continuity of both $\beta(W_i)$ and the distribution of H_i given W_i at a certain threshold point w_0 in the support of W_i. This continuity is implied by the following condition, which is convenient for our purposes.

Assumption 1 *$E[H_i|W_i = w, Z_i]$ and $\beta(w)$ are continuous in w at w_0 almost surely (a.s.).*[3]

Denote the parameter of interest by $\beta_0 = \beta(w_0)$. Assumption 1, equation (2), and the existence of the limits involved imply the equation

$$\delta_Y(Z) = \beta_0' \delta_X(Z), \tag{3}$$

where, for a generic random vector V,

$$\delta_V(Z) = \lim_{w \downarrow w_0} E[V|W = w, Z] - \lim_{w \uparrow w_0} E[V|W = w, Z].$$

Equation (3) is key in our analysis as it relates reduced form effects $\delta_X(Z)$ and $\delta_Y(Z)$ with the structural parameters of interest. To simplify the notation, we drop the dependence on Z of $\delta_V(Z)$.

The following condition guarantees a unique solution of equation (3).

Assumption 2 *$E[\delta_X \delta_X']$ is positive definite.*

Assumption 2 requires that the Xs change across the threshold in a linearly independent manner for different values of Z. Indirectly, it requires that Z assumes at least d values, as we show below. The following result establishes the conditions for identification of β_0.

Theorem 2.1 *Let equation (2) and Assumption 1 hold. Then, β_0 is identified iff Assumption 2 holds.*

Note that Assumption 2 is the minimal condition for identification in this model. Other identification conditions, such as for example the local version of Assumption 2 given by

$$E[\delta_X \delta_X'|W = w_0] \text{ is positive definite,}$$

are sufficient but not necessary for identification, and as such are stronger than Assumption 2. Theorem 2.1 is a special case of a more general identification result derived in Section A of the Appendix. There, we establish a necessary and sufficient condition for identification in a general version of the model in (1) that allows for marginal effects that depend on Z. The conclusion of the Theorem remains true if X and Z are continuous, discrete or mixed, as long as (2) holds. In the following three subsections we develop special cases which clarify the identification requirements as well as why the identification works. All these cases are particularly relevant in practice.

[2] Carpenter and Dobkin (2009) follow a standard RDD approach, imposing a unique marginal effect of drinking on alcohol-related deaths, mostly through vehicle accidents, for young adults around the age of 21.

[3] See Assumption (A1) in Hahn, Todd and van der Klaauw (1990).

2.1 Single Binary Treatment Variable (Standard RDD)

Consider the case of a single binary treatment variable (i.e. $d = 1$), with support $\mathcal{T} = \{0, 1\}$. This is the standard RDD setting, and thus both the standard RDD and our approach can achieve identification of β_0, but the relevance condition is different. In our approach, Assumption 2 requires $E[\delta_X^2] \neq 0$, which means that

$$\delta_X = \lim_{w \downarrow w_0} P(T = 1|, Z, W = w) - \lim_{w \uparrow w_0} P(T = 1|Z, W = w) \neq 0 \text{ with positive probability.}$$

The standard RDD relevance requires $\lim_{w\downarrow w_0} P(T = 1|W = w) - \lim_{w\uparrow w_0} P(T = 1|W = w) \neq 0$ which is equivalent to $\lim_{w\downarrow w_0} E[P(T = 1|Z, W = w)] - \lim_{w\uparrow w_0} E[P(T = 1|Z, W = w)] = E[\delta_X] \neq 0$ (by the law of iterated expectations and the Dominated Convergence Theorem). Thus, the standard RDD relevance is stronger than our Assumption 2.

In other words, our approach uses the variation of the differential in the probability of treatment across the threshold for different values of the covariate Z, whereas the standard RDD relies exclusively on the variation of the average differential in the probability of treatment across the threshold for the different values of Z. To better understand the implications of our relevance condition, consider the following simplified example. Suppose δ_X is a linear function of Z, $\delta_X = \alpha_1 + \alpha_2 Z$, with Z a random variable with zero mean. A standard RDD approach requires $\alpha_1 \neq 0$ for identification, while our design requires $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$. Our approach allows cases in which for a subset of values of Z the probability of treatment increases across the threshold, while for other values it decreases, even if these cancel out so that on average there is no change at all.

Moreover, let q be the cardinality of the support of Z. Then the standard RDD equation $\mathbb{E}[\delta_Y] = \beta_0 \mathbb{E}[\delta_X]$ provides just-identification, while the equation $\delta_Y = \beta_0 \delta_X$ provides in general $q - 1$ over-identification restrictions. Practically, this translates into smaller mean squared errors for estimators based on our identification strategy and mean squared errors that are less sensitive to bandwidths than standard RDD estimators. Our Monte Carlo Section 4 illustrates this point. Additionally, having over-identification restrictions enables researchers to test the constancy of the marginal effects in Z, for example, by estimating and comparing covariate specific marginal effects when sample sizes are large, or using more practical tests that we discuss in Section 2.3 when sample sizes are moderate or small.

2.2 RDD with Two Binary Treatment Variables

Consider the case of two binary treatment variables (i.e. $d = 2$), i.e. $X = (X_1, X_2)'$. It is straightforward to extend the intuition of this case to more complex cases in which there are more binary treatment variables, as well as to the case of multiple arbitrary variables in a linear model described by equation (2). In this setting the standard RDD cannot identify the separate effect of each of the variables, and thus there is no comparison to be made between the two methods.

The main identifying equation reads $\delta_Y = \beta_1 \delta_{X_1} + \beta_2 \delta_{X_2}$, where $\beta_0 = (\beta_1, \beta_2)'$. Assume for exposition ease that the support of Z is finite, given by $\{z_1, ..., z_q\}$. If $\delta_{X_2}(z_j) = 0$ for all $j = 1, ..., q$, then Assumption 2 fails and β_2 is unidentified. In a double treatment level model, Assumption 2 implies that there must be a change in the probability of the second treatment level $\lim_{w\downarrow w_0} P(X_2 = 1|, Z =$

$z_j, W = w) - \lim_{w \uparrow w_0} P(X_2 = 1 | Z = z_j, W = w)$ for at least one value of j. In the more interesting case where there is a discontinuity in X_2, i.e. $\delta_{X_2}(z_j) \neq 0$ for some $j = 1, ..., q$, Assumption 2 holds provided that for at least two values of j there is no λ such that $\delta_{X_1}(z_j) = \lambda \delta_{X_2}(z_j)$. In a double treatment level model, this translates into the requirement that the changes in the probabilities of the first and second treatment levels cannot always be proportional. This is the precise differential effect to which we referred above.

In our application example (see Section 5) the relevance condition translates into the requirement that when crossing the Medicare eligibility age requirement at 65, the increase in insurance coverage from zero to one or more insurance policies and from one to two or more insurance policies cannot be proportional for all the values of education and ethnicity. This condition holds, since indeed minorities and people with less education tend to shift from zero to one insurance policy proportionally more than Whites and people with more education, while the opposite happens from one insurance policy to two or more insurance policies. The idea is thus to explore these linearly independent shifts to disentangle the effect of having one insurance versus none, from the effect of having two or more insurances versus one insurance.

Assumption 2 is thus testable, as its empirical counterpart can be calculated directly. We also recommend visual checks for a sample version of the rank condition $\delta_{X_1}(z_j) \neq \lambda \delta_{X_2}(z_j)$, see for example Figure 7 in Section 5. It is evident from the figure that $\delta_{X_1}(z_1)/\delta_{X_2}(z_1) \neq \delta_{X_1}(z_2)/\delta_{X_2}(z_2)$, where z_1 denotes Whites with some college or more and z_2 denotes Minorities without a high school degree. This implies the lack of proportionality mentioned above, and hence the validity of Assumption 2 in this application.

More generally, if X is d-dimensional and Z is discrete with q points of support, our Assumption 2 translates into a simple rank condition as follows. Let Δ_X denote the $q \times d$ matrix with j-th row equals to $\delta_X(z_j)$, $j = 1, ..., q$, so that $\Delta_Y = \Delta_X \beta_0$. Then, by the Rank Nullity Theorem, a necessary and sufficient condition for identification of β_0 is that

$$rank\left(\Delta_X\right) = d. \tag{4}$$

The necessary order condition is $q \geq d$, and the number of over-identifying restrictions is $q - d$.

To see this, note that

$$E[\delta_X \delta_X'] = \Delta_X' L \Delta_X, \tag{5}$$

where L is a diagonal $q \times q$ matrix with j-th diagonal element $\Pr\left(Z = z_j\right)/q$. It is then clear that Assumption 2 is equivalent to (4), since L has full rank and by (5), $rank\left(E[\delta_X \delta_X']\right) = rank\left(\Delta_X\right)$.

Our empirical example below considers the case of $d = 2$ and $q = 6$. The number of over-identifying restrictions is $q - d = 4$ in this example. As we will see in the next section, having over-identifying restrictions enables researchers to test the hypothesis of constant marginal effects with respect to covariates.

2.3 Variable Marginal Effects

Assume now that marginal effects are a linear function of Z_1, where $Z = (Z_1, Z_2)$. Specifically, assume that the j-th component of β_0 is given by

$$\beta_j = \gamma_{0j} + \gamma_{1j} Z_1,$$

where we assume, for simplicity, that the dimension of Z_1 is one. Other cases can be treated analogously.

It is straightforward to show that this variable marginal effect model can be rewritten as the model in (2) by simply redefining the endogenous variable as $\tilde{X} = (X', X'Z_1)'$, which has dimension $\tilde{d} = 2d$. Thus, the previous identification results can be readily applied to this case. The relevance condition becomes

$$E[\delta_{\tilde{X}} \delta'_{\tilde{X}}] \text{ is non-singular.}$$

So, for example, if Z is discrete with q points of support, the rank condition is

$$rank\left(\Delta_{\tilde{X}}\right) = 2d,$$

with an order condition $q \geq 2d$. More generally, if $\beta_j = \gamma_{0j} + \gamma_{1j} Z_1 + \gamma_{2j} Z_2$, with both Z_1 and Z_2 univariate, the rank condition becomes

$$rank\left(\Delta_{\tilde{X}}\right) = 3d,$$

where now $\tilde{X} = (X', X'Z_1, X'Z_2)'$, and the order condition becomes $q \geq 3d$. So, for example, in our application where we have $q = 6$ and $d = 2$ the order condition is satisfied, even when both marginal effects depend on Z_1 and Z_2 as in the model above.[4] This means that if the rank condition is satisfied, the $\gamma's$ are identified and we can test hypotheses such as, for example, $H_0 : \gamma_{11} = 0$ or $H_0 : \gamma_{12} = \gamma_{22} = 0$. Next section provides an asymptotically normal estimator for the parameters $\tilde{\beta} = (\gamma_{01}, \gamma_{11}, \gamma_{21}, \gamma_{02}, \gamma_{12}, \gamma_{22})'$ in this example, or more generally, for β_0 in model (2), which allows researchers to draw inferences on multiple marginal effects in RDD.

3 Estimation

Our identification strategy builds on the equation

$$E[Y_i|Z_i, W_i = w] = \beta'(w) E[X_i|Z_i, W_i = w] + E[H_i|Z_i, W_i = w], \tag{6}$$

by exploiting the discontinuity in $E[X_i|Z_i, W_i = w]$ and the continuity of $\beta(w)$ and $E[H_i|Z_i, W_i = w]$ at the threshold w_0, which henceforth is taken as $w_0 = 0$ without loss of generality. Implementing this identification strategy requires to fit a model for $E[Y_i|Z_i, W_i = w]$ and $E[X_i|Z_i, W_i = w]$ locally around $w = 0$.

[4] In our application (Section 5), we opt for not assuming that the variable representing the education level, which takes three different values (no high school degree, high school degree and at least some college), enters linearly as in the example above. This allows us to test for the presence of a less restrictive heterogeneity in β_j, but only one dimension (race or education) at a time.

Applied researchers of the RDD methodology often fit locally a reduced form model of the form

$$Y_i = \alpha_y + \pi_y D_i + g^y(W_i) + \gamma_y' Z_i + v_i^y,$$

where $D_i = 1(W_i \geq 0)$, $g^y(w)$ is a smooth function of w, and v_i^y is an error term. Then, they proceed by specifying a similar first stage for the univariate treatment variable X_i,

$$X_i = \alpha_x + \pi_x D_i + g^x(W_i) + \gamma_x' Z_i + v_i^x, \tag{7}$$

and estimating these two equations by Ordinary Least Squares (OLS). Equivalently, they estimate the parameter $\beta_y = \pi_y/\pi_x$ by Two Stage Least Squares (TSLS) using D_i as an instrument for X_i, and treating the smooth function of W_i and Z_i as the non-excluded exogenous variables. This method does not exploit variability of the first stages in covariates, provides just-identification only when π_x is not zero, and does not allow for multivariate treatment variables. We follow our identification strategy above to propose an estimator that overcomes these limitations by exploiting heterogeneity in the first stages in covariates, while allowing for multiple treatments.

We will exploit the variability of the first stages in covariates, but a fully nonparametric estimator of (6) would be affected by the so-called "curse of dimensionality" when Z_i is high dimensional and/or sample sizes are moderate or small. To address the "curse of dimensionality" problem of fully nonparametric methods, while allowing for heterogeneity in covariates, this article proposes a semiparametric varying coefficient specification of the first stage

$$X_i = \alpha_{0X}(W_i) + \alpha_{1X}(W_i) Z_i + \varepsilon_i^x, \tag{8}$$

where $\alpha_{0X}(\cdot)$ and $\alpha_{1X}(\cdot)$ are unknown functions of W and ε_i^x is a prediction error term. This semiparametric specification avoids smoothing in the possibly high dimensional Z, so only smoothing in the one dimensional running variable W is involved.

We could estimate $\alpha_{0X}(w)$ and $\alpha_{1X}(w)$ locally at $w = 0$ by a constant kernel method. However, it is well known that the local constant kernel estimator has generally worse bias properties than the local linear kernel estimator at discontinuity points; see Fan and Gijbels (1996). For this reason, we suggest implementing our identification strategy for the RDD with a local linear estimator. This corresponds to the use of linear (as opposed to just constant) approximations for $\alpha_{0X}(\cdot)$ and $\alpha_{1X}(\cdot)$ around each side of the threshold $w_0 = 0$. Intuitively, this means we approximate the first stage by

$$\begin{aligned} X_i &\approx \alpha_{0X}^+ D_i + \alpha_{1X}^+ Z_i D_i + \dot{\alpha}_{0X}^+ W_i D_i + \dot{\alpha}_{1X}^+ Z_i W_i D_i \qquad (9) \\ &\quad + \alpha_{0X}^- (1 - D_i) + \alpha_{1X}^- Z_i (1 - D_i) + \dot{\alpha}_{0X}^- W_i (1 - D_i) + \dot{\alpha}_{1X}^- Z_i W_i (1 - D_i) + \varepsilon_i^x \\ &= \alpha_{0X}^- + \pi_{0X} D_i + \pi_{1X} Z_i D_i + g^x(W_i, Z_i) + \alpha_{1X}^- Z_i + \varepsilon_i^x, \qquad (10) \end{aligned}$$

where $\pi_{0X} = \alpha_{0X}^+ - \alpha_{0X}^-$, $\pi_{1X} = \alpha_{1X}^+ - \alpha_{1X}^-$ and $g^x(W_i, Z_i) = \dot{\alpha}_{0X}^+ W_i D_i + \dot{\alpha}_{1X}^+ Z_i W_i D_i + \dot{\alpha}_{0X}^- W_i (1 - D_i) + \dot{\alpha}_{1X}^- Z_i W_i (1 - D_i)$. Comparing (7) and (10) we see that exploiting heterogeneity in the first stages permits the use of additional instruments, namely $Z_i D_i$, for estimating the marginal effects.

Define $S_{+i} = D_i \cdot (1, W_i, Z_i, Z_i W_i)'$, $S_{-i} = (1 - D_i) \cdot (1, W_i, Z_i, Z_i W_i)'$, and $S_i = (S'_{+i}, S'_{-i})'$. We formalize our approach as a TSLS which solves the first stage problem in (10)

$$\widehat{\alpha}_X = arg \min_a \sum_{i=1}^{n} \left| X_i - S_i' a \right|^2 k_{h_n}(W_i),$$

where $k_{h_n}(W) = k(W/h_n)$, k is a kernel function and h_n is a bandwidth parameter satisfying some standard conditions in Section B in the Appendix. Then, in the second stage, we run a local linear regression Y_i on $\hat{X}_i = S_i' \widehat{\alpha}_X$ and $C_i = (1, Z_i, W_i, D_i \cdot W_i, Z_i \cdot W_i, Z_i \cdot D_i \cdot W_i)'$, i.e.

$$\begin{bmatrix} \hat{\beta} \\ \hat{\eta} \end{bmatrix} = arg \min_{\beta, \eta} \sum_{i=1}^{n} \left(Y_i - \beta' \hat{X}_i - \eta' C_i \right)^2 k_{h_n}(W_i).$$

Note that C_i contains the set of non-excluded "exogenous" variables in our approach, cf. (10).

The following result establishes the asymptotic normality of the proposed TSLS. Its proof and its regularity conditions are given in the Appendix Section B.

Theorem 3.1 *Let Assumption 1 and Assumption 2 hold. Let also Assumption 4 in Section B in the Appendix hold. Then*

$$\sqrt{n h_n}(\hat{\beta} - \beta_0) \to_d N(0, \Sigma),$$

where Σ is given in Section B in the Appendix.

3.1 Implementation

For implementing our estimator we recommend using the uniform kernel

$$k_{h_n i} = 1(-h_n \leq W_i \leq h_n). \tag{11}$$

This means the estimator can be easily implemented by restricting the sample $\{(Y_i, X_i, Z_i, W_i)\}_{i=1}^n$ to only those observations such that $-h_n \leq W_i \leq h_n$, and running a TSLS regression of Y_i onto X_i and C_i, treating C_i as exogenous, and using $(D_i, Z_i \cdot D_i)'$ as a vector of "instruments" for X_i. When X is d-dimensional and Z is p-dimensional, the necessary order condition is $p \geq d$. We recommend running the first stage and checking for heterogeneity in covariates by means of visual checks and formal tests of significance. Similarly, we also recommend checking for variable marginal effects as a robustness check.

The asymptotic variance of $\hat{\beta}$ can be consistently estimated by the standard TSLS asymptotic variance. Implementation of $\hat{\beta}$ requires the choice of a bandwidth parameter h_n. We can choose h_n along the lines suggested in Calonico, Cattaneo and Titiunik (2014), which can be adapted to the presence of additional covariates (included in the model as in the standard RDD setting) as shown in Calonico, Cattaneo, Farrell and Titiunik (2016). Nevertheless, we have found in all our numerical examples as well as the application that our estimator is not sensitive to the bandwidth parameter, an in particular, it is far less sensitive than standard RDD estimators. See Figure 1 for empirical evidence supporting this claim. In our numerical examples we have found that the simple rule $h_n = 2n^{-1/4}$

performs well. Nevertheless, we recommend sensitivity analysis by varying the bandwidth, see our application for illustration.

Note that the way in which we introduce covariates in our RDD approach is different from how is traditionally done in RDD; see, e.g., Imbens and Lemieux (2008). The traditional RDD with covariates is a TSLS with exogenous variables $C_i^{RDD} = (1, Z_i, W_i, D_i \cdot W_i)'$ and a single excluded instrument D_i for X_i. Our semiparametric varying coefficient specification accounts for the additional exogenous variables $Z_i \cdot W_i$ and $Z_i \cdot D_i \cdot W_i$, while the interaction term $Z_i \cdot D_i$ is used as an additional instrument to help identify multiple effects in our setting.

4 Monte Carlo Simulations

This section studies the finite sample performance of the proposed estimator, in comparison with standard RDD estimators when possible. The first Data Generating Process (DGP) is used to illustrate the finite sample performance of the proposed RDD estimator in the context of the single treatment variable. The data is generated according to:

$$\begin{aligned} Y &= \alpha + \beta X + \gamma Z + u, \\ D &= 1(W \geq 0), \\ X &= \alpha_0 + \alpha_1 D + \alpha_2 Z + \alpha_3 D \cdot Z + \varepsilon_X, \end{aligned}$$

where (u, ε_X, W, Z) are independent and identically distributed as standard normals. Note that with this design

$$\lim_{w \downarrow 0} E[X|W = w, Z] = \alpha_0 + \alpha_1 + (\alpha_2 + \alpha_3) Z$$

and

$$\lim_{w \uparrow 0} E[X|W = w, Z] = \alpha_0 + \alpha_2 Z.$$

Therefore, $\delta_X = \alpha_1 + \alpha_3 Z$. On the other hand, since $E[Z|W] = 0$,

$$\lim_{w \downarrow 0} E[X|W = w] = \alpha_0 + \alpha_1$$

and

$$\lim_{w \uparrow 0} E[X|W = w] = \alpha_0.$$

If we define, for a generic V,

$$\delta_V^{RDD} = \lim_{w \downarrow \bar{w}} E[V|W = w] - \lim_{w \uparrow \bar{w}} E[V|W = w],$$

then $\delta_X^{RDD} = \alpha_1$ controls the level of identification of the standard RDD, whereas α_1 and α_3 simultaneously control the level of identification for the over-identified RDD estimator. We compare the standard RDD estimator ($\widehat{\beta}_{RDD}$), the standard RDD estimator with covariates ($\widehat{\beta}_{RDDcov}$), and the over-identified RDD estimator ($\hat{\beta}$). We consider the parameter values $(\alpha, \beta, \gamma) = (0, 1, 0)$, $(\alpha_0, \alpha_2, \alpha_3) = (0, 1, 1)$ and

Table 1: **Average Bias and MSE: RDD**

α_1	n	*Bias*			*MSE*		
		$\widehat{\beta}_{RDD}$	$\widehat{\beta}_{RDDcov}$	$\widehat{\beta}$	$\widehat{\beta}_{RDD}$	$\widehat{\beta}_{RDDcov}$	$\widehat{\beta}$
0	100	0.233	2.173	-0.014	1,546.280	24,082.300	3.167
	300	-1.554	-0.149	0.002	113,467.000	4,665.060	0.387
	500	7.726	-1.378	0.002	550,357.000	2,730.810	0.158
	1000	-1.020	0.332	-0.002	6,083.13	1,179.240	0.073
1	100	0.030	-0.110	0.005	159.572	291.977	0.699
	300	0.560	0.026	-0.003	590.620	29.011	0.103
	500	-0.027	0.058	0.000	43.261	14.823	0.061
	1000	-0.166	0.007	0.002	170.686	0.095	0.032
2	100	-0.005	-0.135	0.006	190.910	189.136	0.295
	300	-0.059	-0.001	-0.002	19.171	0.046	0.034
	500	-0.001	-0.001	-0.000	0.184	0.030	0.022
	1000	0.003	0.003	0.002	0.017	0.015	0.012

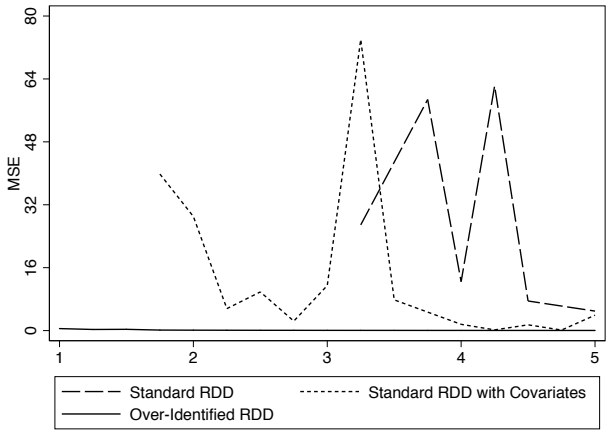
Note: 10,000 Monte Carlo Simulations. Numbers are rounded to the nearest milesimal. When there is a "-" sign in front of a 0.000 it means that the number is between -0.0005 and 0.

several values of α_1. We implement all estimators as in Section 3 with a uniform kernel (11) and a bandwidth $h_n = 2n^{-1/4}$.

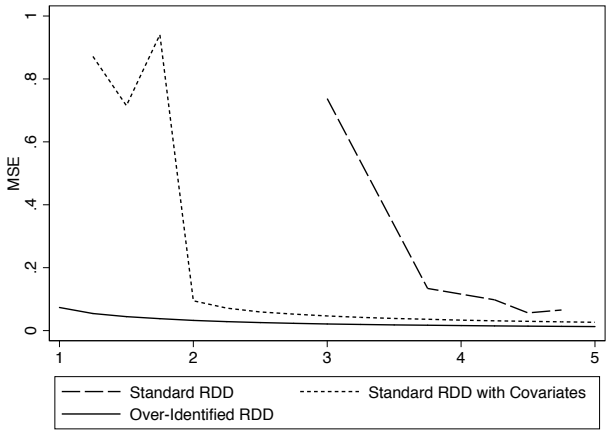
Table 1 reports the average bias and MSE based on 10,000 Monte Carlo simulations. For $\alpha_1 = 0$, the standard RDD is not able to identify the parameter β, and, as expected, has a MSE that remains large for large n. Adding covariates in the standard RDD improves slightly the bias and the MSE for moderate and large samples, but yet does not identify the parameter and leads to large MSE when $\alpha_1 = 0$. In contrast, our estimator $\hat{\beta}$ presents a satisfactory performance, with small MSEs that decrease with the sample size. For $\alpha_1 = 1$ the MSEs for standard RDD, with or without covariates, remain relative large even for large sample sizes as $n = 1000$. This seems to be a case of weak identification by standard RDD, see Feir, Lemieux and Marmer (2016). Indeed, it requires a value of $\alpha_1 = 2$ and a sample size of $n \geq 500$ for the standard RDD to achieve comparable results to those of our estimator.

Next, we study the sensitivity of the proposed estimator to the bandwidth parameter. Figure 1 reports the MSE for this DGP in the "weakly" and "strongly" identified cases (i.e., when $\alpha_1 = 1$ and $\alpha_1 = 2$) for sample sizes 300 and 1,000, and where the bandwidth is chosen as $h_n = cn^{-1/4}$ for a constant $c = 1, 1.25, 1.5, ..., 5$. Note that the vertical scale of the plots is different, since the disparities

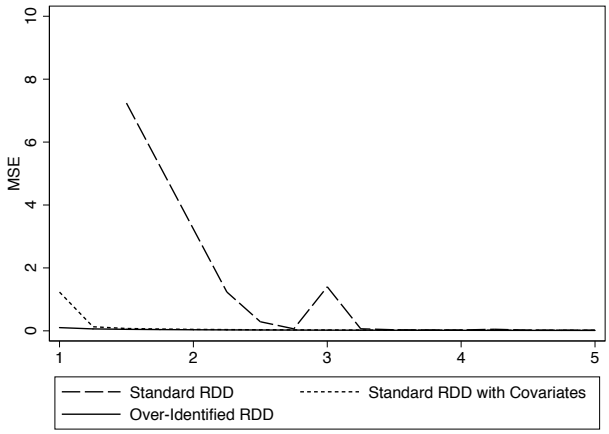
Figure 1: Sensitivity to the Choice of Bandwidth



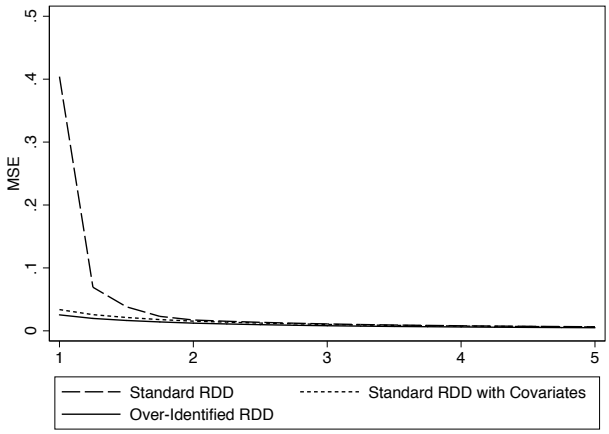
(a) 300 obs., $\alpha_1 = 1$



(b) 1000 obs., $\alpha_1 = 1$



(c) 300 obs., $\alpha_1 = 2$



(d) 1000 obs., $\alpha_1 = 2$

Note: Horizontal axis shows the values of $c = 1, 1.25, 1.5, 1.75, \ldots$, reflecting different bandwidths: $h_n = cn^{-1/4}$. Vertical axis shows the MSE, with different scales depending on the figure. When a curve is only shown for high values of c it is because the values are off the scale for the lower values of c. For example, in Figure 1a the MSE of the standard RDD for $c < 3.25$ is over 80.

between the standard RDD and our method can sometimes be extreme.

The results show that RDD without covariates is particularly sensitive to the choice of the bandwidth. Including covariates helps significantly, but still leads to variable results, particularly for low values of the bandwidth. Our estimator has the smallest MSEs, uniformly over all values of the bandwidth (never above 0.02), and a MSE that is flatter as a function of the bandwidth. These conclusions hold even in the strongly identified case where $\alpha_1 = 2$. This DGP illustrates the potential benefits of our approach even in situations where the standard RDD is applicable.

The second DGP explores all manners of heterogeneity, both in the first stage as well as in the structural equation. The model is

$$
\begin{aligned}
D &= 1(W \geq \bar{w}), \\
X &= \alpha_C + \alpha_D D + \alpha_Z Z + \alpha_W W + \alpha_{ZD} ZD + \alpha_{WD} WD + \alpha_{WZ} WZ + \alpha_{WZD} WZD + \alpha_{VD} VD + U_X, \\
Y &= \beta_C + \beta_X X + \beta_Z Z + \beta_W W + \beta_{ZX} ZX + \beta_{WX} WX + \beta_{WZ} WZ + \beta_{WZX} WZX + \beta_{VX} VX + U_Y
\end{aligned}
$$

In this model, heterogeneity on the observables W and Z in the first stage is determined by α_{ZD}, α_{WD} and α_{WZD}, while heterogeneity on unobservables is given by α_{VD}. Analogously heterogeneity on the observables W and Z in the structural equation is determined by β_{ZX}, β_{WX} and β_{WZX}, while heterogeneity on unobservables is given by β_{VX}. We assume that U_X and U_Y are unobservables which are independent of each other and of W and Z.

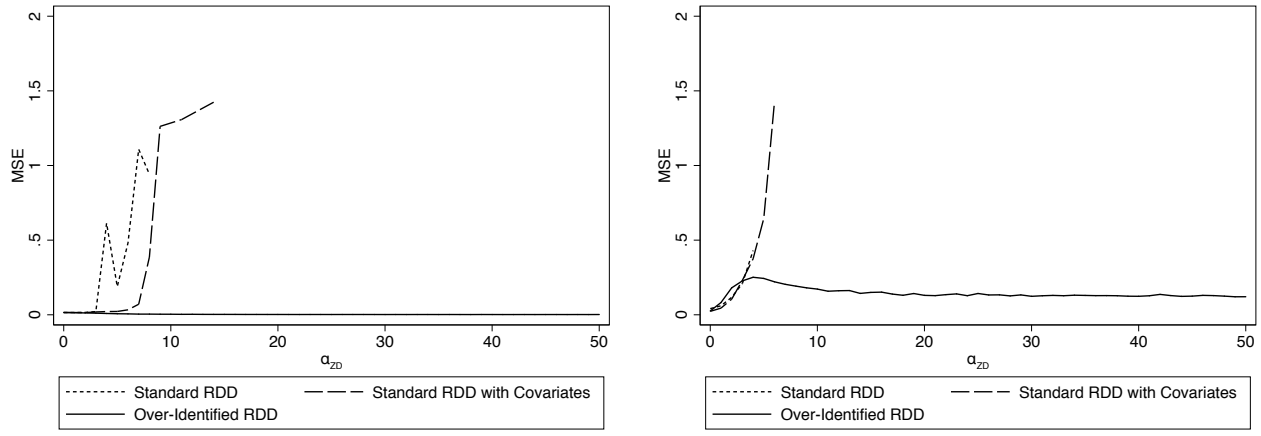
The parameter values we use are: $\bar{w} = 0$, $\alpha_C = -1$, $\alpha_D = 3$, $\alpha_Z = 0$, $\alpha_W = 2$, $\alpha_{WD} = 1$, $\alpha_{WZ} = 0$, $\alpha_{WZD} = 0$, $\alpha_{VD} = 0.1$, $\beta_C = 1$, $\beta_X = 5$, $\beta_Z = 0$, $\beta_W = 0.5$, $\beta_{WX} = 1$, $\beta_{WZ} = 0$, $\beta_{WZX} = 0$, $\beta_{VX} = 0.1$. Additionally (Z, W, V) are jointly normal with mean $(0, 0, 0)$, variances equal to 1 and covariances equal to 0.5. (U_X, U_Y) are independent of other variables and each other, normally distributed with zero mean and variance equal to 1. We vary α_{ZD} and β_{ZX} in our simulations.

The parameters were chosen so as to yield the kinds of plots we normally encounter in applied research. In this model, the strength of identification in the RDD is given by $\alpha_D + \alpha_{ZD}\mathbb{E}[Z|W = \bar{w}] + \alpha_{WD} + \alpha_{WD} = 4.1 + \alpha_{ZD}\mathbb{E}[Z|W = \bar{w}]$. Because we want to understand the relative effects of the identification of the RDD and the heterogeneity in Z, we assume that $\mathbb{E}[Z|W = \bar{w}] \neq 0$. Then the strength of identification in the standard RDD is 4.1, and thus in theory the RDD is identified. The heterogeneity on Z is entirely determined by α_{ZD}.

Figure 2 shows the MSE of the over-identified RDD, the standard RDD and the RDD with covariates in the estimation of β_X when $\beta_{ZX} = 0$ and when $\beta_{ZX} = 0.5$. Note that even when $\beta_{ZX} = 0$ this model has heterogeneous treatment effects that are due to β_{WX} and β_{VX}. The plots show the MSE of the standard RDD and the RDD with covariates coming off the scale for higher levels of heterogeneity in the first stage, α_{ZD}.

Table 2 shows the numerical results. The over-identified RDD seems to converge to a small number as α_{ZD} increases. This is due to the fact that the bandwidth is kept fixed. In contrast, the standard RDD and the RDD with covariates have MSEs of enormous variability.

Figure 2: **MSE in the estimation of the LATE, 1000 obs**



Note: The first plot shows the relative MSEs when $\beta_{ZX} = 0$. The second plot shows the relative MSE when $\beta_{ZX} = 0.5$.

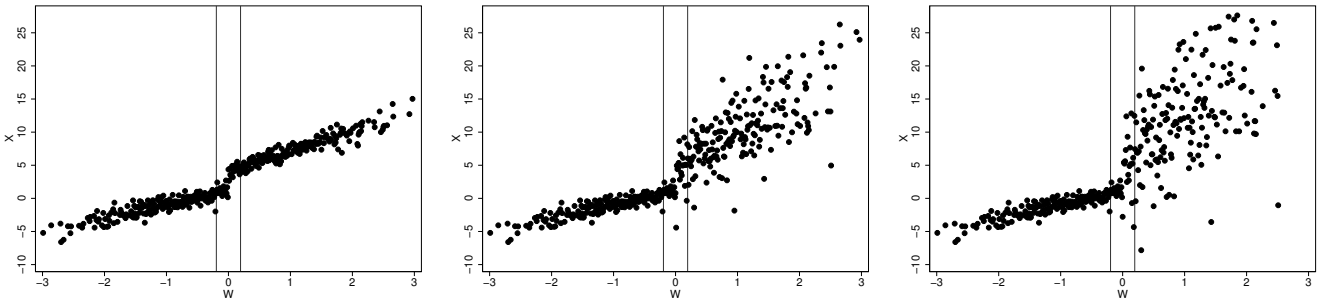
Table 2: **MSE in the estimation of the LATE, 1000 observations**

	$\beta_{ZX} = 0$			$\beta_{ZX} = 0.5$		
α_{ZD}	RDD	RDD w.Cov	Over-Id RDD	RDD	RDD w.Cov	Over-Id RDD
0	0.0151	0.0152	0.0154	0.0412	0.0230	0.0256
1	0.0152	0.0151	0.0141	0.0603	0.0447	0.0820
2	0.0159	0.0147	0.0123	0.1171	0.1047	0.1805
3	0.0213	0.0189	0.0108	0.2096	0.2311	0.2273
4	0.6134	0.0215	0.0087	0.4307	3750	0.2513
5	0.1885	0.0222	0.0070	82.2939	0.6460	0.2435
10	23.9257	27.6922	0.0036	134.5545	4491.3940	0.1721
15	3.0399	138.0055	0.0026	2976.9940	2237.2070	0.1499
20	206.3206	6.3821	0.0021	1989.4200	5768.6810	0.1301
25	2.5779	27.9945	0.0019	203675.7000	3071.0570	0.1427
30	10.9548	1384.8070	0.0017	5248.5900	8053.9880	0.1237
35	4.4075	227.5336	0.0017	1075.6240	702.2343	0.1297
40	5339.5830	9.4090	0.0016	51329.7200	21621.6200	0.1234
45	5.6744	163.7262	0.0017	984.0347	5492.0400	0.1249
50	4.7163	622.1618	0.0018	23167.4400	8036.0820	0.1209

We believe that the bad performance of the standard RDD and the RDD with covariates when the heterogeneity in covariates is high is due to weak instruments. These estimators perceive heterogeneity in covariates as noise. To show this, we plotted the data in one of the simulations, which have 1000 observations in Figure 3. The first plot shows the simulation data when $\alpha_{ZD} = 0$, and thus there is no heterogeneity of the first stage on covariates. There is a clear discontinuity in this case, and thus

all three methods work well. The second plot shows the data when $\alpha_{ZD} = 5$, and the third when $\alpha_{ZD} = 10$. As the heterogeneity increases, the right side of the discontinuity seems noisier. Although the discontinuity is of the exact same size, it is hard for a method that relies exclusively on the average size of the jump to distinguish noise from signal.

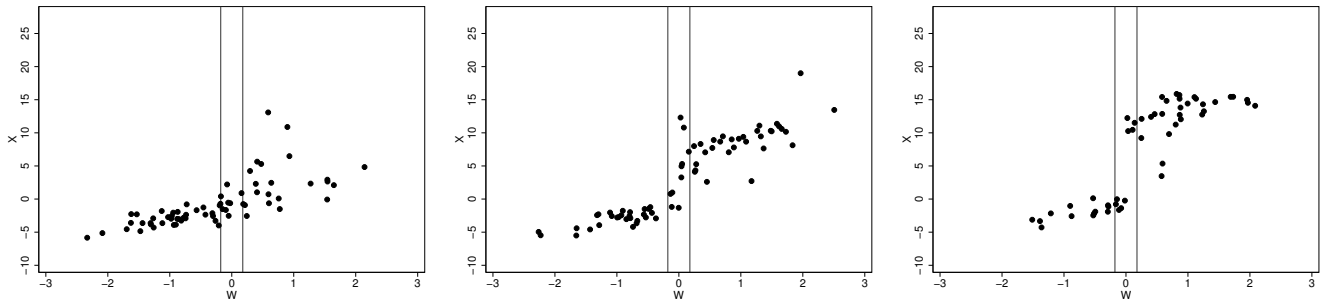
Figure 3: **First-stages for different levels of heterogeneity on Z**



Note: The first plot shows the simulation data when $\alpha_{ZD} = 0$, and thus there is no heterogeneity of the first stage on covariates. The second plot shows the data when $\alpha_{ZD} = 5$, and the third when $\alpha_{ZD} = 10$.

Figure 4 takes the data when $\alpha_{ZD} = 10$ and divides it according to the values of Z. The first plot shows only observations such that $-0.5 < Z < -0.25$, the second plot shows only observations such that $0.25 < Z < 0.5$ and the third plot shows only observations such that $0.75 < Z < 1$. Although there is almost no distinguishable discontinuity for low values of Z, shown in the first plot, the higher values of Z show clear discontinuities. Any subset of the data such as in the second and third plots would be sufficient to identify β_X, but the over-identified RDD uses all the available discontinuities at the same time, thus achieving not only identification, but lower variability as well.

Figure 4: **First-stages conditional on Z**



Note: All three plots shows the simulation data when $\alpha_{ZD} = 10$. The first plot shows only observations such that $-0.5 < Z < -0.25$, the second plot shows only observations such that $0.25 < Z < 0.5$ and the third plot shows only observations such that $0.75 < Z < 1$.

We also consider a third DGP with two treatments:

$$\begin{aligned} Y &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma Z + u, \\ D &= 1(W \geq 0), \\ X_1 &= \alpha_{01} + \alpha_{11} D + \alpha_{21} Z + \alpha_{31} D \times Z + \varepsilon_1, \\ X_2 &= \alpha_{02} + \alpha_{12} D + \alpha_{22} Z + \alpha_{32} D \times Z + \varepsilon_2, \end{aligned}$$

where $(u, W, \varepsilon_1, \varepsilon_2)$ are independent normal random variables with variances 0.25, 0.25, 1 and 1, respectively. In this model $\delta_Y = \beta_1 \delta_{X_1} + \beta_2 \delta_{X_2}$ with

$$\delta_{X_1} = \alpha_{11} + \alpha_{31} Z \text{ and } \delta_{X_2} = \alpha_{12} + \alpha_{32} Z.$$

Thus, our relevance condition is satisfied as long as

$$\det \begin{vmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{12} & \alpha_{32} \end{vmatrix} \neq 0. \tag{12}$$

The parameters in the structural equation are set at $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 2$ and $\gamma = 0$. We also set $\alpha_{01} = \alpha_{02} = \alpha_{21} = \alpha_{22} = 0$. The results reported below are not sensitive to these parameter values. More critical are the values for $(\alpha_{11}, \alpha_{31}, \alpha_{12}, \alpha_{32})$, which determine the rank condition. We consider three different cases: (i) Case 1: $\alpha_{11} = 2, \alpha_{31} = 1, \alpha_{12} = 0, \alpha_{32} = 1$; (ii) Case 2: $\alpha_{11} = 0, \alpha_{31} = -2, \alpha_{12} = 2, \alpha_{32} = 1$; and (iii) Case 3: $\alpha_{11} = -2, \alpha_{31} = 2, \alpha_{12} = 1, \alpha_{32} = 1$. All the three cases satisfy the relevance condition (12). In contrast, the identifying condition for standard RDD varies according to the case considered, since

$$\begin{aligned} \delta_Y^{RDD} &= \beta_1 \delta_{X_1}^{RDD} + \beta_2 \delta_{X_2}^{RDD} \\ &= \beta_1 \alpha_{11} + \beta_2 \alpha_{12}. \end{aligned}$$

Thus, in Case 1, $\alpha_{12} = 0$ and RDD identifies β_1, although it fails to identify β_2; in Case 2, $\alpha_{11} = 0$ and RDD identifies β_2, although it fails to identify β_1; and in the mixed case of Case 3 neither β_1 nor β_2 are identified by standard RDD methods. We implement the standard RDD estimator and the standard RDD estimator with covariates for each treatment separately, and compare these estimates with the over-identified RDD estimator. Table 3 provides the average bias and MSE for β_1 and β_2 based on 10,000 Monte Carlo simulations, and sample sizes $n = 300$, 500, 1,000 and 5,000. We implement all estimators as in Section 3 with a uniform kernel (equation (11)) and bandwidth $h_n = 2n^{-1/4}$.

Table 3: **Average Bias and MSE**

		$Bias\ \beta_1$			$MSE\ \beta_1$			$Bias\ \beta_2$			$MSE\ \beta_2$		
$Case$	n	$\widehat{\beta}_{RDD}$	$\widehat{\beta}_{RDDcov}$	$\widehat{\beta}$	$\widehat{\beta}_{RDD}$	$\widehat{\beta}_{RDDcov}$	$\widehat{\beta}$	$\widehat{\beta}_{RDD}$	$\widehat{\beta}_{RDDcov}$	$\widehat{\beta}$	$\widehat{\beta}_{RDD}$	$\widehat{\beta}_{RDDcov}$	$\widehat{\beta}$
1	300	0.0079	0.0027	0.0018	0.0364	0.0305	0.0268	10.9470	-8.2400	-0.0035	∞	∞	0.4019
	500	0.0043	0.0007	-0.0004	0.0236	0.0197	0.0043	-5.3491	-17.8270	0.0002	∞	∞	0.0263
	1000	0.0034	0.0016	0.0001	0.0136	0.0116	0.0020	-35.3190	-40.2920	0.0006	∞	∞	0.0108
	5000	0.0016	0.0009	-0.0001	0.0038	0.0032	0.0006	6.9921	14.7470	-0.0001	∞	∞	0.0028
2	300	54.463	6.4319	0.0004	∞	∞	0.0071	0.0178	0.0076	0.0006	0.0708	0.0485	0.0054
	500	71.3800	-9.4857	-0.0002	∞	∞	0.0045	0.0096	0.0027	-0.0004	0.0441	0.0302	0.0035
	1000	-1.6590	19.1130	-0.0004	∞	∞	0.0025	0.0078	0.0042	-0.0001	0.0247	0.0170	0.0019
	5000	-74.3120	7.4588	0.0000	∞	∞	0.0007	0.0029	0.0015	-0.0000	0.0072	0.0050	0.0006
3	300	0.5730	0.5366	-0.0003	1.6836	0.3550	0.0027	-2.4689	-2.1373	0.0009	20.5520	180.2900	0.0118
	500	0.5393	0.5232	0.0003	0.3507	0.3095	0.0017	-2.2625	-2.1766	-0.0003	7.0904	5.6605	0.0073
	1000	0.5201	0.5126	0.0003	0.3001	0.2834	0.0010	-2.1365	-2.1005	0.0005	5.1558	4.7913	0.0040
	5000	0.5046	0.5028	-0.0000	0.2622	0.2582	0.0003	-2.0379	-2.0277	-0.0001	4.2752	4.1951	0.0011

Note: 10,000 Monte Carlo Simulations. Columns 3-5 show the bias of the estimation of β_1 using the standard RDD, the standard RDD with covariates, and our method respectively. The remaining columns are analogous.

Table 3 shows that the standard RDD has a satisfactory performance in estimating β_1 in Case 1, where it identifies, but also illustrates its inconsistency in estimating β_1 in Case 2 in and Case 3. The behavior of the standard RDD in identifying β_2 is analogous, only it performs satisfactorily in Case 2, where it identifies, but it is inconsistent in both Case 1 and Case 3. The standard RDD is particularly unreliable for the extreme Cases 1 and 2 when it does not identify (β_2 in Case 1 and β_1 in Case 2). For the mixed case, standard RDD is more stable in estimating both coefficients, but its bias and MSE do not converge to zero as the sample size increases. All of this is true even when covariates are included. In contrast, our approach ($\hat{\beta}_1$ and $\hat{\beta}_2$) performs uniformly well across the three cases considered, and outperforms the standard RDD even in the case where RDD identifies the parameter.

Overall, these Monte Carlo results provide supporting evidence of the robustness of our identification strategy in the case of a single treatment and its ability to identify multiple treatment effects in situations where the standard RDD fails.

5 An Application to the Estimation of the Effect of Insurance Coverage on Health Care Utilization

We apply our approach to the problem of estimating the effects of insurance coverage on health care utilization with a regression discontinuity design, as in Card, Dobkin and Maestas (2008). They exploit the fact that Medicare eligibility varies discontinuously at age 65. Medicare eligibility may affect health care utilization via two channels. First, it provides coverage to people who were previously uninsured. Second, it provides more generous coverage to people who were previously insured by other, less generous insurance policies. Let Y be a measure of health care use (e.g., whether the person did not get care for cost reasons last year). The two main explanatory variables of interest are whether the person has any insurance coverage (i.e., one or more policies) (X_1), and whether the person has insurance coverage by two or more policies (X_2). The running variable W is defined to be the age (measured in quarters of a year) relative to the threshold of 65 years of age, and the treatment status D is whether the person is eligible to Medicare. We want to identify β_1 and β_2 in the following model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma' Z + \varepsilon \tag{13}$$

were Z represents the variables which we will explore to generate variation in the first stage (e.g., race, education level).

Table 4 presents the summary statistics for the key variables of our sample, obtained from the National Health Interview Survey (NHIS) from 1999 to 2003.[5] We consider three different outcome variables (Y): whether the person delayed care last year for cost reasons, whether the person did not get care last year for cost reasons, and whether the person went to the hospital last year.[6]

[5] Card, Dobkin and Maestas (2008) focus their main analysis on years 1992-2003, however the dual insurance coverage variable is observed only from 1999 onwards. For this reason, our analysis is restricted to years 1999 to 2003.

[6] There is another variable used by Card, Dobkin and Maestas (2008): whether the person saw a Doctor last year. This variable has many missing data throughout the sample period, so we opted to drop it from our analysis. However, our conclusions do not change when this variable is included in our study.

Table 4: **Summary Statistics**

Variable	All	Non-Hispanic White			Minority		
		HS Dropout	HS Graduate	>HS Graduate	HS Dropout	HS Graduate	>HS Graduate
Delayed Care (Y)	0.07	0.09	0.06	0.05	0.10	0.07	0.07
	(0.25)	(0.29)	(0.24)	(0.23)	(0.30)	(0.25)	(0.25)
Did Not Get Care (Y)	0.05	0.07	0.04	0.03	0.09	0.06	0.05
	(0.22)	(0.25)	(0.20)	(0.18)	(0.29)	(0.23)	(0.22)
Hospital Stay (Y)	0.13	0.17	0.12	0.12	0.14	0.12	0.12
	(0.34)	(0.38)	(0.33)	(0.32)	(0.35)	(0.33)	(0.32)
1+ Coverage (X_1)	0.92	0.91	0.95	0.96	0.78	0.87	0.91
	(0.27)	(0.28)	(0.22)	(0.19)	(0.41)	(0.33)	(0.29)
2+ Coverage (X_2)	0.33	0.44	0.38	0.32	0.24	0.23	0.24
	(0.47)	(0.50)	(0.49)	(0.47)	(0.43)	(0.42)	(0.42)
Medicare Eligible (D)	0.42	0.55	0.45	0.37	0.46	0.36	0.34
	(0.49)	(0.50)	(0.50)	(0.48)	(0.50)	(0.48)	(0.47)
Observations	63,165	8,337	16,037	21,352	8,293	4,302	4,844

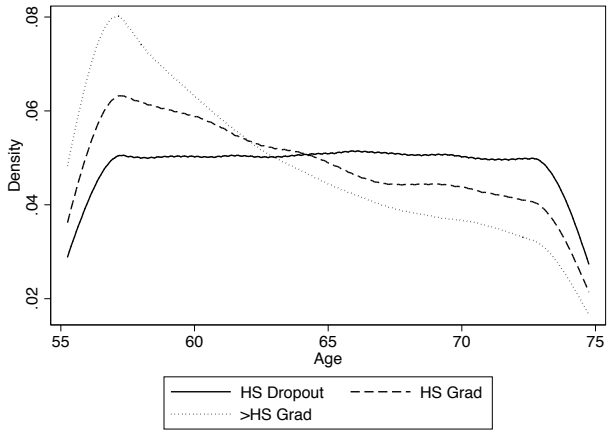
Note: Source: NHIS 1999-2003. Standard deviations in parentheses. "HS Dropout" represents people who have less than a high school degree. "HS Graduate" represents those who have a high school degree. "> HS Graduate" represents those who have some college or more.

The table shows that non-Hispanic Whites are less likely to delay or ration care because of cost relative to minorities. The results are analogous when comparing people with higher vs. lower education levels. However, people with lower education levels go more often to Hospitals than people with higher education levels, and non-Hispanic Whites go as often as minorities, except for less educated ones.

The table also shows that non-Hispanic Whites tend to be insured with a higher likelihood than minorities, and at the same time are more likely to carry a second insurance policy. Additionally, people with more education are more likely to have some insurance than people with less education, irrespective of the race. However, less educated Whites are more likely to carry two or more insurance policies than their more educated counterparts. This counter-intuitive correlation is better understood in the context of age as an important confounder. As seen in the second to last row of the table, people with more education are less likely to be eligible to Medicare. Indeed, Figure 5 shows that people with less education tend to be older than people with more education, so they are more likely to be eligible to Medicare. Following Card, Dobkin and Maestas (2008), we use a regression discontinuity design to circumvent this and other endogeneity concerns.[7]

[7]Of course, age is simply one potential confounder. There are many unobserved determinants of X_1 and X_2 that should also affect Y directly, as for example the person's income, work status and health status (see Section II.A in Card,

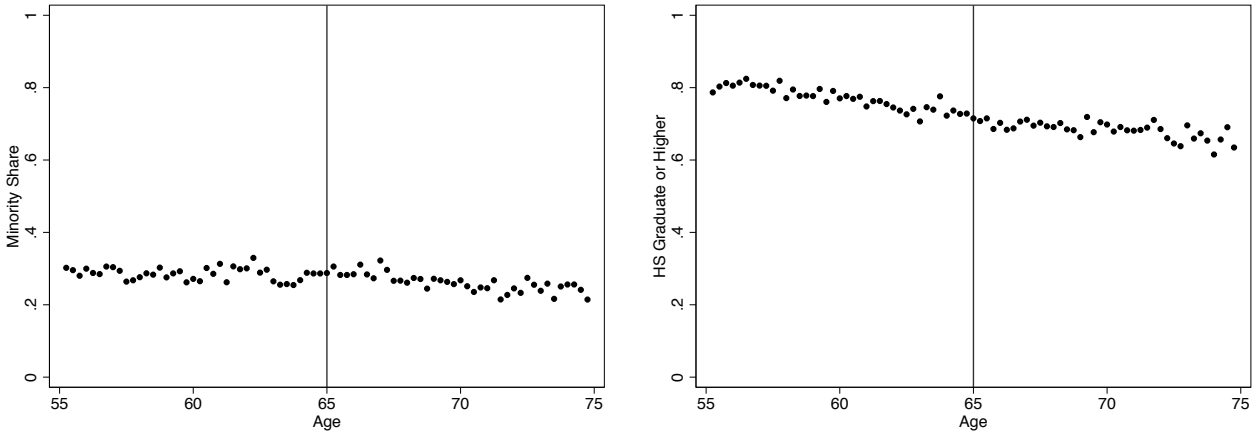
Figure 5: Age Distribution by Level of Education



Note: For each education level, this Figure shows the Kernel density plot of the age distribution (measured in quarters of a year). Kernel: Epanechnikov. Bandwidth: 1.

To see how an RDD is reasonable in this context, Figure 6 shows plots suggestive that people just younger and just older than 65 years of age are comparable in terms of race and education. Similarly to Card, Dobkin and Maestas (2008), we find no evidence of discontinuity at 65 years of age for a wide range of covariates, suggesting that variables included in Z are likely exogenous conditional on being close enough to the threshold of 65 years of age.

Figure 6: **Validity Plots**



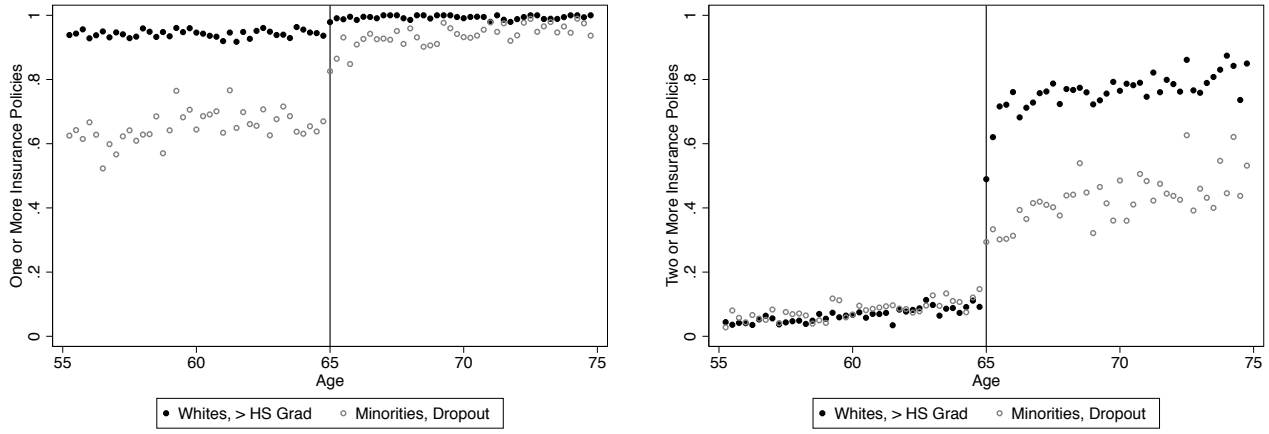
Note: The scatter plot shows the average of the variable described in the vertical axis for each age level (measured in quarters of a year).

Figure 7 shows the first stage discontinuity plots for different subgroups of people based on race

Dobkin and Maestas (2008) for a more detailed discussion of potential confounders in this context.) An RDD approach aims at avoiding all such confounders.

and education. While the discontinuity in X_1 (Panel (a)) is larger for Minorities with low levels of education, the discontinuity in X_2 (Panel (b)) is larger for Whites with high levels of education. These results imply that eligibility to Medicare affects X_1 and X_2 for Whites with more education and for minorities with less education in a linearly independent way, which allows us to identify both β_1 and β_2 in equation (13). In reality, there is further linearly independent heterogeneity in the first stage for other combinations of race-education levels, which allows us to augment the number of over-identifying restrictions.

Figure 7: **First Stage By Race and Education**



Note: The scatter plot shows the average of the variable described in the vertical axis for each age level (measured in quarters of a year) and for each subsample represented in the legend.

Table 5 shows the intention to treat or "reduced-form" estimates obtained from a local linear regression using a standard RDD.[8] As expected, results differ by race and education level. While minorities with lower levels of education tend to benefit from eligibility by discontinuously not delaying and not rationing care for cost reasons, Whites with higher levels of education tend to benefit from eligibility by discontinuously going more to hospitals. As discussed in Card, Dobkin and Maestas (2008), the heterogeneous results found in both the first stage and the intention to treat estimates suggest that perhaps Medicare eligibility might be affecting different outcome variables through different channels. Intuitively, insurance coverage (X_1) might influence health decisions that are more recurrent and less expensive (such as delaying or rationing care) more than insurance generosity (X_2), while the opposite might happen for more expensive and sporadic decisions such as the decision to get hospitalized. Next, we use our approach to make this analysis more formal.

In Table 6, we report the Two Stage Least Square estimates of β_1 and β_2 in equation (13) using different approaches. Column I shows the standard RDD estimate of β_1 under the restriction $\beta_2 = 0$, and column II shows the analogous estimate of β_2 under the restriction $\beta_1 = 0$. Columns III and IV

[8] We use a uniform kernel and add no controls to the regression besides W, $W.D$ and D. Results are robust to the choice of kernel.

Table 5: **Intention to Treat Results - Standard RD**

			Whites			Minorities		
Y	h	All	Dropout	HS Grad	>HS Grad	Dropout	HS Grad	>HS Grad
Delayed Care	1	0.002	0.017	0.025	-0.017	-0.046	0.013	0.051*
		(0.005)	(0.014)	(0.018)	(0.016)	(0.027)	(0.012)	(0.026)
	3	-0.002	0.006	0.008	-0.003	-0.048*	0.003	0.019
		(0.005)	(0.018)	(0.012)	(0.012)	(0.020)	(0.014)	(0.023)
	5	-0.009*	0.001	0.002	-0.006	-0.042*	-0.027	-0.006
		(0.004)	(0.016)	(0.010)	(0.008)	(0.015)	(0.015)	(0.021)
	7	-0.009*	0.005	-0.010*	-0.002	-0.040*	-0.015	-0.002
		(0.004)	(0.014)	(0.008)	(0.007)	(0.012)	(0.013)	(0.016)
	10	-0.018*	-0.008	-0.014*	-0.012	-0.053*	-0.027*	-0.001
		(0.004)	(0.012)	(0.007)	(0.006)	(0.010)	(0.012)	(0.014)
Rationed Care	1	0.005	0.032	-0.000	0.003	-0.047*	0.037	0.044*
		(0.005)	(0.018)	(0.016)	(0.005)	(0.016)	(0.022)	(0.013)
	3	0.000	0.017	0.003	-0.003	-0.023*	-0.008	0.006
		(0.004)	(0.020)	(0.009)	(0.006)	(0.017)	(0.019)	(0.016)
	5	-0.008	0.010	-0.012	-0.007	-0.026*	-0.029	0.016
		(0.004)	(0.016)	(0.008)	(0.005)	(0.013)	(0.017)	(0.015)
	7	-0.008*	0.010	-0.019*	-0.005	-0.030*	-0.015	0.020
		(0.003)	(0.013)	(0.007)	(0.004)	(0.011)	(0.014)	(0.013)
	10	-0.016*	-0.004	-0.019*	-0.012*	-0.038*	-0.022*	0.004
		(0.003)	(0.012)	(0.006)	(0.004)	(0.010)	(0.012)	(0.011)
Went To Hospital	1	0.007	-0.030	0.017	0.030*	-0.034	0.032*	-0.020
		(0.006)	(0.058)	(0.010)	(0.012)	(0.036)	(0.014)	(0.021)
	3	0.008	-0.008	0.017	0.013	-0.018	0.008	0.026
		(0.007)	(0.029)	(0.012)	(0.013)	(0.027)	(0.021)	(0.027)
	5	0.012	0.008	0.019	0.018	-0.019*	0.019	0.022
		(0.007)	(0.023)	(0.012)	(0.011)	(0.022)	(0.018)	(0.021)
	7	0.015*	0.005	0.027*	0.017*	-0.014	0.019	0.021
		(0.006)	(0.018)	(0.011)	(0.009)	(0.018)	(0.016)	(0.019)
	10	0.010*	0.006	0.020*	0.015*	-0.016	-0.000	0.019
		(0.005)	(0.016)	(0.010)	(0.008)	(0.016)	(0.016)	(0.018)

Note: This table shows standard "reduced form" or intention to treat estimates of the effect of Medicare eligibility on Y for the whole sample and for subsamples based on race/ethnicity and education. These were obtained from a local linear regression with uniform kernel and bandwidth h as described in the second column of the table. Standard errors clustered by values of the running variable W are shown in parentheses.

show analogous results to columns I and II using our over-identified RDD approach and each of the six combinations of race and education as elements of Z. Finally, columns V and VI show joint estimates of β_1 and β_2 using our over-identified RDD approach under no restriction in equation (13). The standard RDD univariate estimates suggest no pattern similar to the intuition discussed above. Based on these results, one would conclude that X_1 or X_2 affecting Y are similarly plausible interpretations irrespective of Y. The reason for this is that the standard RDD is not exploiting the over-identifying restrictions which are generated by the heterogeneity in the first stage by values of Z, as we exploit in our approach. Our approach in the univariate version, which exploits the heterogeneity in the first stage but does not add both X_1 and X_2 in the same regression, makes this pattern sharper: while we find that β_1 is significant mostly for "Delayed Care" and "Rationed Care", we find that β_2 is significant mostly for "Hospital Stay".

One important concern with the univariate approaches is that their exclusion restriction might be invalid. In columns I and III, it is assumed that X_2 does not affect Y, so the effect of Medicare eligibility on Y is allowed to operate only through X_1. In contrast, in columns II and IV it is assumed that X_1 does not affect Y, so the effect of Medicare eligibility on Y is allowed to operate only through X_2. If these restrictions are not valid, then these estimates will be biased. In columns V and VI, we present our approach in a multivariate setting in order to avoid making these exclusion restrictions. As expected, the multivariate estimates suggest an even sharper pattern than the results of our approach in the univariate setting: (a) insurance coverage (X_1) reduces the probability of delaying care for cost reasons, or rationing care for cost reasons, in about 15 to 20 percentage points, depending on the bandwidth, but seem to have little or no effect on hospital stays; (b) having a second, more generous insurance (X_2) increases the probability of hospital stays in about 5 percentage points, but seems to have little or no effect on delaying or rationing care for cost reasons.

It is interesting to note that the univariate approaches perform better when the underlying restrictions are more plausible. To see this, consider the results from our multivariate approach (columns (V) and (VI)). They suggest that it is more plausible to assume that $\beta_2 = 0$ for the first two outcome variables in the table (Delayed and Rationed Care) than to assume that $\beta_1 = 0$. Indeed, the univariate estimates are more similar to the multivariate estimates when it is assumed $\beta_2 = 0$ (columns I and III) than when it is assumed $\beta_1 = 0$ (columns II and IV). Analogously, multivariate estimates suggest that it is more plausible to assume that $\beta_1 = 0$ for the third outcome variable in the table (Went to Hospital) than to assume that $\beta_2 = 0$. Indeed, the univariate estimates are closer to the multivariate estimates when it is assumed $\beta_1 = 0$ (columns II and IV) than when it is assumed $\beta_2 = 0$ (columns I and III). Moreover, irrespective of the plausibility of the restriction, the estimates of our over-identified univariate approach are more similar to our preferred estimates in columns V and VI than the standard RDD estimates. Intuitively, this happens because our univariate approach exploits the heterogeneity in the first stage while the standard RDD does not.

Table 6: **Two Stage Least Square Results**

		Univariate				Multivariate	
		Standard RDD		Our Approach		Our Approach	
		(I)	(II)	(III)	(IV)	(V)	(VI)
Y	h	$\beta_2 = 0$, $\hat{\beta}_1$	$\beta_1 = 0$, $\hat{\beta}_2$	$\beta_2 = 0$, $\hat{\beta}_1$	$\beta_1 = 0$, $\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Delayed Care	1	0.030	0.006	-0.053	0.016	-0.235	0.059*
		(0.053)	(0.011)	(0.060)	(0.009)	(0.144)	(0.026)
	3	-0.022	-0.004	-0.109	0.001	-0.200*	0.032*
		(0.057)	(0.011)	(0.059)	(0.009)	(0.078)	(0.013)
	5	-0.092*	-0.019*	-0.118*	-0.011	-0.160*	0.015
		(0.043)	(0.009)	(0.041)	(0.008)	(0.059)	(0.012)
	7	-0.091*	-0.019*	-0.116*	-0.012	-0.158*	0.014
		(0.036)	(0.008)	(0.035)	(0.007)	(0.053)	(0.011)
	10	-0.190*	-0.037*	-0.198*	-0.029*	-0.211*	-0.004
		(0.036)	(0.007)	(0.033)	(0.007)	(0.049)	(0.010)
Rationed Care	1	0.067	0.014	-0.037	0.020*	-0.227*	0.061
		(0.062)	(0.013)	(0.051)	(0.010)	(0.098)	(0.015)
	3	0.003	0.001	-0.043	0.002	-0.086	0.015
		(0.052)	(0.010)	(0.048)	(0.008)	(0.065)	(0.011)
	5	-0.082	-0.017	-0.079*	-0.013*	-0.076	-0.001
		(0.042)	(0.009)	(0.036)	(0.007)	(0.049)	(0.010)
	7	-0.087*	-0.018*	-0.096*	-0.016*	-0.099*	0.001
		(0.033)	(0.007)	(0.033)	(0.006)	(0.047)	(0.008)
	10	-0.169*	-0.033*	-0.156*	-0.027*	-0.133*	-0.007
		(0.032)	(0.006)	(0.032)	(0.005)	(0.044)	(0.007)
Went to Hospital	1	0.089	0.187	-0.082	0.022	-0.344	0.084*
		(0.072)	(0.015)	(0.081)	(0.017)	(0.185)	(0.040)
	3	0.103	0.020	-0.007	0.021	-0.119	0.040
		(0.084)	(0.016)	(0.069)	(0.018)	(0.138)	(0.033)
	5	0.130	0.027	0.025	0.030*	-0.106	0.047
		(0.070)	(0.015)	(0.060)	(0.015)	(0.101)	(0.024)
	7	0.155*	0.032*	0.056	0.034*	-0.099	0.050*
		(0.058)	(0.012)	(0.051)	(0.012)	(0.090)	(0.021)
	10	0.112*	0.022*	0.026	0.024*	-0.112	0.042*
		(0.056)	(0.011)	(0.050)	(0.011)	(0.082)	(0.017)

Note: This table shows 2SLS estimates of the effect of (a) X_1 (columns I and III) or (b) X_2 (columns II and IV), or (c) both X_1 and X_2 (columns V and VI) on Y using different approaches. In the univariate case, each column refers to estimates obtained from a different regression, while in the multivariate case both columns refer to estimates obtained from the same regression. These estimates were obtained from a local linear regression with uniform kernel and bandwidth h. Standard errors clustered by values of the running variable W are shown in parentheses.

Table 7: **Testing for Heterogeneity**

Z_1	h	Delayed		Rationed		Went to Hospital		First Stage	
		X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
Race	1	1.51	2.33	5.23	8.06	1.61	2.76	114.67	2,504.62
		(0.22)	(0.13)	(0.02)	(0.00)	(0.20)	(0.10)	(0.00)	(0.00)
	3	1.16	1.55	1.36	0.72	0.02	0.30	26.91	46.48
		(0.28)	(0.21)	(0.24)	(0.40)	(0.88)	(0.58)	(0.00)	(0.00)
	5	1.08	0.03	2.37	1.68	0.16	0.69	34.33	50.22
		(0.30)	(0.87)	(0.12)	(0.19)	(0.68)	(0.41)	(0.00)	(0.00)
	7	1.59	0.54	3.64	5.90	0.07	0.51	23.17	35.26
		(0.21)	(0.46)	(0.06)	(0.02)	(0.79)	(0.47)	(0.00)	(0.00)
	10	3.68	1.85	3.12	3.21	0.19	0.35	27.75	48.64
		(0.06)	(0.17)	(0.08)	(0.07)	(0.67)	(0.55)	(0.00)	(0.00)
Education	1	0.25	2.82	0.98	1.03	0.46	0.60	114.67	2,504.62
		(0.88)	(0.24)	(0.61)	(0.60)	(0.80)	(0.74)	(0.00)	(0.00)
	3	1.27	1.05	0.61	1.32	0.83	0.16	26.91	46.48
		(0.53)	(0.59)	(0.74)	(0.52)	(0.66)	(0.92)	(0.00)	(0.00)
	5	0.79	1.39	3.35	3.93	1.07	0.33	34.33	50.22
		(0.67)	(0.50)	(0.19)	(0.14)	(0.59)	(0.85)	(0.00)	(0.00)
	7	0.32	1.95	2.03	3.18	0.76	0.24	23.17	35.26
		(0.85)	(0.38)	(0.36)	(0.20)	(0.69)	(0.89)	(0.00)	(0.00)
	10	1.49	3.46	2.22	2.94	0.73	0.75	27.75	48.64
		(0.47)	(0.18)	(0.33)	(0.23)	(0.70)	(0.69)	(0.00)	(0.00)

Note: This table shows for each bandwidth h the F-test along with corresponding pvalue of the test for whether all elements of the β_1 (odd columns) or β_2 (even columns) are equal to each other. Each panel allows for unrestricted heterogeneity of β_1 and β_2 for each different value of the variable Z_1, referenced in the first column of the table. For comparison, the last two columns show an analogous test about the 6 coefficients of the first stage (defined by all combinations of race and education).

Finally, we provide some evidence in favor of our key assumption that the parameters of interest (β_1 and β_2 in equation (13)) do not vary with race or education. As discussed in Section 2.3, we can relax this assumption by allowing for heterogeneity in covariates. In Table 7, we allow for heterogeneity along one of these two dimensions (race, education) at a time, and we perform F-tests for whether the parameters of either X_1 (odd columns) or X_2 (even columns) are the same for all levels of race (first panel) or education (second panel). For comparison, in the last two columns we show results of an analogous test for whether the parameters of the first stage regression of X_1 or X_2 are the same for all six combinations of race and education. Perhaps with the exception of heterogeneity by race for

the outcome variable rationed care, all results point to homogeneity of β_1 and β_2. In all cases, the evidence of heterogeneity in the first stage is overwhelming relative to any evidence of heterogeneity in β_1 and β_2. Taken together, the results suggest that the main idea of exploiting heterogeneity in the first stage effects along these two dimensions while maintaining the assumption of homogeneity in the main effects of interest along these same dimensions fits this application well.

Overall, our results corroborate in a sharper way the intuition suggested by Card, Dobkin and Maestas (2008) that Medicare eligibility has generated different effects on health behavior depending on whether the person would otherwise have no insurance, or would otherwise have some insurance, albeit a less generous one.

6 Conclusions

This article has proposed a new identification and estimation strategy for RDD. It explores the heterogeneity in the "first stage" discontinuities for different values of a covariate to generate over-identifying restrictions. This allows us to identify quantities which cannot be identified with the standard RDD method, including the effects of multiple endogenous variables, multiple marginal effects of a multivalued endogenous variable, and heterogeneous effects conditional on covariates. For a linear model with a single endogenous variable, when both standard RDD and our approach are applicable, our method yields smaller MSE that are less sensitive to bandwidths than standard RDD procedures.

For multivalued treatments none of the existing RDD procedures identify the marginal effects. We have provided a method that achieves identification in a variety of situations, including situations with variable marginal effects. We have proposed a simple TSLS estimator implementing our identification strategy. Monte Carlo simulations confirm its excellent finite sample performance relative to standard RDD estimators, and the empirical application to health utilization shows its ability to identify multiple marginal effects in a situation of practical interest.

Section A in the Appendix generalizes the identification result to nonparametric variable marginal effects. It should be possible to extend our TSLS procedure to nonparametric cases by using, for example, sieve methods or local polynomial methods. Similarly, nonparametric tests of homogeneous effects could be developed based on these nonparametric estimators. These methods would certainly require larger sample sizes to achieve a similar level of precision to those developed in the paper. Furthermore, these would involve the choice of additional bandwidth parameters. Similarly, it is possible to extend our identification result to the nonparametric continuous treatment case. This would entail a new completeness condition between the continuous treatment and the vector of covariates, similar to the covariance completeness condition studied in Caetano and Escanciano (2017).

The proposed methods enable researchers to estimate multiple marginal effects in situations that inherently have limited exogenous variation. We have shown how covariates bring new possibilities that were currently unexploited. We hope this research will foster a deeper understanding of the effects of policies on economic outcomes and a more efficient use of the data, in particular, covariates.

A Appendix A: A General Identification Result

This section establishes an identification result that includes the identification result in the main text as a special case. This section also contains proofs for these identification results. Set $Z = (Z_1, Z_2)$, where Z_1 and Z_2 have dimensions q_1 and q_2, respectively (we allow for the possibility of $q_1 = 0$, meaning Z_1 is empty). We consider the following generalization of model (1)

$$Y_i = g(T_i, Z_{1i}, W_i) + h(Z_i, W_i, \varepsilon_i), \tag{14}$$

where now the unknown function g depends on Z_{1i} in addition to the dependence on the multivalued treatment variable T_i and W_i. Arguing as in the main text, we can write (14) as the varying coefficients model

$$Y_i = \beta(Z_{1i}, W_i)' X_i + H_i, \tag{15}$$

where $H_i = g(t_0, Z_{1i}, W_i) + h(Z_i, W_i, \varepsilon_i)$, and now the marginal effects depend on Z_{1i} and W_i,

$$\beta(Z_{1i}, W_i) = (g(t_1, Z_{1i}, W_i) - g(t_0, Z_{1i}, W_i), ..., g(t_d, Z_{1i}, W_i) - g(t_{d-1}, Z_{1i}, W_i))'.$$

The following condition generalizes the relevance condition (2).

Assumption 3 *$E[\delta_X \delta_X' | Z_1]$ is positive definite a.s.*

Theorem A.1 *Let (14) and Assumption 1 hold. Then, $\beta_0(z_1) \equiv \beta(z_1, w_0)$ is identified if and only if Assumption 3 holds.*

Proof. *We first prove the "if" part. Taking limits as in the main text we obtain the equation*

$$\delta_Y(Z) = \delta_X'(Z)\beta_0(Z_1).$$

Multiplying by $\delta_X(Z)$ both sides, and taking conditional means on Z_1, we arrive at

$$E[\delta_X(Z)\delta_Y(Z)|Z_1] = E[\delta_X(Z)\delta_X'(Z)|Z_1]\beta_0(Z_1).$$

This and the generalized relevance condition yield identification, i.e.

$$\beta_0(Z) = \left(E[\delta_X(Z)\delta_X'(Z)|Z_1]\right)^{-1} E[\delta_X(Z)\delta_Y(Z)|Z_1].$$

For the necessity part, we suppose the generalized relevance condition does not hold. This means there exists a non-trivial measurable function $\lambda(Z_1)$ such that, a.s.,

$$\lambda(Z_1)' E[\delta_X(Z)\delta_X'(Z)|Z_1]\lambda(Z_1) = 0.$$

Hence,

$$E[\left(\lambda(Z_1)'\delta_X(Z)\right)^2] = 0,$$

and thus

$$\lambda(Z_1)'\delta_X(Z) = 0 \ a.s. \tag{16}$$

Let $\beta_0(Z_1)$ denote the true value that generated the data, and define $\tilde{\beta}(Z_1) = \beta_0(Z_1) + \lambda(Z_1)$ and $\tilde{H}_i = H_i - \lambda'(Z_{1i})X_i$. Note that if H_i satisfies Assumption 1, $\tilde{H}_i$ also does it because by (16),

$$\lim_{w \downarrow w_0} E[\tilde{H}|W = w, Z] - \lim_{w \uparrow w_0} E[\tilde{H}|W = w, Z] = \lim_{w \downarrow w_0} E[H|W = w, Z] - \lim_{w \uparrow w_0} E[H|W = w, Z] - \lambda'(Z_1)\delta_X(Z) = 0.$$

Hence, the triple $(\alpha(Z), \tilde{\beta}(Z_1), \tilde{H})$ satisfies the same conditions as the triple $(\alpha(Z), \beta_0(Z), H)$, and $\beta_0(\cdot)$ is not identified. ■

Theorem 2.1 is a special case of Theorem A.1 (the case $q_1 = 0$). In this case Z_1 is empty, $\beta_0(Z_1)$ is a constant and

$$E[\delta_X(Z)\delta_X'(Z)|Z_1] = E[\delta_X(Z)\delta_X'(Z)].$$

Another special case is $q_2 = 0$. In this case $Z = Z_1$ and

$$E[\delta_X(Z)\delta_X'(Z)|Z_1] = \delta_X(Z)\delta_X'(Z).$$

Within this case, there are two separate subcases: (i) $d = 1$ and (ii) $d > 1$. In (i) the relevance condition holds iff $\delta_X(Z) \neq 0$ with positive probability. In (ii), since the rank of $\delta_X(Z)\delta_X'(Z)$ is at most one, the relevance condition fails, and hence by Theorem A.1, identification fails. We summarize this impossibility result in the following corollary.

Corollary A.1 *Let (14) and Assumption 1 hold. Then, if $q_2 = 0$ and $d > 1$, nonparametric identification is not possible.*

This impossibility result shows that in order to identify variable marginal effects, some semiparametric restriction is needed. One such restriction is $q_2 > 0$. To better explain the meaning of the relevance condition in the $q_2 > 0$ case, suppose Z_j is discrete with support $\{z_{j1}, ..., z_{jm_j}\}$, for $j = 1, 2$. For each b fixed, $1 \leq b \leq m_1$, define the $m_2 \times d$ matrix Δ_b with $l-$th row $\delta_X(z_{1b}, z_{2l})$, for $1 \leq l \leq m_2$. Then, the relevance condition holds, iff for all b, $1 \leq b \leq m_1$,

$$rank\,(\Delta_b) = d.$$

This requires the order condition $m_2 \geq d$.

B Appendix: Proof of Asymptotic Results

In this section we establish the asymptotic normality of the proposed estimator $\hat{\beta}$. For simplicity, we consider the case where X and Z are univariate, although the extension to the multivariate case only involves further notation. Without loss of generality assume hereinafter that $w_0 = 0$. We introduce some further notation and assumptions. Let $\varepsilon_{V_i} = V_i - E[V_i|W_i, Z_i]$ denote the regression errors for $V = Y$ and $V = X$. Define $\varepsilon_{U_i} = \varepsilon_{Y_i} - \beta_0\varepsilon_{X_i}$. Assume, for $V = Y$ and $V = X$,

$$E[V_i|W_i, Z_i] = \alpha_{0V}(W_i) + \alpha_{1V}(W_i)Z_i, \tag{17}$$

We investigate the asymptotic properties of $\hat{\beta}$ under the following assumptions, which parallel those of Hahn, Todd and van der Klaauw (1990):

Assumption 4 *Suppose that*

1. *The sample* $\{\chi_i\}_{i=1}^n$ *is an iid sample, where* $\chi_i = (Y_i, X_i, Z_i, W_i)$.

2. *The density of* W, $f(w)$, *is continuous and bounded near* $w = 0$. *It is also bounded away from zero near* $w = 0$. Γ *defined below is positive definite.* Δ_X *is full column rank.*

3. *The kernel* k *is continuous, symmetric and nonnegative-valued with compact support.*

4. *The functions* $\mu_j(w) = E[Z^j|W = w]$, $\mu_{Xj}(w) = E[Z^j X|W = w]$, $\mu_{Yj}(w) = E[Z^j Y|W = w]$, $\sigma_j^2(w) = E[Z_i^{2j}|W = w]$, $\sigma_{Xj}^2(w) = E[Z_i^{2j}X_i^2|W = w]$, $\sigma_{Yj}^2(w) = E[Z_i^{2j}Y_i^2|W = w]$, $q_j(w) = E[Z^{2j}\varepsilon_{U_i}^2|W = w]$, *and* $s_j(w) = E[Z^{3j}\varepsilon_{U_i}^3|W = w]$ *are uniformly bounded near* $w = 0$, *with well-defined and finite left and right limits to* $w = 0$, *for* $j = 0, 1$ *and* 2.

5. *The bandwidth satisfies* $nh_n^5 \to 0$.

6. *For* $V = Y$ *and* $V = X$: *(i) let 17 hold; (ii) for* $w > 0$ *or* $w < 0$, $\alpha_{0V}(w)$ *and* $\alpha_{1V}(w)$ *are twice continuously differentiable; (iii) there exists some* $M > 0$ *such that* $\dot{\alpha}_{jV}^+(w) = \lim_{u\downarrow w} \partial\alpha_{jV}(u)/\partial u$ *and* $\ddot{\alpha}_{jV}^+(w) = \lim_{u\downarrow w} \partial^2\alpha_{jV}(u)/\partial u^2$ *are uniformly bounded on* $(0, M]$, *for* $j = 0, 1$. *Similarly,* $\dot{\alpha}_{jV}^-(w) = \lim_{u\uparrow w} \partial\alpha_{jV}(u)/\partial u$ *and* $\ddot{\alpha}_{jV}^-(w) = \lim_{u\uparrow w} \partial^2\alpha_{jV}(u)/\partial u^2$ *are uniformly bounded on* $[-M, 0)$, *for* $j = 0, 1$.

For a measurable function of the data $g(\chi_i)$, define the local linear sample mean

$$\hat{E}[g(\chi_i)] = \frac{1}{nh_n}\sum_{i=1}^n g(\chi_i)k_{ih_n}.$$

Let S_{in} and C_{in} be defined the same as S_i and C_i but with W_i replaced by W_i/h_n. Define $\tilde{X}_i = (X_i, C_i')'$, $\tilde{X}_{in} = (X_i, C_{in}')'$, and $\theta = (\beta_0, \eta')'$, where $\eta = (\eta_1, ..., \eta_6)'$ has the same dimension as C_{in}. Define $\theta_n = (\beta_0, \eta_1, \eta_2, h_n\eta_3, h_n\eta_4, h_n\eta_5, h_n\eta_6)'$. With this notation in place, the TSLS is the first component of

$$\begin{aligned}\hat{\theta}_n &= \left(\hat{E}\left[\tilde{X}_{in}S_{in}'\right]\left(\hat{E}\left[S_{in}S_{in}'\right]\right)^{-1}\hat{E}\left[S_{in}\tilde{X}_{in}'\right]\right)^{-1}\hat{E}\left[\tilde{X}_{in}S_{in}'\right]\left(\hat{E}\left[S_{in}S_{in}'\right]\right)^{-1}\hat{E}\left[S_{in}Y_i\right] \\ &= \theta_n + \left(\hat{E}\left[\tilde{X}_{in}S_{in}'\right]\left(\hat{E}\left[S_{in}S_{in}'\right]\right)^{-1}\hat{E}\left[S_{in}\tilde{X}_{in}\right]\right)^{-1}\hat{E}\left[\tilde{X}_{in}S_{in}'\right]\left(\hat{E}\left[S_{in}S_{in}'\right]\right)^{-1}\hat{E}\left[S_{in}U_i\right],\end{aligned}$$

where $U_i = Y_i - \tilde{X}_{in}'\theta_n = Y_i - \tilde{X}_i'\theta$.

We show in Lemma B.11 that

$$\sqrt{nh_n}\left(\hat{\theta}_n - \theta_n\right) \to_d N(0, \Omega), \tag{18}$$

and provide an expression for Ω. The asymptotic normality for $\sqrt{nh_n}\left(\hat{\beta}_n - \beta_0\right)$ then follows as

$$\sqrt{nh_n}\left(\hat{\beta}_n - \beta_0\right) \to_d N(0, \Sigma),$$

where $\Sigma = e_0'\Omega e_0$, and e_0 has a one in the first entry, corresponding to X, and zero everywhere else.

We introduce some notation that will be used throughout,

$$\gamma_l = \int_0^\infty u^l k(u)du,$$

$$\mu_j^+ = \lim_{w\downarrow 0} E[Z^j|W=w] \qquad \mu_j^- = \lim_{w\uparrow 0} E[Z^j|W=w],$$

$$k_{ih_n}^+ = k(W/h_n)1(W \geq 0) \qquad k_{ih_n}^- = k(W/h_n)1(W<0),$$

$$b_{lj}^+ = \gamma_l\mu_j^+,\ b_{lj}^- = \gamma_l\mu_j^-,\ b_{xlj}^+ = \gamma_l\mu_{Xj}^+,\ b_{xlj}^- = \gamma_l\mu_{Xj}^-.$$

Lemma B.1 *Under Assumption 4 1-5,*

$$\hat{E}\left[\tilde{X}_{in}S_{in}'\right] \to_p \Delta_X,$$

where

$$\Delta_X = f(0)\begin{bmatrix} b_{x00}^+ & b_{x01}^+ & b_{x10}^+ & b_{x11}^+ & b_{x00}^- & b_{x01}^- & b_{x10}^- & b_{x11}^- \\ b_{00}^+ & b_{01}^+ & b_{10}^+ & b_{11}^+ & b_{00}^- & b_{01}^- & b_{10}^- & b_{11}^- \\ b_{01}^+ & b_{02}^+ & b_{11}^+ & b_{12}^+ & b_{01}^- & b_{02}^- & b_{11}^- & b_{12}^- \\ b_{10}^+ & b_{11}^+ & b_{20}^+ & b_{21}^+ & 0 & 0 & 0 & 0 \\ b_{11}^+ & b_{12}^+ & b_{21}^+ & b_{22}^+ & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b_{10}^- & b_{11}^- & b_{20}^- & b_{21}^- \\ 0 & 0 & 0 & 0 & b_{11}^- & b_{12}^- & b_{21}^- & b_{22}^- \end{bmatrix}.$$

Proof. *Let*

$$\theta_{lj}^+ = \frac{1}{nh_n}\sum_{i=1}^n \left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^+, \qquad l,j = 0,1,2.$$

Then, by the change of variables $u = w/h_n$,

$$\begin{aligned} E[\theta_{lj}^+] &= h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^+\right] \\ &= \int_0^\infty u^l k(u)\mu_j(uh_n)f(uh_n)du \\ &= f(0)\gamma_l\mu_j^+ + o(1), \end{aligned}$$

where $\mu_j(w) = E[Z^j|W=w]$ and the convergence follows by the Dominated Convergence theorem. As for the variance

$$\begin{aligned} Var(\theta_{lj}^+) &\leq \left(nh_n^2\right)^{-1} E\left[\left(\frac{W_i}{h_n}\right)^{2l} Z_i^{2j} k_{ih_n}^{+2}\right] \\ &= (nh_n)^{-1}\int_0^\infty u^{2l}k^2(u)\sigma_j^2(uh_n)f(uh_n)du \\ &= o(1), \end{aligned}$$

again by the Dominated Convergence theorem.

Similarly, define

$$\theta^+_{xlj} = \frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^l Z_i^j X_i k^+_{ih_n}, \qquad l,j = 0,1,2.$$

Then, by the change of variables $u = w/h_n$,

$$\begin{aligned} E[\theta^+_{xlj}] &= h_n^{-1} E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j X_i k^+_{ih_n}\right] \\ &= \int_0^\infty u^l k(u)\mu_{Xj}(uh_n)f(uh_n)du \\ &= f(0)\gamma_l\mu^+_{Xj} + o(1), \end{aligned}$$

by the Dominated Convergence theorem. As for the variance

$$\begin{aligned} Var(\theta^+_{xlj}) &\leq \left(nh_n^2\right)^{-1} E\left[\left(\frac{W_i}{h_n}\right)^{2l} Z_i^{2j} X_i^2 k^{+2}_{ih_n}\right] \\ &= (nh_n)^{-1}\int_0^\infty u^{2l}k^2(u)\sigma^2_{Xj}(uh_n)f(uh_n)du \\ &= o(1), \end{aligned}$$

again by the Dominated Convergence theorem. The proof for θ^-_{lj} *and* θ^-_{xlj}*, which replace* $k^+_{ih_n}$ *by* $k^-_{ih_n}$*, is analogous, and hence omitted.* ■

Lemma B.2 *Under Assumption 4 1-5,*

$$\hat{E}\left[S_{in}S_{in}'\right] \to_p \Gamma$$

where

$$\Gamma = f(0)\begin{bmatrix} b^+_{00} & b^+_{01} & b^+_{10} & b^+_{11} & 0 & 0 & 0 & 0 \\ b^+_{01} & b^+_{02} & b^+_{11} & b^+_{12} & 0 & 0 & 0 & 0 \\ b^+_{10} & b^+_{11} & b^+_{20} & b^+_{21} & 0 & 0 & 0 & 0 \\ b^+_{11} & b^+_{12} & b^+_{21} & b^+_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b^-_{00} & b^-_{01} & b^-_{10} & b^-_{11} \\ 0 & 0 & 0 & 0 & b^-_{01} & b^-_{02} & b^-_{11} & b^-_{12} \\ 0 & 0 & 0 & 0 & b^-_{10} & b^-_{11} & b^-_{20} & b^-_{21} \\ 0 & 0 & 0 & 0 & b^-_{11} & b^-_{12} & b^-_{21} & b^-_{22} \end{bmatrix}.$$

Proof. *The proof is analogous to that of Lemma B.1 and hence is omitted.* ■

Lemma B.3 *Under Assumption 4 1-5,*

$$\hat{E}\left[S_{in}Y_i\right] \to_p \Delta_Y,$$

where $\Delta_Y = f(0)(b^+_{y00}, b^+_{y10}, b^+_{y01}, b^+_{y11}, b^-_{y00}, b^-_{y10}, b^-_{y01}, b^-_{y11})'$*,* $b^+_{ylj} = \gamma_l\mu^+_{Yj}$*,* $b^-_{ylj} = \gamma_l\mu^-_{Yj}$*.*
Proof. *The proof is analogous to that of Lemma B.1 and hence is omitted.* ■

Lemma B.4 *Assume 1, 2 and 4. Then,*

$$\Delta_Y = \Delta_X' \theta_0,$$

where $\theta_0 = (\beta_0, \eta_1, \eta_2, 0, 0, 0, 0)'$.

Proof. *By Assumption 4 6, we can write*

$$E[H_i | Z_i, W = w] := \alpha_{0H}(w) + \alpha_{1H}(w) Z_i.$$

Hence, this last display and Assumption 1 yield

$$\mu_{Yj}(w) = \beta(w)\mu_{Xj}(w) + \alpha_{0H}(w)\mu_j(w) + \alpha_{1H}(w)\mu_{j+1}(w).$$

Taking right and left limits at $w = 0$ *imples the desired equality* $\Delta_Y = \Delta_X' \theta_0$, *where* $\eta_1 = \alpha_{0H}(0)$ *and* $\eta_2 = \alpha_{1H}(0)$. ■

The previous Lemmas show the consistency of $\hat{\beta}$, since

$$\hat{\theta}_n \to_p \left(\Delta_X \Gamma^{-1} \Delta_X'\right)^{-1} \Delta_X \Gamma^{-1} \Delta_Y = \theta_0$$

We prove several Lemmas that will yield the asymptotic normality of $\sqrt{nh_n}\left(\hat{\theta}_n - \theta_n\right)$, and hence of $\sqrt{nh_n}\left(\hat{\beta}_n - \beta_0\right)$.

Define the function

$$\begin{aligned}\zeta_H(w, z) = \alpha_{0H}(w) + \alpha_{1H}(w) z - \alpha_{0H}^+ + \alpha_{1H}^+ z \\ - \left(\dot{\alpha}_{0H}^+ + \dot{\alpha}_{1H}^+ z\right) w - \frac{1}{2}\left(\ddot{\alpha}_{0H}^+ + \ddot{\alpha}_{1H}^+ z\right) w^2,\end{aligned}$$

where $\dot{\alpha}_{0H}^+ = \lim_{w \downarrow 0} \partial \alpha_{0H}(w)/\partial w$ and $\ddot{\alpha}_{0H}^+ = \lim_{w \downarrow 0} \partial^2 \alpha_{0H}(w)/\partial w^2$, and similarly for α_{1H}. We use later that

$$\sup_{0 < w < M h_n} |\zeta_H(w, z)| = o(h_n^2)(1 + |Z|).$$

The function $\zeta_H(w, z)$ is the Taylor's remainder of order two of $E[H_i | Z_i, W = w] := \alpha_{0H}(w) + \alpha_{1H}(w) Z_i$ around $w = 0$. We can also relate the coefficients in this expansion with the coefficients in η. Following the arguments above, it can be shown that

$$\tilde{\eta} = arg \min_{\beta, \eta} \sum_{i=1}^{n} \left(H_i - \eta' C_i\right)^2 k_{h_n}(W_i).$$

estimates consistently η. Thus,

$$\begin{aligned}&\eta_1 = \alpha_{0H}(0),\ \eta_2 = \alpha_{1H}(0),\ \eta_3 = \dot{\alpha}_{0H}(0), \\ &\eta_4 = 0,\ \eta_5 = \dot{\alpha}_{1H}(0),\ \eta_6 = 0.\end{aligned}$$

Recall $U_i = Y_i - \tilde{X}_i' \theta$. Then, from the definitions above

$$\begin{aligned}E[U_i | Z_i, W] &= E[H_i | Z_i, W] - \alpha_{0H}^+ + \alpha_{1H}^+ Z - \dot{\alpha}_{0H}^+ W + \dot{\alpha}_{1H}^+ ZW \\ &= \frac{1}{2}\left(\ddot{\alpha}_{0H}^+ + \ddot{\alpha}_{1H}^+ Z_i\right) W_i^2 + \zeta_H(W_i, Z_i).\end{aligned}$$

The following Lemmas make use of Assumptions 1, 2 and 4.

Lemma B.5 (Numerator: Expectation)

$$E\left[\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}U_i k_{ih_n}\right] \to_p \frac{1}{2}f(0)h_n^2(b_U + o(1)),$$

where

$$b_U = \begin{bmatrix} \gamma_2(\ddot{\alpha}_{0H}^+ + \ddot{\alpha}_{0H}^+\mu_1^+) \\ \gamma_2(\ddot{\alpha}_{0H}^+\mu_1^+ + \ddot{\alpha}_{0H}^+\mu_2^+) \\ \gamma_3(\ddot{\alpha}_{0H}^+ + \ddot{\alpha}_{0H}^+\mu_1^+) \\ \gamma_3(\ddot{\alpha}_{0H}^+\mu_1^+ + \ddot{\alpha}_{0H}^+\mu_2^+) \\ \gamma_2(\ddot{\alpha}_{0H}^- + \ddot{\alpha}_{0H}^+\mu_1^-) \\ \gamma_2(\ddot{\alpha}_{0H}^-\mu_1^- + \ddot{\alpha}_{0H}^-\mu_2^+) \\ \gamma_3(\ddot{\alpha}_{0H}^- + \ddot{\alpha}_{0H}^-\mu_1^+) \\ \gamma_3(\ddot{\alpha}_{0H}^-\mu_1^+ + \ddot{\alpha}_{0H}^-\mu_2^+) \end{bmatrix}.$$

Proof. *Let*

$$u_{lj}^+ = \frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^l Z_i^j U_i k_{ih_n}^+, \qquad l,j = 0,1.$$

Then, write

$$\begin{aligned} E[u_{lj}^+] &= h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j U_i k_{ih_n}^+\right] \\ &= h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j \left(\frac{1}{2}\left(\ddot{\alpha}_{0H}^+ + \ddot{\alpha}_{1H}^+ Z_i\right)W_i^2 + \zeta_H(W_i,Z_i)\right)k_{ih_n}^+\right] \\ &= h_n^{-1}\frac{1}{2}\ddot{\alpha}_{0H}^+ E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j W_i^2 k_{ih_n}^+\right] + h_n^{-1}\frac{1}{2}\ddot{\alpha}_{1H}^+ E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^{j+1} W_i^2 k_{ih_n}^+\right] \\ &+ h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j \zeta_H(W_i,Z_i) k_{ih_n}^+\right]. \end{aligned}$$

By the change of variables $u = w/h_n$,

$$\begin{aligned} h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j W_i^2 k_{ih_n}^+\right] &= h_n^2\int_0^\infty u^{l+2}k(u)\mu_j(uh_n)f(uh_n)du \\ &= h_n^2\mu_j^+ f(0^+)\gamma_{l+2} + o(1), \end{aligned}$$

and similarly

$$h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^{j+1} W_i^2 k_{ih_n}^+\right] = h_n^2\mu_{j+1}^+ f(0^+)\gamma_{l+2} + o(1).$$

On the other hand, assume without loss of generality that $[-M,M]$ contains the support of k, so that

$$h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j \zeta_H(W_i,Z_i) k_{ih_n}^+\right] = o(h_n^2).$$

The proof for the left limit version is analogous, and hence omitted. ■

Lemma B.6 (Numerator: Conditional Expectation)

$$\frac{1}{nh_n}\sum_{i=1}^{n} E[S_{in}U_i k_{ih_n}|W_i, Z_i] = \frac{1}{nh_n}\sum_{i=1}^{n} E[S_{in}U_i k_{ih_n}] + o_p(h_n^2).$$

Proof. *We have*

$$\begin{aligned}\frac{1}{nh_n}\sum_{i=1}^{n} E[S_{in}U_i k_{ih_n}|W_i, Z_i] &= \frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}\left(\frac{1}{2}\left(\ddot{\alpha}_{0H}^{+} + \ddot{\alpha}_{1H}^{+}Z_i\right)W_i^2 + \zeta_H(W_i, Z_i)\right)\\ &= \frac{1}{2}\ddot{\alpha}_{0H}^{+}\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}W_i^2 + \frac{1}{2}\ddot{\alpha}_{1H}^{+}\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}Z_iW_i^2\\ &+ \frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}\zeta_H(W_i, Z_i).\end{aligned}$$

Observe that

$$\begin{aligned}Var\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}W_i^2\right) &= \left(nh_n^2\right)^{-1} Var\left(S_{in}k_{ih_n}W_i^2\right)\\ &\leq C\left(nh_n\right)^{-1} h_n^{-1} E\left[S_{in}S_{in}'k_{ih_n}^2W_i^4\right]\\ &= O\left(\left(nh_n\right)^{-1}h_n^4\right)\\ &= o(1),\end{aligned}$$

since for $l, j = 0, 1, 2$

$$\begin{aligned}h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^{+2}W_i^4\right] &= h_n^4\int_0^\infty u^l k^2(u)\mu_j(uh_n)f(uh_n)du\\ &= h_n^4\mu_j^{+}f(0^+)v_l + o(1),\end{aligned}$$

where

$$v_l = \int_0^\infty u^l k^2(u)du,$$

and similarly

$$h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^{-2}W_i^4\right] = h_n^4\mu_j^{-}f(0^-)v_l + o(1).$$

Likewise,

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}Z_iW_i^2\right) = o(1).$$

and

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}\zeta_H(W_i, Z_i)\right) = o(1).$$

Note that

$$\frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}^{+}\left(U_i - E[U_i|W_i, Z_i]\right) = \frac{1}{nh_n}\sum_{i=1}^{n} S_{in}k_{ih_n}^{+}\varepsilon_{U_i},$$

where $\varepsilon_{U_i} = U_i - E[U_i|W_i, Z_i]$ *denotes the regression error. Then, we have the following result.* ■

Lemma B.7 (Numerator: Conditional Variance)

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{+in}k_{ih_n}\varepsilon_{U_i}\right) = \frac{1}{nh_n}\Sigma_{U^+} + o(1),$$

where

$$\Sigma_{U^+} = f(0^+)\begin{bmatrix} v_0 & v_0q_1^+ & v_2 & v_2q_1^+ \\ v_0q_1^+ & v_0q_2^+ & v_2q_1^+ & v_2q_2^+ \\ v_2 & v_2q_1^+ & v_4 & v_4q_1^+ \\ v_2q_1^+ & v_2q_2^+ & v_4q_1^+ & v_4q_2^+ \end{bmatrix},$$

$$v_l = \int_0^\infty u^l k^2(u)du \text{ and } q_j^+ = \lim_{w\downarrow 0} E[Z^{2j}\varepsilon_{U_i}^2|W=w].$$

Proof. *Consider the generic term, for* $l, j = 0, 1$

$$\frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^+\varepsilon_{U_i},$$

and its variance, which equals

$$\begin{aligned}(nh_n)^{-1}h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^{2l} Z_i^{2j}k_{ih_n}^{+2}\varepsilon_{U_i}^2\right] &= (nh_n)^{-1}\int_0^\infty u^{2l}k^2(u)q_j(uh_n)f(uh_n)du \\ &= (nh_n)^{-1}f(0^+)q_j^+v_{2l} + o(1)\end{aligned}$$

where $q_j(w) = E[Z^{2j}\varepsilon_{U_i}^2|W=w]$. ■

Similarly, we have the following result, which proof is the same as in the previous lemma.

Lemma B.8

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{-in}k_{ih_n}\varepsilon_{U_i}\right) = \frac{1}{nh_n}\Sigma_{U^-} + o(1),$$

where

$$\Sigma_{U^-} = f(0^-)\begin{bmatrix} v_0 & v_0q_1^- & v_2 & v_2q_1^- \\ v_0q_1^- & v_0q_2^- & v_2q_1^- & v_2q_2^- \\ v_2 & v_2q_1^- & v_4 & v_4q_1^- \\ v_2q_1^- & v_2q_2^- & v_4q_1^- & v_4q_2^- \end{bmatrix},$$

$$q_j^- = \lim_{w\uparrow 0} E[Z^{2j}\varepsilon_{U_i}^2|W=w].$$

Define

$$\Sigma_U = \begin{bmatrix} \Sigma_{U^+} & 0 \\ 0 & \Sigma_{U^-} \end{bmatrix}$$

Lemma B.9 (Numerator: Conditional CLT)

$$(nh_n)^{-1/2}\sum_{i=1}^{n} S_{in}k_{ih_n}\varepsilon_{U_i} \to_d N\left(0,\Sigma_U\right).$$

Proof. *Consider a generic term for* $l,j=0,1$

$$\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^{+}\varepsilon_{U_i}.$$

We apply Lyapounov with third absolute moment. By the lemma on the asymptotic variance, we need to establish

$$(nh_n)^{-1/2}\,h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^{3l} Z_i^{3j}k_{ih_n}^{+3}\varepsilon_{U_i}^3\right]=o(1).$$

But note that, defining $s_j(w)=E[Z^{3j}\varepsilon_{U_i}^3|W=w]$,

$$\begin{aligned}h_n^{-1}E\left[\left(\frac{W_i}{h_n}\right)^{3l} Z_i^{3j}k_{ih_n}^{+3}\varepsilon_{U_i}^3\right] &= \int_0^\infty u^{3l}k^3(u)s_j(uh_n)f(uh_n)du\\ &= O(1).\end{aligned}$$

The same holds for the left limit part. ■

Lemma B.10 (Numerator: Unconditional CLT)

$$(nh_n)^{-1/2}\sum_{i=1}^{n} S_{in}U_i k_{ih_n} - \frac{(nh_n)^{1/2}h_n^2}{2}f(0)b_U \to_d N\left(0,\Sigma_U\right).$$

Proof. *It follows from previous Lemmas.* ■

Lemma B.11 (Main CLT)

$$\sqrt{nh_n}(\hat{\theta}_n-\theta_n)\to_d N\left(0,\Omega\right),$$

where

$$\Omega=\left(\Delta\Gamma^{-1}\Delta'\right)^{-1}\Delta\Gamma^{-1}\Sigma_U\Gamma^{-1}\Delta'\left(\Delta\Gamma^{-1}\Delta'\right)^{-1}.$$

Proof. *It follows from previous Lemmas.* ■

References

Brollo, F., T. Nannicini, R. Perotti and G. Tabellini (2013): "The Political Resource Curse", *American Economic Review*, 103(5): 1759-96.

Angrist, J.D. and Lavy, V. (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement.", *The Quarterly Journal of Economics*, 114(2), pp.533-575.

Caetano, C., and Escanciano, J. C. (2017): "Identifying Multiple Marginal Effects with a Single Instrument", working paper.

Calonico, S., Cattaneo, M.D., Farrell, M.H and Titiunik, R. (2016): "Regression-Discontinuity Designs using Covariates", working paper.

Calonico, S., Cattaneo, M.D. and Titiunik, R. (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs", *Econometrica*, 82, 2295-2326.

Card, David and Dobkin, Carlos and Maestas, Nicole (2008): "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare," *The American Economic Review*, vol. 98(5), 2242-2258.

Carpenter, C. and Dobkin, C. (2009): "The Effect of Alcohol Access on Consumption and Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age", *American Economic Journal: Applied Economics*, Vol. 1, Issue 1, pp. 164-82.

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

Chay, K., and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113(2), 376–424.

DiNardo, J., and D. S. Lee. (2011): "Program Evaluation and Research Designs," In Handbook of Labor Economics, ed. O. Ashenfelter and D. Card, vol. 4A, 463-536. Elsevier Science B.V.

Fan, J., and Gijbels, I. (1996): *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

Feir, D., T. Lemieux, and V. Marmer (2016): "Weak Identification in Fuzzy Regression Discontinuity Designs." *Journal of Business & Economic Statistics*, 34(2), 185–196.

Frolich, M. (2007): "Regression Discontinuity Design with Covariates," IZA Discussion Paper N. 3024..

Hahn, J., Todd, P., and van der Klaauw, W. (1999): "Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design." National Bureau of Economic Research Working Paper 7131.

Hahn, J., Todd, P., and van der Klaauw, W. (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–09.

Imbens, G. W. and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–35.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature* 48, 281-355.

Ludwig, Jens, and Douglas L. Miller (2007): "Does Head Start improve children's life chances? Evidence from a regression discontinuity design," *The Quarterly journal of economics* 122.1, 159-208.