

Essays on Estimation of a Regression Jump: A Generalized Reflection Approach

by

SIHONG XIE

B.S., Nanjing University of Aeronautics and Astronautics, 1988

M.S., Nanjing University of Aeronautics and Astronautics, 1991

M.B.A., Colorado State University, 2011

M.S., University of Colorado, 2014

A thesis submitted to the  
Faculty of the Graduate School of the  
University of Colorado in partial fulfillment  
of the requirement for the degree of Doctor of Philosophy  
Department of Economics

2019

This thesis entitled:  
Essays on Estimation of a Regression Jump: A Generalized Reflection Approach  
written by Sihong Xie  
has been approved for the Department of Economics

---

Carlos Martins-Filho(Chair)

---

Xiaodong Liu

---

Keith Maskus

---

Adam McCloskey

---

Eric Vance

Date \_\_\_\_\_

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Sihong, Xie (Ph.D., Economics)

Essays on Estimation of a Regression Jump: A Generalized Reflection Approach

Thesis directed by Professor Carlos Martins-Filho

## **Abstract**

Regression Discontinuity (RD) designs are popular models in economics used by researchers to evaluate the effects of policy interventions. In the past two decades, a great number of papers on RD applications and methodology have been published in leading economic journals. However, research on RD estimators, which is fundamental to RD models, has been few and far between. The main estimation approach is to apply local linear (LL) or local polynomial estimators on both sides of the known discontinuity point and then to estimate the jump. Most developments have focused on amendments to and improvements of LL, but there are almost no competitive alternatives for LL estimators. This dissertation adopts a novel approach by providing a completely new class of RD estimators taking a generalized reflection approach by using the extension of Hestenes (1941). My estimators have simple analytical representations, desirable asymptotic properties, and are computationally easy to implement. Having boundary properties that are as good as LL estimators and performing better than LL estimators in finite samples, my estimators offer a competitive alternative for LL estimators in RD models.

In Chapter 1, I review major theoretical developments in RD design in the econometrics literature, focusing on estimators for regression discontinuity. In Chapter 2, I introduce my Hestenes-based RD estimators. Focusing on properties at boundary points, I provide results on the bias, variance and asymptotic distribution of my estimators. I compare the finite sample properties of my estimators with popular regression estimators – the Nadaraya-Watson and LL estimators – using Monte Carlo studies, empirical examples, and empirically motivated simulations. Chapter 3 extends the estimation of univariate regression with a discontinuity to multivariate regression settings. I consider an additive model and propose four two-stage estimators: at the first stage, I use a marginal integration, instrument variable, backfitting, or B-splines estimator for the continuous components of the regression; at the second stage, I use the Hestenes estimator developed in Chapter 2 to estimate the jump discontinuity. Monte Carlo studies show my estimators outperform the local linear RD estimators in an additive linear model that are commonly used in empirical research.

## **Acknowledgments**

I am very grateful to my advisor, Carlos Martins-Filho. His brilliant work has inspired me and he helps me to produce work that I can be proud of.

I would like to thank the other members of the committee, Xiaodong Liu, Keith Maskus, and Adam McCloskey, for their careful reading of this manuscript and helpful suggestions. I also want to thank William Ridley for his diligent proofreading of many parts of this manuscript. Last, but not least, I would like to thank my cohorts and office mates, whose support and encouragement has created a wonderful experience for my graduate studies at CU.

## Contents

### Chapter

|            |  |     |
|------------|--|-----|
| <b>1.</b>  | <b>A Literature Review of Regression Discontinuity Designs (RDD)</b> .....                                 | 1   |
| 1.1        | Estimators for Treatment Effects .....   | 4   |
| 1.2        | Including Additional Covariates .....  | 9   |
| 1.3        | Other Theoretical Developments .....   | 10  |
| 1.4        | Extensions of RDD .....  | 15  |
| 1.5        | Conclusion .....   | 19  |
| <b>2.</b>  | <b>Estimation of a Regression Jump: a Generalized Reflection Approach</b> ..                               | 20  |
| 2.1        | A General Reflection Approach: Hestenes-based Regression Estimator ..                                      | 24  |
| 2.2        | Estimators for a Jump Discontinuity .....  | 53  |
| 2.3        | Monte Carlo Studies .....  | 59  |
| 2.4        | Empirical Illustration .....   | 76  |
| 2.5        | Empirically Motivated Simulations .....  | 87  |
| 2.6        | Conclusions and Future Studies.....  | 91  |
| <b>3.</b>  | <b>Estimation of a Regression Jump with Additional Covariates in a Non-parametric Additive Model</b> ..... | 93  |
| 3.1        | Additional covariates in RDD .....   | 95  |
| 3.2        | Condition for Identification of the Model.....   | 97  |
| 3.3        | Estimators Under the Study .....   | 97  |
| 3.4        | Simulation .....   | 107 |
| 3.5        | Conclusions .....  | 112 |
| References | .....  | 114 |
| Appendix   |  |     |
| <b>A.</b>  | <b>Supporting Lemmas</b> .....   | 119 |
| <b>B.</b>  | <b>Matlab Codes for Simulations</b> .....  | 128 |
| B.1        | Regression Estimators .....  | 128 |
| B.2        | Density Estimators .....   | 132 |
| B.3        | Estimators for the Additive Model .....  | 138 |

Tables

**Chapter 2**

|      |   |    |
|------|---|----|
| 2.1  | Regression Models . . . . .                                       | 62 |
| 2.2  | Simulation Results for the Default Setting . . . . .              | 65 |
| 2.3  | Varying Sample Size . . . . .                                     | 66 |
| 2.4  | Different Density Function of the Regressor $X$ . . . . .         | 68 |
| 2.5  | Select Bandwidth by a Plug-in Method . . . . .                    | 70 |
| 2.6  | Select Bandwidth by a Jackknife Cross-Validation Method . . . . . | 71 |
| 2.7  | Choose a Epanechnikov Kernel . . . . .                            | 73 |
| 2.8  | Choose a Triangular Kernel . . . . .                              | 74 |
| 2.9  | Choose a Rectangular Kernel . . . . .                             | 75 |
| 2.10 | Univariate Regression Without Additional Covariates . . . . .     | 84 |
| 2.11 | Multivariate Regression With Additional Covariates . . . . .      | 85 |
| 2.12 | Density Estimates on Population . . . . .                         | 87 |
| 2.13 | Results from the Empirical Motivated Simulation . . . . .         | 90 |

**Chapter 3**

|     |                              |     |
|-----|------------------------------|-----|
| 3.1 | Simulation Results . . . . . | 111 |
|-----|------------------------------|-----|

**Appendix A**

## Figures

|                  |   |     |
|------------------|---|-----|
| 1.1              | The RDD Model .....   | 2   |
| <b>Chapter 2</b> |   |     |
| 2.1              | Kernel functions in NW and Hestenes Estimators .....                      | 25  |
| 2.2              | Coefficients of Leading Terms of Asymptotic Distribution .....            | 57  |
| 2.3              | Regression Functions Estimated by Three Estimators .....                  | 63  |
| 2.4              | Impacts on Total Spending, Other Revenue, and Own Revenue .....           | 81  |
| 2.5              | Impacts on Schooling, Literacy, Poverty, and Party Reelection .....       | 82  |
| 2.6              | Population Density Estimated by LL Estimator and Hestenes Estimator ..... | 86  |
| 2.7              | The Histogram of Original Samples and the Generated Samples .....         | 88  |
| 2.8              | The Estimated Jump in the Empirically Motivated Simulation .....          | 89  |
| <b>Chapter 3</b> |   |     |
| 3.1              | Regression Components of the True Model .....                             | 109 |

## Chapter 1

### A Literature Review of Regression Discontinuity Designs (RDD)

This dissertation studies regression discontinuity models, of which my main point of interest is the estimation of a jump discontinuity in a nonparametric regression. In Chapter 2, I propose a class of reflection estimators based on [Hestenes' \(1941\)](#) extension. In chapter 3, I extend the estimation of univariate regression with a discontinuity to multivariate regression with a discontinuity. I start with a basic model.

Consider a nonparametric regression model where, without loss of generality, there is a known discontinuity at  $x = 0$ . For a regressor  $X \in \mathbb{R}$ ,

$$Y_i = m(X_i) + \epsilon_i,$$

where  $E(\epsilon_i|X_i) = 0$ ,  $V(\epsilon_i|X_i) = \sigma^2$ , and

$$m(x) = \begin{cases} m^+(x) & \text{if } x \geq 0 \\ m^-(x) & \text{if } x < 0 \end{cases}.$$

The jump discontinuity of the conditional expectation is given by

$$\alpha = \lim_{x \downarrow 0} m^+(x) - \lim_{x \uparrow 0} m^-(x).$$

One area where the estimation of  $\alpha$  is of interest in economics is regression discontinuity designs. [Thistlethwaite and Campbell \(1960\)](#) first developed RDD with an application in educational psychology to verify if students who earn National Merit Scholarships were more likely to enroll in graduate study. In their seminal paper, [Hahn et al. \(2001\)](#) establish the identification of RDD: assuming that the regression function of the potential outcome is continuous at the threshold, the jump discontinuity of the regression function at the threshold is the average treatment effect (ATE).



**Figure 1.1:** The RDD Model

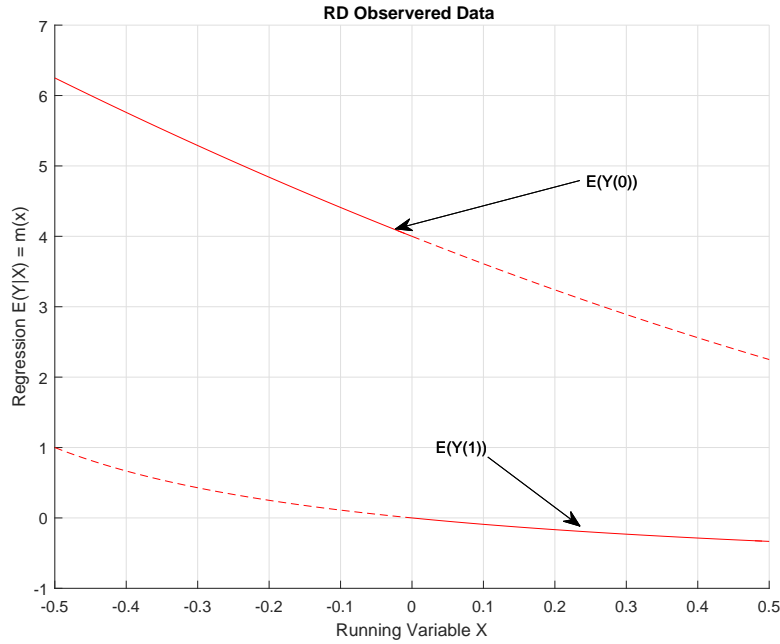


Figure 1.1 describes the basic elements of a RDD model. Let  $Y(D)$  be potential outcome variables. The binary variable  $D \in \{0, 1\}$  stands for some kind of treatment.  $Y(1)$  is the potential outcome when the individual is treated;  $Y(0)$  is the potential outcome when the individual is untreated. The running variable  $X$  determines if an individual receives the treatment. When  $X \in \mathbb{R}$  is above a threshold  $c \in \mathbb{R}$ , the individual receives the treatment and  $D = 1$ , otherwise, the individual does not receive the treatment and  $D = 0$ . Without loss of generality, I take  $c = 0$ .  $D = 1[X \geq 0]$ . The observed outcome is

$$Y = DY(1) + (1 - D)Y(0).$$

The framework is used to identify and conduct inference on ATE,  $J(0) \equiv E[Y(1) - Y(0)|D = 0]$ . Under the assumption that regression functions of the two potential outcomes  $Y(0)$  and  $Y(1)$  are continuous at the threshold,  $J(0) = \lim_{x \downarrow 0} E[Y|X = x] - \lim_{x \uparrow 0} E[Y|X = x]$ .

In general, the ATE can be denoted as

$$J(0) = \frac{\lim_{x \downarrow 0} m^+(x) - \lim_{x \uparrow 0} m^-(x)}{\lim_{x \downarrow 0} E(D|X = x) - \lim_{x \uparrow 0} E(D|X = x)}. \quad (1.1)$$

Trochim (1984) divided RDD into sharp designs and fuzzy designs depending on whether or not the treatment variable is a deterministic function of the running variable. For sharp RDD, since  $\lim_{x \downarrow 0} E(D|X = x) - \lim_{x \uparrow 0} E(D|X = x) = 1$ ,

$$J(0) = \lim_{x \downarrow 0} m^+(x) - \lim_{x \uparrow 0} m^-(x) = \alpha.$$

It is apparent that the jump discontinuity  $\alpha$  in this general statistic model is the ATE  $J(0)$  in a sharp RDD. In other words,  $\alpha$  corresponds to the jump in the conditional expectation as the running variable  $X$  approaches the threshold for an individual to receive the treatment.

To estimate  $\alpha$ , I use observations to the left of the threshold to estimate regression on the left and to use observations to the right to estimate regression on the right and then difference the two estimates at the threshold. The challenge comes from the discontinuity. Most of the known properties of conventional nonparametric estimators, such as the Nadaraya-Watson (NW) or local linear estimators (LL), are derived for points at which the regression is continuous and typically sufficiently smooth. When considering a discontinuity, traditional kernel regression estimators such as NW estimators have undesirable boundary properties. Specifically, the bias convergence rate of the estimator at boundary points is slower than that at interior points. The properties of these estimators at boundary points, a vicinity of the discontinuity point defined by the bandwidth, are affected by the discontinuity, while at interior points – which are away from the discontinuity point and where the regression is smooth – the estimators’ properties are not affected by the discontinuity. Since the jump is the object of interest, I am concerned with the properties of estimators at points in the vicinity of the discontinuity point – that is, the properties of estimators at the boundary.

Since boundary problems in RDD are a special case of the model that I am considering, it is important to first present a review of the RDD literature. In this chapter, I start by discussing the literature that deals directly with the estimation of the ATE in RDD. I then review major theoretical

developments in RDD over the last 20 years. In Section 1.1, I discuss the estimators of the ATE in a univariate regression setup as a prelude to the results in the second chapter. In Section 1.2, I expand the model to include additional covariates other than the running variable, which motivates the multivariate analysis in the third chapter. In Section 1.3, I cover other major developments in RDD. In Section 1.4, I cover some significant RDD extensions. Section 1.5 concludes with how my dissertation relates to this growing body of literature on RDD.

## 1.1 Estimators for Treatment Effects

RDD are popular models for the statistical evaluation of policy interventions because they do not impose a specific functional form on the regression, and further allow for endogeneity between the outcome variable and the running variable. [Lee and Lemieux \(2010\)](#) explain that as long as individuals do not have exact control over the running variable, their manipulations of the running variable (including self-sorting) do not affect the identification of the treatment effect. The work of [Hahn et al. \(2001\)](#) has arguably inspired the flourishing use of RDD in economics – thousands of empirical papers have been written employing the approach. At the same time, theoretical developments in RDD in the econometrics literature have branched out in many different directions. In 2008, the Journal of Econometrics dedicated a special issue to regression discontinuity, in which [Cook \(2008\)](#) gives a historical introduction of RDD in the fields of psychology, statistics, and economics and [Imbens and Lemieux \(2008\)](#) give practical guidance for implementing RDD in empirical research. Other works by [Lee and Lemieux \(2010\)](#), [van der Klaauw \(2008\)](#) and [Cattaneo and Escanciano \(2017\)](#) provide theoretical and empirical overviews of the existing literature. These papers helped promote the broad adoption of RDD in economic analysis.

The main objective of RDD is the estimation of treatment effects. Since RDD does not specify a functional form for the regression function, the estimators of the treatment effect are, in essence, nonparametric estimators. The boundary problems that confront traditional kernel regression estimators, such as NW estimators, are well-known. Solutions have been proposed. In practice, it is very common for people to use local linear or local polynomial on both sides of the point of discontinuity and then to estimate the jump. This is because [Fan \(1992\)](#) shows that the local

linear estimators have desirable boundary properties. In their seminal paper, [Hahn et al. \(2001\)](#) recommend the LL estimators for RDD. In an influential review paper, [Lee and Lemieux \(2010\)](#) give guidance on estimating RDD using the LL estimators.

Compared to the fast growing body of literature on applications and adaptations of the LL estimators to fit various situations in RDD, the alternative RD estimators to the LL estimators are rare. But there are a few exceptions. [Porter \(2003\)](#) propose partial linear estimators from and [Imbens and Wager \(2019\)](#) propose finite-sample-minimax linear estimators. In this section, I discuss these RD estimators as well as local linear and local polynomial estimators in the basic setup, where there is a known discontinuity point, and both the running variable and outcome variable are continuous.

### 1.1.1 Local Linear and Local Polynomial Estimators

Let  $\{X_i, Y_i\}_{i=1}^n$  be a sample in which observations to the left and right of the cutoff point are IID. An estimator for the discontinuity jump size  $\alpha$  is defined as

$$\hat{\alpha} = \hat{m}^+(0) - \hat{m}^-(0), \tag{1.2}$$

where  $\hat{m}^+(0) = \lim_{x \downarrow 0} \hat{m}^+(x)$  and  $\hat{m}^-(0) = \lim_{x \uparrow 0} \hat{m}^-(x)$ , and  $\hat{m}^+$  and  $\hat{m}^-$  are estimators for  $m^+$  and  $m^-$ , and each  $X_i$  has the same marginal density function  $f$  across the cutoff point in  $\mathbb{R}$ .

A local polynomial of order  $p$  estimator  $\hat{m}_p(x)$  of  $m(x)$  is given by  $\hat{m}_p^+(x) = \hat{a}$ , whereas

$$\begin{pmatrix} \hat{a} \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix} = \underset{a, b_1, \dots, b_p, i=1}{\operatorname{argmin}} \sum^n k_h(X_i - x) D_i [Y_i - a - b_1(X_i - x) - \dots - b_p(X_i - x)^p]^2,$$

where  $D_i = I(X_i \geq 0)$ ,  $I(A)$  is the indicator function of event  $A$ ,  $k_h(\cdot) = \frac{1}{h}k(\frac{\cdot}{h})$ ,  $h$  is a bandwidth, and  $k(\cdot)$  is a kernel function. Similarly, I define  $\hat{m}_p^-(x)$  by replacing  $D_i$  with  $1 - D_i$ .

When  $p = 0$ , the estimators are Nadaraya-Watson estimators. They can be used to estimate

the regression to the left and right of the threshold, and then to estimate the jump. Nevertheless, NW estimators have boundary problems. Simulations show that in finite samples, NW estimators have larger biases at boundary points than at interior points. In practice, they are rarely used.

When  $p = 1$ , the estimators are local linear (LL) estimators. LL estimators are commonly used to estimate regressions in domains with boundaries. This is because [Fan \(1992\)](#) shows that LL estimators have desirable boundary properties, in that LL estimators have the same order of bias and variance at boundary points as those at interior points. [Hahn et al. \(2001\)](#) recognize the problem with NW estimators and thus recommend the use of LL estimators in RDD. [Imbens and Lemieux \(2008\)](#) and [Lee and Lemieux \(2010\)](#) give guidance on their implementation in empirical RDD research.

[Porter \(2003\)](#) proposes local polynomial estimators, where  $p > 1$ , that can achieve biases of smaller order at boundary points by using a high-order polynomial. However, [Gelman and Imbens \(2014\)](#) warn against the use of high-order polynomials because regressors are more likely to have collinearity problem and they assign undue influence to outliers, more importantly, points far away from the discontinuity point. In practice, first and second order local polynomials (where  $p = 1$  or  $p = 2$ ) are the most commonly used approaches. Methodologically, the bias reduction is achieved by the polynomial “correction.” That is, local polynomial estimation directly accounts for the shape of the regression when estimating its level.

### 1.1.2 Partially Linear Estimators

In a regular regression setting where the regression is continuous, there is a simple link between the bias order and the smoothness of the density function of the regressor and the regression, but this link is broken at the boundary or the point of discontinuity. The following method strives to recover this relationship so the bias reduction can be achieved by higher-order bias-reducing kernels.

[Porter \(2003\)](#) proposes a partial linear estimator. Consider the model

$$Y_i = m(X_i) + D_i\alpha + \epsilon_i \quad \text{for } i = 1, 2, \dots, n,$$

where  $\alpha$ , the jump at the threshold, is the parameter of interest. Then,

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - \alpha D_i - \sum_{j=1}^n w_j^i (Y_j - \alpha D_j)]^2,$$

where  $w_j^i = \frac{k_h(X_i - X_j)}{\sum_{l=1}^n k_h(X_i - X_l)}$ .

Partially linear estimators can achieve biases of a smaller order at boundary points but they require the regression functions to be identical (separated by a jump) on both sides of the threshold, a restriction that is not required by typical nonparametric estimators when applied to data to the left and right sides of a point of discontinuity.

### 1.1.3 Finite-Sample-Minimax Linear Estimator

[Imbens and Wager \(2019\)](#) point out that local linear regression estimators (LL) are not the optimal linear estimators in terms of minimax mean-squared error. They propose a linear estimator for treatment effects that is a numeric solution to a convex optimization problem, subject to bounds on the second derivative of the regression function. Both LL estimators and finite-sample-minimax linear estimators are linear estimators, in that the estimator is a weighted average of the observations of the dependent variable and the weights are determined by the distance between the sample of the regressor and the evaluation point. Unlike LL estimators, which are obtained by solving weighted least square problems and they have closed-form solutions, finite-sample-minimax linear estimators are obtained by using numerical convex optimization tools and they do not have analytic representations. The advantages of their estimators are that they can be seamlessly extended to the case of discrete running variables as well as multiple running variables, and outperform LL estimators in terms of minimax mean-squared error.

Define the treatment effect as  $\tau(c) = \mu_1(c) - \mu_0(c)$ , where  $c$  is the threshold.  $\mu_w$  is the conditional expectation of potential outcomes,  $\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x]$ , where  $w$  corresponds to the treatment,  $w \in \{0, 1\}$ . Subsequently, a linear treatment estimator is defined as  $\hat{\tau} = \sum_{i=1}^n \hat{\gamma}_i Y_i$ .

The weights  $\hat{\gamma}$  are obtained by solving the follow minimax problem:

$$\hat{\gamma} = \operatorname{argmin}_{\gamma} \left\{ \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + I_B^2(\gamma) \right\}.$$

The first term is the variance  $\sum_{i=1}^n \gamma_i^2 \sigma_i^2$  where  $\sigma_i^2 = \operatorname{Var}[Y_i|X_i]$ . The second term is the worst case bias over the class of functions with bounded second derivatives:

$$I_B(\gamma) := \sup_{\mu_0(\cdot), \mu_1(\cdot)} \left\{ \sum_{i=1}^n \gamma_i \mu_{w_i}(X_i) - (\mu_1(c) - \mu_0(c)) : |\mu_w''(x)| \leq B \text{ for all } w, x \right\},$$

where  $\mu_0$  and  $\mu_1$  belong to the class of functions with second derivative bounded by  $B$ . This method does not require choosing a kernel or bandwidth. Instead, it requires choosing  $\sigma_i^2$  and  $B$ . The authors show that under certain conditions,  $\hat{\tau}$  is asymptotically normally distributed. The results can be used to construct confidence intervals.

#### 1.1.4 Generalized Reflection Estimators

LL estimators have desirable adaptive boundary properties. They can automatically adjust to the boundary, but they do not address the boundary problem head-on. One way to look at this is that LL estimators adapt to the boundary by adjusting a regular kernel to an effective kernel that contains partially integrated kernels, so the bias at the boundary has the same order of decay as the bias at interior points. However, it does not solve the boundary problem: the bias expression at boundary points is different from that at interior points.

In Chapter 2, I propose a new class of estimators for the size of a jump using Hestenes' extension. [Hestenes \(1941\)](#) shows that functions that are sufficiently smooth and are defined on a subset of the real line can be extended to the entire real line, inheriting the smoothness of the original function. Hestenes provides a simple method to obtain the extensions. I construct regression estimators defined on bounded domains by reflecting the observations to the other side of the boundary multiple times with scaling parameters obtained through Hestenes' extension method. The number of times reflection is performed is decided by the assumed smoothness of the regression function of the regressand and the density function of the regressor. In contrast to LL estimators,

my estimators have the bias expression at the boundary points that are the same as that at the interior points. Furthermore, by assuming the same order of smoothness of the regression function, my estimators have the same bias and variance order as LL estimators. [Schuster \(1985\)](#) proposes a reflection estimator for density estimation, where the reflection is made once. In this context, my method can be thought of as a generalized reflection approach. Hence, my estimators can take advantage of high orders of smoothness of regression and density functions and achieve better (or faster) rates of convergence for the bias.

## 1.2 Including Additional Covariates

Researchers in empirical RDD often include covariates other than the running variable in the regression to guard against the misspecification of the model and to increase the precision of the estimates, yet theoretic works on the properties of these estimators are rare. One exception is a paper by [Calonico et al., 2018](#), which studies the properties of RDD estimators in an additive separable linear model.

[Calonico et al., 2018](#) propose an additive separable linear model that aligns with the models typically employed in empirical RDD studies:

$$Y = m_1(X_1) + Z_2\beta_2 + \dots + Z_D\beta_D + \epsilon,$$

where  $m_1(x_1)$  is discontinuous. The estimation of  $m_1$  does not suffer from the problem known as the curse of dimensionality in which the optimal convergence rate slows down geometrically as the dimension  $D$  increases. This is achieved because it is an additive model. Besides, only the regression component of the running variable is estimated locally while the regression components of the other covariates are estimated globally. However, the model is not flexible enough because covariates enter in a linear-in-parameters form.

In Chapter 3, I extend the estimation of a univariate regression with discontinuities in Chapter 2 to a multivariate regression with discontinuities. I consider a nonparametric additive model where



one of its components is discontinuous whereas all others are continuous, i.e.,

$$Y = m_1(X_1) + m_2(Z_2) + \cdots + m_D(Z_D) + \epsilon. \quad (1.3)$$

My estimation strategy is a two-step procedure in which I first estimate the components of the regression associated with the continuous covariates, then use the Hestenes-based estimators I have developed to estimate the jump discontinuity of the regression associated with the running variable in the second stage. The first stage estimators are constructed using different methods such as marginal integration, instrument variable, backfitting, and B-splines estimators. I conduct numerical simulations to compare the finite sample performance of my two-stage multivariate RD estimators, with univariate RD estimators that do not account for additional covariates, and with [Calonico et al., 2018](#)'s multivariate RD estimators in which additional covariates enter in linear forms to show that my Hestenes RD estimators are well suited for a flexible additive model that accounts for additional covariates.

### 1.3 Other Theoretical Developments

Commonly used treatment effect estimators are based on kernel estimation and thus require the selection of bandwidths and kernels. A number of researches have been conducted on selecting optimal bandwidths and kernels as well as making correct inference. In addition, researchers have been concerned with the discontinuity of the density function of the running variable and heterogeneous treatment effects. Last, but not the least, the methods of optimally plotting sample data are explored.

#### 1.3.1 Optimal Bandwidth

A critical aspect of nonparametric estimation is bandwidth selection. Since I am only interested in the estimation of the jump at the discontinuity point, a locally optimal bandwidth suits most purposes better than a globally optimal bandwidth. [Imbens and Kalyanaraman \(2012\)](#) provides an optimal data-adaptive bandwidth for LL estimators to minimize the mean-squared error of the

regression estimator. Improving on their work, [Arai and Ichimura \(2018\)](#) achieve smaller mean-squared error by choosing different bandwidths for the regression estimators to the right and to the left of the discontinuity point.

### 1.3.2 Optimal Kernel

The kernel can also make a difference in nonparametric estimation. [Cheng et al. \(1997\)](#) show that the triangular kernel is the minimax optimal boundary kernel; over all regression functions of the same smoothness and among all possible choices of kernels for linear estimators, the triangular kernel minimizes the asymptotic mean-squared error. [Gao \(2018\)](#) characterizes the asymptotically optimal kernel for the minimax linear RD estimators under the restriction that the first-order derivative of the regression function is Lipschitz continuous.

### 1.3.3 Optimal Inference

Testing the null hypothesis that the treatment effect  $\alpha = 0$  is critically important. However, the optimal bandwidth obtained by minimizing mean squared error, such as the method by [Imbens and Kalyanaraman \(2012\)](#), often leads to a non-negligible bias. Namely, the optimal bandwidth converges to zero with the order  $h_{MSE} \propto n^{-1/5}$ . For the leading terms of bias to vanish so that the estimator has an asymptotically normal distribution, the condition  $nh^5 \rightarrow 0$  together with  $nh \rightarrow \infty$ , implies  $h_{norm} \propto n^\delta$  where  $-1 < \delta < -1/5$ . This is a well known problem with kernel estimation. As a result, it leads to biased confidence intervals and low coverage. To correct the problem, [Calonico et al. \(2014\)](#) propose bias-corrected RD estimators. Together with a standard error estimator, they construct confidence intervals that are robust to bandwidth choice and make correct inferences of the treatment effects in sharp RD, fuzzy RD, sharp kink RD, and fuzzy kink RDD (more explanation about kink RD designs is provided in Section 1.4.1).

In contrast with the above bias-corrected method, [Armstrong and Kolesár \(2018\)](#) base their method of obtaining optimal confidence intervals on the characteristics of the bias – i.e., the smoothness of the regression function – rather than undersmoothing. Assuming the regression function satisfies a Lipschitz condition of a certain order, they construct an optimal confidence interval for

LL estimators by directly solving an optimization problem that maximizes their minimum test power under the alternative hypothesis. Moreover, [Armstrong and Kolesár \(2017\)](#) provide a simple adjustment to the critical value commonly used to construct confidence intervals for local polynomial estimators to achieve correct coverage for finite samples. Finally, [Kolesár and Rothe \(2018\)](#) construct robust confidence intervals for LL estimators and achieve uniform asymptotic coverage for discrete running variables.

Researchers often find that the asymptotic properties of estimators at the boundary are complex and the results difficult to use – requiring estimation of conditional variance, covariance, and the density of the running variable. [Otsu et al. \(2015\)](#) circumvent the estimation of asymptotic variances of LL estimators by employing empirical likelihood ratios for the treatment effects estimated by LL estimators, so they can use empirical likelihood methods to construct confidence intervals for inference in sharp and fuzzy RDD.

Another strain of the literature on robust inference relates to randomization. One way to understand and justify the identification of RDD is to look at the treatment assignment at the threshold as a local randomized trial. [Cattaneo et al. \(2015\)](#) argue that the conventional assumption of the continuity of the regression function of potential outcome variables cannot justify randomization. They formally define the condition that justifies the randomized trial interpretation and they provide a finite-sample inference procedure. Their inference is based on the limited number of observations around the threshold rather than the standard large-sample procedure. [Li et al. \(2015\)](#) further develop the idea of local randomization by defining a hypothetical experiment and assignment mechanism: there exists a subgroup around the threshold whose treatment status is assigned by chance, i.e. the local overlap assumption holds. The authors develop a Bayesian hierarchical modeling approach to select the target subgroups.

In RDD, the running variable can be seen as an instrument variable (IV) ([Hahn et al. \(2001\)](#)), so it is not surprising that anything that can impair IV estimation or inference, such as a weak IV, can negatively affect the estimation in RDD as well. [Feir et al. \(2016\)](#) propose modified t-tests for the weak-IV robust inference in fuzzy RD designs.

As RDD are frequently used for evaluating the effects of policy interventions, testing the null

hypothesis that the treatment effect is zero is important. [Kamat \(2018\)](#) points out that under the current regularity assumptions, the test permits a large set of possible distributions that the distributions of the null are too close to the distributions of the alternatives to distinguish them, which results in the power of the test is bounded by the size of the test. This implies one can't construct valid finite sample tests and uniformly asymptotically tests. The test is only pointwise asymptotically valid but not uniformly asymptotically valid. [Kamat \(2018\)](#) reminds empirical researchers of the need to strengthen the current regularity assumption so that the test can be uniformly asymptotically valid.

### 1.3.4 Tests to Guard Against Identification Failure

Since RDD exploits the jump in the regression of the treatment on the running variable, a discontinuity in the density function of the running variable could cause the failure of identification. One source from which a discontinuity might arise could be that an individual endogenously manipulates the running variable (to be complete, manipulation could also cause a kink or fluctuation, but not necessarily a discontinuity, in the density function such as heaping in the vicinity of the threshold). [McCrary \(2008\)](#) advocates testing for the discontinuity of the density function of the running variable at the threshold. The jump density estimators are based on [Cheng's \(1994\)](#) local linear binning density estimators. Recent work from [Mynbaev and Martins-Filho \(2018\)](#) provides a simple and elegant solution to estimate the density at the boundary. Using a simple mathematical result from [Hestenes \(1941\)](#) to construct the density estimator, the approach generalizes [Schuster's \(1985\)](#) reflection method.

[Bugni and Canay \(2018\)](#) provide a test for the continuity of the density function without imposing the smoothness requirement as most other estimators do – their approach only requires the continuity of the density function. A feature of their paper is that rather than relying on the asymptotic property of the estimator, they derive the properties of their estimators for a fixed sample size. This is close to the reality in the RDD setup; the number of observations in the vicinity of the threshold does not necessarily grow as the sample size increases.

Otsu et al. (2013) share the same concern: the low density of the running variable around the threshold in RDD. Their solution is to use local likelihood estimators rather than local linear estimators. In addition, they construct empirical local likelihood tests and confidence intervals for inference, which avoids the estimation of the asymptotic variance.

### 1.3.5 Treatment Effect Heterogeneity

When the assumption of a constant treatment effect is not appropriate, heterogeneous treatment effects must be considered. Hahn et al. (2001) show that in fuzzy designs, the ratio  $J(0)$  in equation (1.1) is the local average treatment effect (LATE) at the threshold of a subgroup of individuals, namely, policy compliers (individuals that take the treatment when their treatment assignment status  $D_i = 1$  and do not take the treatment otherwise). Critics often complain that RDD possess a high degree of internal validity but are lacking in external validity. To convince an audience of the external validity of findings generated by an RDD (namely, the LATE), the LATE must be extrapolated to the points away from the threshold, and to other subgroups: the always-takers who take the treatment regardless of treatment assignment and never-takers who never take the treatment regardless of treatment assignment.

First, consider the extrapolation of the treatment effect to other subgroups. Taking advantage of fuzzy design, Bertanha and Imbens (2014) argue that to assess the plausibility of generalizing the LATE to other sub-populations, researchers should test whether the distribution of treatment effects for the always-takers is the same as that for the treated compliers. If they are, it is plausible to assume never-takers (who are not observed) will have the same treatment effect, so that the generalization of the treatment effect from the compliers to the population is plausible.

Second, consider the extrapolation of the treatment effect away from the threshold. Dong and Lewbel (2015) ask how the treatment effect would change if the threshold changed. To verify external validity, they suggest researchers estimate the first derivative of the regression function with respect to the running variable. If this treatment effect derivative is small (close to zero), the extrapolation of the treatment effect for individuals with the value of the running variable away from the threshold is credible; otherwise, this extrapolation is questionable. In the context of a sharp

design, taking advantage of the availability of additional covariates other than the running variable, [Angrist and Rokkanen \(2015\)](#) argue that conditional on additional covariates, if the running variable and the outcome variable are uncorrelated, then it is more plausible to extrapolate the treatment effect to points away from the threshold.

Lastly, consider the quantile regression of the treatment effect on the outcome variables. [Abadie et al. \(2002\)](#) and [Chernozhukov and Hansen \(2005\)](#) develop approaches to IV quantile treatment effects by which the method of [Abadie et al. \(2002\)](#) can be applied to a sub-population of compliers with binary treatment variables, while the method by [Chernozhukov and Hansen \(2005\)](#) can be applied to the whole population with both discrete and continuous treatment variables. To estimate the local quantile treatment effects, [Frandsen et al. \(2012\)](#) use a local linear estimator to estimate the marginal distributions of the potential outcomes. [Shen and Zhang \(2016\)](#) test the treatment effect of a policy intervention on the whole distribution of a response outcome with a Kolmogorov-Smirnov-type test, which is free of any distributional assumptions. These examples show that quantile regression serves as an ideal approach for characterizing the treatment effect on the distribution of the outcome variable.

### 1.3.6 Optimal RD Plots

As an important tool for exploratory data analysis, graphical representation plays a critical role in validating the suitability of the RDD approach for a given context. Most RDD plots are ad hoc and use an evenly spaced binning strategy to show a scatter plot representation of the sample. [Calonico et al. \(2015\)](#) introduce optimal RD plots by providing several data-driven methods to optimally choose the number of bins according to the goals of the researchers. This helps researchers to uncover and communicate the underlying statistical properties of data.

## 1.4 Extensions of RDD

Beyond their original use for investigating policy intervention with cross-sectional independent data, one running variable, and a known discontinuity point, RDD have been extended to a variety of situations: Kink Regression Designs (KRD), unknown thresholds, multiple thresholds, categorical

outcomes, dependent samples, and the running variable with rounding and measurement errors. These examples show a variety of uses and adaptability of RDD in non-standard situations.

#### 1.4.1 Kink Regression Designs

RDD identify the ATE in settings where the regression function has a known discontinuity. Similarly, KRDD exploits the opportunity to identify the ATE where the regression function is kinked. That is the regression function is continuous but its first order derivative is discontinuous. RDD is called upon when there exists endogeneity between the treatment variable  $D$  and the outcome variable  $Y$ , and I cannot find a proper IV. The running variable  $X$  is not an IV in the conventional sense – that is, a proper IV is exogenous and does not play a role in the regression of  $Y$ . In RDD, I allow  $X$  and  $Y$  to correlate as long as the regression of  $Y$  on  $X$  is continuous. If there is a discontinuity in the regression of  $D$  on  $X$ , I expect to see the discontinuity in the regression of  $Y$  on  $X$  if there is a causal effect of  $D$  on  $Y$ . The same logic can be used in the kink design: if there is a kink in the regression of  $D$  caused by  $X$ , then I should expect to see a kink in the regression  $Y$  if there is a causal effect of  $D$  on  $Y$ .

[Dong \(2010\)](#) entertains the idea that even if there is no jump, a kink can be used to identify treatment effects. [Card et al. \(2015\)](#) describe the conditions and identification for kink designs and compare estimation performance by local linear and local polynomial estimators from [Fan \(1992\)](#), estimators from [Imbens and Kalyanaraman \(2012\)](#) that select an optimal bandwidth, and estimators from [Calonico et al. \(2014\)](#) that have bias correction and produce robust confidence interval. [Chiang and Sasaki \(2019\)](#) extend the kink design to quantile kink design and provide the asymptotic properties of the estimators for inference.

#### 1.4.2 Unknown Thresholds

RDD are primarily used in situations where the discontinuity points are known. [Porter and Yu \(2015\)](#) consider unknown thresholds. First, they use a unified test to check if there are selection or treatment effects. Once the effects are confirmed, they use a difference kernel estimator to estimate the position of the jump, then estimate the size of the jump as if the discontinuity is known.

Surprisingly, although the position of the jump is estimated rather than known, the efficiency of the estimator is not affected. That is, the treatment estimator with an unknown discontinuous point has the same limiting distribution as the treatment estimator with a known discontinuous point.

### 1.4.3 Multiple Thresholds

When the running variables are associated with multiple thresholds, researchers often fold them into one threshold and estimate one treatment effect. [Cattaneo et al. \(2016\)](#) study this process of obtaining one treatment effect through pooled and normalized regression and give guidance on the interpretation of the results. [Bertanha \(2017\)](#) argues the results generated by this method rely on restrictive assumptions – either requiring the same LATE or only being valid for individuals near the cutoff – and proposes consistent and asymptotically normal estimators. In these cases, heterogeneity is informed by one of the following processes: a nonparametric distribution from the data, a deterministic function from economic theory in the case of a sharp design, or a parametric distribution in the case of a fuzzy design.

### 1.4.4 Multiple Running Variables

RDD also allow for the existence of multiple running variables. In manner similar to [Imbens and Kalyanaraman \(2012\)](#), who study the optimal bandwidth for a single running variable, [Imbens and Zajonc \(2011\)](#) study optimal bandwidth selection for multiple running variables given that the second derivative of the regression function is bounded. Focusing on identification, [Papay et al. \(2011\)](#), [Wong et al. \(2013\)](#), and [Reardon and Robinson \(2012\)](#) study optimal inference for LL estimators with multiple running variables.

### 1.4.5 Categorical and Duration Outcomes

When outcome variables are categorical, LL estimators are often inappropriate. For example, in the case of binary outcomes, the estimated probability implied by LL estimators is not guaranteed to fall in the unit interval. [Xu \(2017\)](#) studies estimators for categorical outcomes in RDD settings.



To estimate the treatment effect, the author proposes a new nonparametric estimator based on the extension of the multinomial logit model and addresses the issues of optimal bandwidth selection and robust interference through a local likelihood method.

A special type of categorical outcome relates to duration outcomes, such as unemployment duration (i.e. survival analysis). [Xu \(2018\)](#) uses a discrete-time framework to estimate the unconditional duration effect and conditional hazard effect for each discrete level. In this setup, the objects of interest are the local risk function for each discrete level and the global baseline function. The author uses a maximum likelihood estimator for the local risk function and series approximation for the global baseline function. This work shows that although the LL estimators are often used in a standard RDD setup, other types of estimators can be better suited for extended RDD setups – such as this, in the situation of categorical outcomes.

#### 1.4.6 Dependent Samples

RDD are most often applied to IID samples, but there are many situations where the assumption of IID does not hold. The following papers extend RDD to dependent samples, such as clustered samples, time series samples, or spatially correlated geographic samples. [Bartalotti and Brummet \(2017\)](#) study the effects of clustering on inference and bandwidth selection in RDD. They extend the optimal bandwidth selection by [Imbens and Kalyanaraman \(2012\)](#) and the optimal inference by [Calonico et al. \(2014\)](#) for IID samples to clustered samples. [Kuersteiner et al. \(2018\)](#) use an RDD to estimate the non-linear impulse response function of a time series. [Keele and Titiunik \(2015\)](#) apply RDD to a geographic sample where estimation needs to account for multiple running variables, different measures of the distance from the cutoffs, and spatial correlation.

#### 1.4.7 Rounding and Measurement Error

In the standard setup, the running variable is continuous, but in reality, many running variables are often measured in discrete units. [Dong \(2015\)](#) considers inconsistent estimates caused by rounding errors in discrete running variables. A more general case of rounding errors is heaping. The heaping of the running variable at a particular values can be induced by a variety of reasons,

such as the unit of measurement used or data rounding. [Barreca et al. \(2016\)](#) study the heaping-induced bias in RDD. Finally, [Davezies and Le Barbanchon \(2017\)](#) consider continuous errors rather than discrete errors in the running variable – a mixture of the true running variable and a noisy proxy – and come up with a consistent nonparametric estimator.

## 1.5 Conclusion

The past two decades have seen a flourishing of developments on RDD: over 2,000 empirical and 50 methodology papers have been published on RD models in leading economic journals. In this chapter, I review the recent methodological developments on RDD in the econometrics literature with a focus on RD estimators in both standard and extended setups. Pioneered by [Fan \(1992\)](#) for their desirable properties at the boundary and recommended by [Hahn et al. \(2001\)](#) for RDD, the local linear or local polynomial estimators are primary estimators. While some papers have made important amendments to LL estimates and others have employed different types of estimators for RDD used in non-traditional setups, research on alternative RD estimators has been few and far between. There are almost no competitive alternatives for LL estimators. In this dissertation, I propose a complete new class of RD estimators using general reflection approach. In Chapter 2, I introduce my RD estimators for univariate regression and in Chapter 3, I introduce my RD estimators for multivariate regression. Having boundary properties as good as LL estimators and performing better than LL estimators in finite sample estimations, my estimators offer compelling alternatives for LL estimators in RD models.

## Chapter 2

### Estimation of a Regression Jump: a Generalized Reflection Approach

As a method for empirically evaluating the effects of policy or experimental interventions, regression discontinuity (RD) designs have been widely used in economics, political science and other social and behavioral sciences. See [Imbens and Lemieux \(2008\)](#), [Lee and Lemieux \(2010\)](#) and [Cattaneo and Escanciano \(2017\)](#) for theoretical and empirical overviews of the existing literature. RD designs are inherently nonparametric models, as identification typically relies only on smoothness assumptions on the relevant conditional expectations at a known threshold or cut-off point in the set where the conditioning covariate (regressor) takes values. It is well known that traditional nonparametric kernel regression estimators, such as Nadaraya-Watson (NW), suffer from boundary problems (see, *inter alia*, [Gasser and Müller, 1979](#), [Gasser and Müller, 1984](#), [Fan, 1992](#), [Härdle and Linton, 1994](#)). Specifically, these estimators have slower rates of convergence for bias at boundary points than at interior points in the regression domain. Under typical assumptions on the smoothness of the regression and regressor density, the traditional Nadaraya-Watson estimator constructed with bandwidth  $h > 0$  has bias of order  $O(h)$  at boundary points, compared to  $O(h^2)$  at interior points of the regression domain. This problem is particularly relevant for RD designs, as the estimation of regression functions at boundary points is precisely the object of interest.

The problem can be aggravated in RD designs, see [Porter \(2003\)](#), as an estimator for the jump discontinuity at the threshold may compound the poor bias behavior of nonparametric estimators of the regression to the right and to left of the threshold. While there is a vast literature in econometrics and statistics that attempts to address this issue, (see, *inter alia*, [Fan, 1992](#), [Hahn et al., 2001](#), [Porter, 2003](#), [Imbens and Lemieux, 2008](#), [Lee and Lemieux, 2010](#), [Imbens and Kalyanaraman, 2012](#)) the main approach in RD designs is to estimate local polynomial (mostly local linear) approximations for the regressions on both sides of the discontinuity and use these to produce an estimate for the jump discontinuity at the threshold. This approach is justified by [Fan and Gijbels \(1992\)](#), where it is shown that local linear estimators, under standard smoothness assumptions,

have bias of order  $O(h^2)$  at boundary points. [Porter \(2003\)](#) proposes RD estimators – partially linear local polynomial estimators – that can achieve smaller order biases at boundary points by using high order kernels. However, these estimators require identical regression functions (separated by a jump) on both sides of the threshold, a restriction that is not required by typical nonparametric estimators when applied to data to the left and right sides of a point of discontinuity.

In this chapter, I adopt a novel approach to construct estimators for a regression jump. The basic idea behind my estimation procedure is to extend regressions beyond the boundary of their domains to the entire real line, using an extension proposed by [Hestenes \(1941\)](#). These extended regressions are then estimated using a generalized reflection approach and used to estimate a jump discontinuity. The inspiration for my method comes from [Mynbaev and Martins-Filho \(2018\)](#), where a simple and elegant solution to boundary problems in density estimation is obtained using the same extension principle. Their solution can be applied not only to densities but also to any sufficiently smooth function, such as suitably defined regressions. I apply Hestenes’ extension to estimate regressions that have a jump discontinuity and can thus be viewed as comprising two regimes with boundaries: one to the left and one to the right of the point of discontinuity. Regression functions on each side of the discontinuity can be different and, in particular, can have different degrees of smoothness. In essence, instead of using higher order polynomial functions to reduce bias, I use a generalized reflection kernels to estimate the extended the regression functions across discontinuity points to reduce bias. I call my estimators – the HT estimators.

In fact, my estimation strategy produces a *class* of jump discontinuity estimators. What distinguishes the elements in the class are the types of Hestenes’ extension used. My estimators are constructed based on the algebraic structure of the classical NW estimator. However, contrary to the NW estimator that suffers from the aforementioned boundary problems (slow rates of bias decay and , in some cases, inconsistency), my estimators have boundary behavior that is completely analogous to that at interior points of the regression domain. Thus, I restore bias behavior at boundary points to be the same as that at interior points. The estimators I propose are easy to construct, require no modification to commonly used kernels and allow for a common bandwidth over the entire domain of the regressions.

Focusing on properties at boundary points, I derive the bias, variance and asymptotic distribution of my estimators. In addition, I provide a theoretical comparison between my estimator and the popular local linear (LL) estimator. These two types of estimators have the same unconditional bias of order  $O(h^2)$  and variance order  $O((nh)^{-1})$ , where  $n$  is the sample size, but with different magnitudes for both bias and variance. My estimators solve the boundary problems by completing partial integration of kernels through reflection method: the bias at the boundary is the same as that at the interior points. The variance at the boundary is affected by the chosen parameters of [Hestenes \(1941\)](#) extension. On the other hand, the LL estimators solve the boundary problems by adapting to the boundary with effective kernels: the bias is smaller but the variance is larger at the boundary point than those at interior points, and the bias and variance of LL estimators are affected by partial integration of kernels. Given typical smoothness assumptions on regression and regressor density functions, such as they have continuous second order derivatives, my estimators always have smaller variances of asymptotic distributions than LL estimators, which leads to smaller mean squared errors (MSE) in most cases.

I have conducted extensive Monte Carlo studies to shed light on the finite sample behavior of my estimators. I compare bias, variance, and root mean square error (RMSE) of my estimators with those of the NW and LL estimators using different bandwidth selection methods, kernels, sample sizes, and data generation processes. The simulation results show that my estimators perform better than NW estimators in all cases and have smaller RMSE than LL estimators in most cases.

To illustrate the applicability of my estimators in empirical settings, I apply them to data used in [Litschig and Morrison \(2013\)](#), where a RD design is used to examine how government spending affects education and poverty by taking advantage of an intergovernmental transfers program in Brazil. The authors use the ordinary least squares (OLS) procedure to implement a local linear estimator to estimate jumps, which is a typical approach in empirical RD design studies. Although this approach is straightforward and has avoided the intricacy of choosing kernels and bandwidth for a typical nonparametric estimation, it is not flexible. First, the bandwidth is arbitrarily chosen. When researchers choose the percentage of the sample around the vicinity of the discontinuity for estimation, they are arbitrarily choosing the bandwidth. Second, they restrict themselves to the

uniform kernel choice. As a result, after they have tried different bandwidths and different orders of polynomial functions to produce several sets of estimates, they cannot tell which set of estimates is preferred. To be clear, these are not limitations of LL estimators, but the trade-off – exchanging flexibility for convenience – by using the OLS procedure.

I apply my estimators and the LL estimators to estimate the jump in the conditional mean of the treatment and outcome variables, allowing for a variety of kernels and choosing an optimal bandwidth. I compare estimates from the HT and LL estimates to their estimates. The results in this case are qualitatively the same but quantitatively different. The subsequent data-calibrated simulations using the same empirical example show that my estimators have reduced the RMSE of the LL estimator by more than a half. This means that my estimators provide more precise estimations and powerful tests. That is, there could be situations that tests based on my estimators reject the null hypothesis that there is no treatment effect while tests based on LL estimators fail to reject the null. In another word, my estimators could detect the treatment effect that otherwise would be dismissed.

My estimators have simple and intuitive expressions that are easy to implement using any statistic software. They are also computationally simpler than the LL estimators. The LL estimators have a singularity problem with their matrix while my estimators don't have a singularity issue and always provide stable estimates. Considering the popularity of RDD in empirical studies, I have developed a full set of nonparametric RDD estimation packages in Matlab which includes LL and Hestenes estimators for both regression and density estimation. The empirical example has demonstrated that my estimation procedure is straightforward and my estimation package is readily available to use.

The rest of this chapter is organized as follows. In Section 2.1, I introduce the Hestenes-based estimators for regressions with a domain that includes boundary points. Then I derive the bias, variance and asymptotic distribution of my estimator for points in (and outside) a vicinity of the boundary. I compare these properties with those of LL estimators. In Section 2.2, I connect regression estimators to estimation of a jump discontinuity within the context of the RD design literature. I define estimators for the jump discontinuity and establish consistency and asymptotic normality

of the estimator. In Section 2.3, I conduct Monte Carlo simulations that compare my estimators to NW and LL estimators. In Section 2.4, I give an empirical illustration of my methodology. In Section 2.5, I conduct empirical motivated simulations to verify the finite sample performance of my estimators. Section 2.6 gives concluding remarks and topics for future study. Supporting lemmata are collected in Appendix A. MATLAB codes that implement the NW, LL, and HT regression estimators, as well as LL and HT density estimators, for the simulation and empirical application are collected in Appendix B.

## 2.1 A General Reflection Approach: Hestenes-based Regression Estimator

The general reflection approach solves the boundary problem by forming a reflection kernel at the boundary. To show how it works, I start by explicitly comparing the regular kernel functions of the NW estimator and the reflection kernel functions of the HT estimator at the vicinity of boundary. Suppose I have a random sample  $\{X_i, Y_i\}$  of size  $n$  from a random vector  $(Y, X)$ .  $X$  has a density  $f$  defined on  $[0, \infty)$ . The classical Nadaraya-Watson estimator is defined as follows:

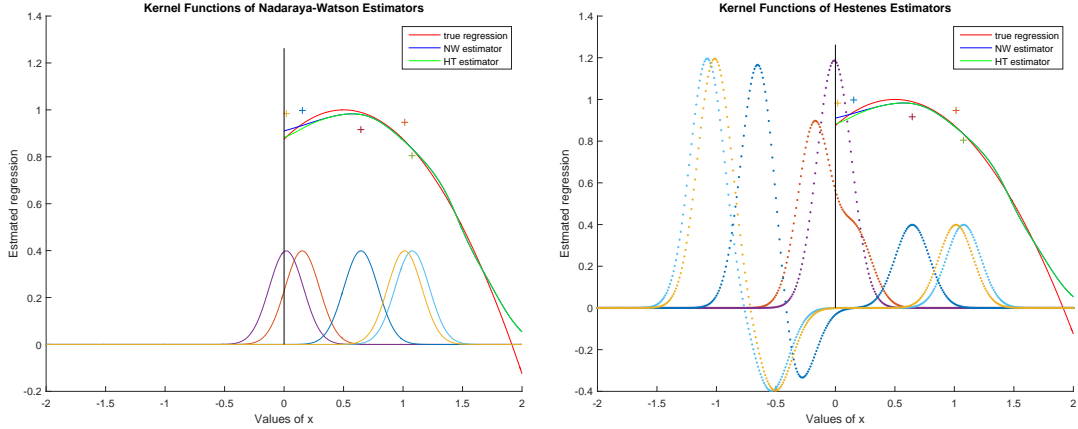
$$\hat{m}_{NW}(x) = \frac{\frac{1}{nh} \sum_{i=1}^n K_i(x) Y_i}{\frac{1}{nh} \sum_{i=1}^n K_i(x)} = \sum_{i=1}^n w_{in}(x) Y_i,$$

where  $K_i(x) = K\left(\frac{X_i - x}{h}\right)$ .  $h$  is a bandwidth, which is a function of  $n$ .  $w_{in}(x)$  is the weight, which denotes the contribution of each observation to the regression estimate at  $x$ . Thus,  $\hat{m}_{NW}(x)$  is a weighted average of the observations  $Y_1, \dots, Y_n$ .

For a fixed  $n$ , boundary points are those points in the range of  $X$  where  $x < h$  and interior points are those where  $x \geq h$ . As  $n \rightarrow \infty$ ,  $h \rightarrow 0$ . The only boundary point is  $x = 0$ . At the boundary point, the estimator has a larger bias than those at interior points. This is the well-known boundary problem. Figure 2.1 illustrates what causes the boundary problem in the NW estimator and how the general reflection approach implemented by the Hestenes estimator solves the problem.

The two graphs show the same estimation with kernel functions for the NW estimator and the HT estimator respectively. At the boundary, the estimate from the HT estimator stays close to the true regression while the estimate from the NW estimator is away from the true regression. This

**Figure 2.1:** Kernel functions in NW and Hestenes Estimators



results from the fact that the two estimators have different kernel functions for observations near the boundary.

The figure on the left shows the five observations (denoted by + around the regression function) at the vicinity of zero and their correspondent kernel functions  $K_i(x)$  (five solid curves underneath the regression function) of the NW estimator. Since the observations can only occur on the positive side of the real line, the kernel function should only be on the positive side of the real line. However, the kernel functions cross over the point zero to the area that is not part of the domain, so the weights given by the kernel function to observations at the vicinity of the boundary are too small, which causes a larger bias at the boundary.

The figure on the right shows the same five observations (denoted by + around the regression function) at the vicinity of zero and their correspondent kernel functions (five dotted curves underneath the regression function) from the HT estimator. Here, the kernel functions are extended to the negative side of the real line to estimate the regression function that is extended to the whole real line (the negative side is not drawn). The kernel can even have negative values, which adds flexibility. Compared to the kernel function of the NW estimator, the kernel functions (the two curves in purple and red near the origin) of observations that are close to the boundary have changed to reflect the increased weights for these observations, whereas the kernel functions (the three curves in blue, orange, light blue on the right side) of observations that are away from the origin gradually merge to be the same kernels of the NW estimator. This is the general reflection



method at work: the observations at the vicinity of zero are reflected to the other side in order to estimate the regression defined in the whole real line, so the lost weights for observations at the vicinity of the boundary are added back, which reduces the bias at the boundary and solves the boundary problem.

### 2.1.1 Hestenes Extensions

Now, I describe how Hestenes extensions are defined. Consider an independent and identically distributed sequence of random vectors in  $\mathbb{R} \times [0, \infty)$ , denoted by  $\{(Y_i, X_i)\}_{i=1}^n$ , each of which is distributed as  $(Y, X)$ , with

$$E(Y|X = x) = m(x). \quad (2.1)$$

I assume that the marginal density  $f$  of  $X$  exists and that  $m, f \in C_b^s([0, \infty))$ , where  $C_b^s([0, \infty))$  the class of functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  whose support is  $[0, \infty)$  and which is  $s$ -times differentiable with  $|f^{(s)}(x)| \leq C$  for some  $0 < C < \infty$ ,  $s \in \mathbb{N}$  and  $f^{(s)}(x)$  denotes the derivative of order  $s \in \mathbb{N}$  of  $f$ . To overcome boundary problems when estimating  $m$ , I smoothly extend the function  $r(x) \equiv m(x)f(x)$  from its original domain  $[0, \infty)$  to  $(-\infty, 0)$ . Its smooth extensions to  $(-\infty, 0)$  will be denoted by  $\phi(x)$  for  $x < 0$  and are given as  $\phi(x) = \sum_{i=1}^{s+1} k_i r(-w_i x)$ . In words, this expression means that I scale  $r$  and reflect it over from the nonnegative to the negative side of the real line up to  $s+1$  times. The extension  $\phi$  is a linear combination of the reflection of scaled  $r$  functions which satisfy sewing conditions that preserve continuity and derivatives up to order  $s$ . I will then use the observations on  $(Y, X)$ , which include only nonnegative values of  $X$  to estimate  $\mu(x)$ , which is defined on the whole real line by piecing together  $r$  and  $\phi$ .

$$\mu(x) \equiv \begin{cases} m(x)f(x), & x \geq 0 \\ \sum_{i=1}^{s+1} k_i m(-w_i x)f(-w_i x), & x < 0. \end{cases} \quad (2.2)$$

Specifically, let  $w_1, \dots, w_{s+1}$  be pairwise distinct positive numbers for  $s = 0, 1, \dots$  such as,  $w_i = 1/i$ , or  $w_i = i$ , for  $i = 1, \dots, s+1$ . Also, let the numbers  $k_1, \dots, k_{s+1}$  be defined by the

following system of equations

$$\sum_{i=1}^{s+1} (-w_i)^j k_i = 1, \quad j = 0, \dots, s. \quad (2.3)$$

The determinant of this system (Vandermonde) is nonsingular, i.e.,

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ -w_1 & -w_2 & \cdots & -w_{s+1} \\ \vdots & \vdots & \vdots & \vdots \\ (-w_1)^s & (-w_2)^s & \cdots & (-w_{s+1})^s \end{vmatrix} \neq 0,$$

and consequently  $k_1, \dots, k_{s+1}$  are uniquely defined for any choice of  $\{w_i\}_{i=1}^{s+1}$ . Then, it follows immediately that sewing conditions are satisfied due to (2.3).

$$\phi^{(j)}(0-) = \sum_{i=1}^{s+1} (-w_i)^j k_i r^{(j)}(0+) = r^{(j)}(0+), \quad j = 0, \dots, s,$$

where for an arbitrary function  $g : \mathbb{R} \rightarrow \mathbb{R}$  and  $\varepsilon > 0$ ,  $g(x+) = \lim_{\varepsilon \downarrow 0} g(x+\varepsilon)$  and  $g(x-) = \lim_{\varepsilon \downarrow 0} g(x-\varepsilon)$ .

$\mu(x)$  is  $s$ -times differentiable. This follows from  $s$ -times differentiability of  $f$  and  $m$  as

$$[m(x)f(x)]^{(s)} = \sum_{j=0}^s \binom{s}{j} m^{(s-j)}(x) f^{(j)}(x).$$

Note that in my context, differentiability of  $m(x)$  at  $x = 0$  must be understood as differentiability from the right, i.e.,  $m^{(s)}(0+)$ .<sup>1</sup> In the following subsections, I construct estimators for  $m(x)$  where  $x \in [0, \infty)$  and study their biases, variance and asymptotic distributions for points in a vicinity of

<sup>1</sup>As an example, the sewing condition for  $s = 1$  is satisfied as

$$\begin{aligned} \phi^{(1)}(0-) &= \lim_{\varepsilon \downarrow 0} \phi^{(1)}(0-\varepsilon) = \lim_{\varepsilon \downarrow 0} \sum_{j=1}^2 k_j r^{(1)}(-w_j(0-\varepsilon)) \\ &= \lim_{\varepsilon \downarrow 0} \sum_{j=1}^2 k_j [m^{(1)}(-w_j(0-\varepsilon))f(-w_j(0-\varepsilon)) + m(-w_j(0-\varepsilon))f^{(1)}(-w_j(0-\varepsilon))] \\ &= \sum_{j=1}^2 (-w_j)k_j [m^{(1)}(0+)f(0) + m(0+)f^{(1)}(0)] = r^{(1)}(0+). \end{aligned}$$

zero, and compare them to cases where  $x$  is an interior point, i.e., outside this vicinity. It is well known that the classical NW estimators have poor properties at boundary points than those at interior points. I show that my estimators have the same bias expression at the boundary points as that of interior points.

Schuster (1985) proposes a boundary folding method to reduce bias for *density* estimation at the boundary, which multiplies the sample by two and flips the weights from one side of the discontinuous point to the other side. As pointed out by Mynbaev and Martins-Filho (2018), Schuster's reflection estimator does not use the smoothness beyond continuity and only works well when the first derivative of the density function at the discontinuity point  $f^{(1)}(0) = 0$ , so Schuster's reflection estimator can be seen as a special case of the general reflection method that are based on Hestenes extension.

### 2.1.2 Infeasible Hestenes-based Regression Estimators

I define an infeasible Hestenes-based regression estimator  $m_H(x)$  for  $m(x)$  when  $x \geq 0$  which assumes that  $f$  is known. A kernel function  $K$  satisfies the following condition.

**Assumption 1.**  $K$  is uniformly bounded and symmetric with  $\int_{\mathbb{R}} K(u)du = 1$ ,  $\int_{\mathbb{R}} uK(u)du = 0$  and  $\int_{\mathbb{R}} |u^i K(u)| du < C$  for  $i = 1, 2, 3, 4$  and some  $0 < C < \infty$ .

If  $f(x) > 0$ , I define

$$m_H(x) = \frac{1}{f(x)} \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] Y_i, \quad (2.4)$$

where  $n$  is taken to be the size of a random sample of observations  $\{(Y_i, X_i)\}_{i=1}^n$  and  $h > 0$  is a sequence of nonstochastic bandwidths that depend on  $n$  such that  $h \rightarrow 0$  as  $n \rightarrow \infty$ .

The algebraic structure of the estimator is motivated by Mynbaev and Martins-Filho (2018) where density estimators are constructed based on Hestenes' extension. Since  $m_H$  depends on  $s$  and the sequence  $\{w_j\}_{j=1}^{s+1}$ , equation (2.4) defines a *class* of estimators whose elements are indexed by  $\{w_j\}_{j=1}^{s+1}$ . For instance,  $w_j = 1/j$  or  $w_j = j$  have been suggested in Mynbaev and Martins-Filho

(2018) and will produce different estimators in the class. Once  $\{w_j\}_{j=1}^{s+1}$  is chosen, the sequence  $\{k_j\}_{j=1}^{s+1}$  is uniquely defined by (2.3) so every estimator in the class is uniquely indexed by  $\{w_j\}_{j=1}^{s+1}$ .

The following theorem gives an integral representation for the bias of  $m_H$ . Its proof gives the mathematical motivation for the algebraic structure of the estimator. In what follows I adopt the following notations:  $\kappa_i = \int_{\mathbb{R}} u^i K(u) du$ ,  $\mu_{i,x} = \int_{-\infty}^x u^i K(u) du$ ,  $\kappa_{i,x} = \int_x^{\infty} u^i K(u) du$ ,  $\lambda_i = \int_{\mathbb{R}} u^i K^2(u) du$ ,  $\eta_{i,x} = \int_{-\infty}^x u^i K^2(u) du$  and  $\lambda_{i,x} = \int_x^{\infty} u^i K^2(u) du$  for  $i = 0, 1, 2, 3, 4$ .

**Theorem 1.** *Suppose that  $K$  satisfies Assumption 1. Then,*

$$E(m_H(x)) - m(x) = \frac{1}{f(x)} \int_{\mathbb{R}} K(\psi) [\mu(x - h\psi) - \mu(x)] d\psi, \quad (2.5)$$

where  $\mu(x)$  is as defined in equation (2.2). If, in addition,  $f, m \in C_b^4([0, \infty))$ ,

$$E(m_H(x)) - m(x) = h^2 \frac{\kappa_2}{2} \left[ m^{(2)}(x) + \frac{2m^{(1)}(x)f^{(1)}(x)}{f(x)} + \frac{f^{(2)}(x)m(x)}{f(x)} \right] + O(h^4). \quad (2.6)$$

*Proof.* Since  $\{Y_i, X_i\}_{i=1}^n$  is independent and identically distributed sequence

$$\begin{aligned} E(m_H(x)|X_1, \dots, X_n) &= \frac{1}{f(x)} \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] m(X_i) \\ E(m_H(x)) &= \frac{1}{f(x)} \frac{1}{h} E \left[ K\left(\frac{X_1 - x}{h}\right) m(X_1) \right] + \frac{1}{f(x)} \frac{1}{h} E \left[ \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_1}{w_j} + x}{h}\right) m(X_1) \right] \\ &= \frac{1}{f(x)} \frac{1}{h} \int_0^{\infty} K\left(\frac{X_1 - x}{h}\right) m(X_1) f(X_1) dX_1 \\ &\quad + \frac{1}{f(x)} \frac{1}{h} \int_0^{\infty} \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_1}{w_j} + x}{h}\right) m(X_1) f(X_1) dX_1. \end{aligned}$$

Since  $K$  is symmetric, in the first integral let  $\psi = \frac{x - X_1}{h}$  and in the second integral, let  $\psi = \frac{\frac{X_1}{w_j} + x}{h}$ .

Then,

$$\begin{aligned} E(m_H(x)) &= \frac{1}{f(x)} \int_{-\infty}^{\frac{x}{h}} K(\psi) m(x - h\psi) f(x - h\psi) d\psi \\ &\quad + \frac{1}{f(x)} \int_{\frac{x}{h}}^{\infty} \sum_{j=1}^{s+1} k_j K(\psi) m(-(x - h\psi)w_j) f(-(x - h\psi)w_j) d\psi. \end{aligned}$$

In the first integral, I have  $x - h\psi \geq 0$  and in the second integral, I have  $x - h\psi < 0$ . Hence,

$$E(m_H(x)) = \frac{1}{f(x)} \int_{-\infty}^{\infty} K(\psi) \mu(x - h\psi) d\psi. \quad (2.7)$$

Since  $\int_{\mathbb{R}} K(\psi) d\psi = 1$ ,

$$E(m_H(x)) - m(x) = \frac{1}{f(x)} \int_{-\infty}^{\infty} K(\psi) [\mu(x - h\psi) - \mu(x)] d\psi \text{ for } x \geq 0. \quad (2.8)$$

For the second part of the theorem, note that since  $f, m \in C_b^4([0, \infty))$  for  $x \geq 0$ , I have  $\mu^{(4)}(x) = \sum_{j=0}^4 \binom{4}{j} m^{(4-j)}(x) f^{(j)}(x)$ . Using differentiability of  $\mu$ , I have

$$\begin{aligned} E(m_H(x)) - m(x) &= \frac{1}{f(x)} \int_{-\infty}^{\infty} K(\psi) \left( \mu^{(1)}(x)(-h\psi) + \frac{1}{2} \mu^{(2)}(x)(-h\psi)^2 \right. \\ &\quad \left. + \frac{1}{6} \mu^{(3)}(x)(-h\psi)^3 + \frac{1}{24} \mu^{(4)}(\bar{x})(-h\psi)^4 \right) d\psi \\ &= \frac{1}{f(x)} \int_{-\infty}^{\infty} K(\psi) \left( \frac{1}{2} \mu^{(2)}(x)(-h\psi)^2 + \frac{1}{24} \mu^{(4)}(\bar{x})(-h\psi)^4 \right) d\psi, \end{aligned}$$

where  $\bar{x} = \alpha x + (1 - \alpha)(x - h\psi)$ ,  $\alpha \in [0, 1]$ , and  $K$  is symmetric.

Now,

$$\left| \int_{-\infty}^{\infty} K(\psi) \psi^4 \mu^{(4)}(\bar{x}) d\psi \right| \leq \int_{-\infty}^{\infty} |K(\psi)| \psi^4 |\mu^{(4)}(\bar{x})| d\psi \leq C,$$

for some  $C < \infty$ , provided  $|\mu^{(4)}(x)| < C < \infty$ , and  $\int_{-\infty}^{\infty} |K(\psi)| \psi^4 d\psi < C$ . Consequently,

$$\begin{aligned} E(m_H(x)) - m(x) &= \left( \frac{\mu^{(2)}(x)}{2f(x)} \int_{-\infty}^{\infty} K(\psi) \psi^2 d\psi \right) h^2 + O(h^4) = \frac{\mu^{(2)}(x)}{2f(x)} \kappa_2 h^2 + O(h^4) \\ &= \frac{h^2}{2} \left[ m^{(2)}(x) + \frac{2m^{(1)}(x)f^{(1)}(x)}{f(x)} + \frac{f^{(2)}(x)m(x)}{f(x)} \right] \kappa_2 h^2 + O(h^4). \end{aligned}$$

□

The integral representation for the bias of  $m_H(x)$  given in equation (2.5) is the key insight in constructing the estimator. It shows that the representation obtained for traditional estimators, such as NW, can be obtained with the function  $\mu$  in place of the regressor density. The algo-

braic structure of  $m_H$  permits the unification of the bias representations for  $x$  in and outside a neighborhood of zero.

As pointed out [Mynbaev and Martins-Filho, 2018](#), there is a simple link between the degree of smoothness of the density and the order of estimator's bias. This link also exists for regression estimators, such as NW estimators, in a regular setting where the regression is continuous, but this link is broken for the NW estimator at the boundary. As a direct consequence of the method of the proof of Theorem 1, if, as usually done in the nonparametric kernel literature, the kernel  $K$  is of order  $s$  and  $m, f \in C_b^s([0, \infty))$  then the bias of  $m_H$  is  $O(h^s)$ . It is in this sense I claim that the HT estimator has recovered the missing link and restored the bias at the boundary points to be the same as that of interior points. At the boundary point  $x = 0$ , I have

$$E(m_H(0)) - m(0) = \frac{h^2}{2} \left[ m^{(2)}(0) + \frac{2m^{(1)}(0)f^{(1)}(0)}{f(0)} + \frac{f^{(2)}(0)m(0)}{f(0)} \right] \kappa_2 + O(h^4).$$

It is instructive to compare this expression to that of the bias for an infeasible NW estimator given by

$$m_{NW}(x) = \frac{1}{nhf(x)} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i.$$

The following corollary follows directly from Theorem 1.

**Corollary 1.** *Under the assumptions of Theorem 1, for  $x \geq 0$ , the bias of  $m_{NW}(x)$  is given by*

$$\begin{aligned} E(m_{NW}(x)) - m(x) &= -m(x)\kappa_{0, \frac{x}{h}} - h \left( \frac{m(x)f^{(1)}(x)}{f(x)} + m^{(1)}(x) \right) \mu_{1, \frac{x}{h}} \\ &+ h^2 \left( \frac{m(x)f^{(2)}(x)}{2f(x)} + \frac{m^{(1)}(x)f^{(1)}(x)}{f(x)} + \frac{m^{(2)}(x)}{2} \right) \mu_{2, \frac{x}{h}} + O(h^3). \end{aligned}$$

The slower order of the remainder term results from the fact that the symmetry of  $K$  can no longer be used to eliminate the term of order  $h$ . Hence, at  $x = 0$ ,

$$\begin{aligned} E(m_{NW}(0)) - m(0) &= -\frac{1}{2}m(0) - h \left( \frac{m(0)f^{(1)}(0)}{f(0)} + m^{(1)}(0) \right) \mu_{1,0} \\ &+ \frac{h^2}{2} \left( m^{(2)}(0) + \frac{2m^{(1)}(0)f^{(1)}(0)}{f(0)} + \frac{f^{(2)}(0)m(0)}{f(0)} \right) \mu_{2,0} + O(h^3). \end{aligned}$$

This expression suggests that  $m_{NW}$  is inconsistent at the boundary. Compared to the bias of  $m_H$ , it has two extra terms that are of larger order,  $O(1)$  and  $O(h)$ . This results in a large bias for the NW estimator.

The following theorem provides approximations for the variance of  $m_H$ .

**Theorem 2.** *Suppose  $E((Y - m(X))^2|X = x) = \sigma^2$ ,  $m, f \in C_b^4([0, \infty))$  and  $K$  satisfies Assumption 1. Then, if  $nh \rightarrow \infty$  and  $x \geq 0$ , the variance of  $m_H(x)$  is given by*

$$V(m_H(x)) = \begin{cases} \frac{1}{nh} \frac{m^2(0) + \sigma^2}{f(0)} \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du + o((nh)^{-1}), & x = 0 \\ \frac{1}{nh} \frac{m^2(x) + \sigma^2}{f(x)} \lambda_0 + o((nh)^{-1}), & x > 0 \end{cases}.$$

*Proof.* Write  $m_H(x) = \frac{\hat{g}(x)}{f(x)}$  where  $f(x) \neq 0$  and

$$\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] Y_i.$$

and write

$$u_i = \frac{1}{h} \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] Y_i.$$

Then let  $w_0 = -1$ , and  $k_0 = -1$ , and write

$$u_i = \frac{1}{h} \sum_{j=0}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) Y_i.$$

Then,  $\hat{g}(x) = \frac{1}{n} \sum_{i=1}^n u_i$ . Consequently,

$$V(m_H(x)) = \frac{1}{nf^2(x)} \left( E u_1^2 - E(u_1)^2 \right)$$

Now,

$$\begin{aligned}
E(u_1^2) &= E \left[ \sum_{j=0}^{s+1} \frac{1}{h} \frac{k_j}{w_j} K \left( \frac{X_1 + x}{h} \right) Y_1 \right]^2 \\
&= \frac{1}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} E \left[ K \left( \frac{X_1 + x}{h} \right) K \left( \frac{X_1 + x}{h} \right) Y_1^2 \right] \\
&= \frac{1}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} E \left[ K \left( \frac{X_1 + x}{h} \right) K \left( \frac{X_1 + x}{h} \right) (m(X_1) + \epsilon_1)^2 \right]
\end{aligned}$$

where  $\epsilon_1 = Y_1 - m(X_1)$ .

$$\begin{aligned}
&= \frac{1}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \left\{ E \left[ K \left( \frac{X_1 + x}{h} \right) K \left( \frac{X_1 + x}{h} \right) m^2(X_1) \right] \right. \\
&\quad \left. + \sigma^2 E \left[ K \left( \frac{X_1 + x}{h} \right) K \left( \frac{X_1 + x}{h} \right) \right] \right\} \\
&= \frac{1}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \int_0^\infty K \left( \frac{t + x}{h} \right) K \left( \frac{t + x}{h} \right) m^2(t) f(t) dt \\
&\quad + \frac{\sigma^2}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \int_0^\infty K \left( \frac{t + x}{h} \right) K \left( \frac{t + x}{h} \right) f(t) dt \\
&= T_1 + T_2.
\end{aligned}$$

I first study  $T_1$  and consider two cases  $x > 0$  and  $x = 0$ .

Case ( $x > 0$ ): Let

$$I_{ij} = \frac{1}{h} \int_0^\infty K \left( \frac{t + x}{h} \right) K \left( \frac{t + x}{h} \right) m^2(t) f(t) dt.$$



Note that  $hT_1 = I_{00} + \sum_{i+j>0} \frac{k_i}{w_i} \frac{k_j}{w_j} I_{ij}$ . Now,

$$\begin{aligned}
\left| I_{00} - m^2(x)f(x) \int_{\mathbb{R}} K^2(u)du \right| &= \left| \int_{-\infty}^{\frac{x}{h}} K^2(u)m^2(x-hu)f(x-hu)du - m^2(x)f(x) \int_{\mathbb{R}} K^2(u)du \right| \\
&= \left| \int_{\mathbb{R}} K^2(u) \left[ m^2(x-hu)f(x-hu) - m^2(x)f(x) \right] du \right. \\
&\quad \left. - \int_{\frac{x}{h}}^{\infty} K^2(u)m^2(x-hu)f(x-hu)du \right| \\
&\leq \left| \int_{\mathbb{R}} K^2(u) \left[ m^2(x-hu)f(x-hu) - m^2(x)f(x) \right] du \right| \\
&\quad + \left| \int_{\frac{x}{h}}^{\infty} K^2(u)m^2(x-hu)f(x-hu)du \right| \\
&\leq \left| \int_{|u|\leq C} K^2(u) \left[ m^2(x-hu)f(x-hu) - m^2(x)f(x) \right] du \right| \\
&\quad + \left| \int_{|u|>C} K^2(u) \left[ m^2(x-hu)f(x-hu) - m^2(x)f(x) \right] du \right| \\
&\quad + \left| \int_{\frac{x}{h}}^{\infty} K^2(u)m^2(x-hu)f(x-hu)du \right|.
\end{aligned}$$

Let  $\bar{p}(\delta, x) = \sup_{|y|\leq\delta} |f(x-y) - f(x)|$ , and since  $f \in C_b^4([0, \infty))$  I have  $f(x-hu) - f(x) \leq \bar{p}(Ch, x)$ .

Thus,

$$\left| \int_{|u|\leq C} K^2(u) \left[ m^2(x-hu)f(x-hu) - m^2(x)f(x) \right] du \right| \leq C\bar{p}(Ch, x) \int_{|u|\leq C} K^2(u)du.$$

Consequently, since  $m \in C_b^4([0, \infty))$ ,

$$\left| I_{00} - m^2(x)f(x) \int_{\mathbb{R}} K^2(u)du \right| \leq C\bar{p}(Ch, x) \int_{|u|\leq C} K^2(u)du + C \int_{|u|>C} K^2(u)du + C \int_{\frac{x}{h}}^{\infty} K^2(u)du.$$

For  $C$  be sufficiently large and  $h, \epsilon$  sufficiently small, by continuity of  $f$ ,  $\bar{p}(Ch, x) < \epsilon$ . Since  $\int_{\mathbb{R}} |K(u)|^2 du < C$ ,  $\int_{|u|>C} K^2(u)du < \epsilon$  and  $\int_{\frac{x}{h}}^{\infty} K^2(u)du < \epsilon$ . Therefore, for all  $\epsilon > 0$ ,

$$\left| I_{00} - m^2(x)f(x) \int_{\mathbb{R}} K^2(u)du \right| \leq \epsilon. \tag{2.9}$$

Having bounded  $I_{00}$ , now I turn attention to  $I_{ij}$  where  $i+j > 0$ , and, without loss of generality,

take  $w_i > 0$ . Change variables by setting  $u = \frac{1}{h}(t + xw_i)$ .

$$I_{ij} = \frac{1}{h} \int_{\frac{xw_i}{h}}^{\infty} K\left(\frac{u}{w_i}\right) K\left(\frac{x}{h}\left(1 - \frac{w_i}{w_j}\right) + \frac{u}{w_j}\right) m^2(hu - xw_i) f(hu - xw_i) du.$$

Given the uniform boundedness of  $K(x)$ ,  $f(x)$  and  $m^2(x)$ ,

$$I_{ij} \leq Cw_i \int_{\frac{x}{h}}^{\infty} K(u) du, \quad (2.10)$$

where  $\int_{\frac{x}{h}}^{\infty} K(u) du < \epsilon$  for sufficiently small  $h$ . Consequently, inequalities (2.9) and (2.10) give

$$\left| hT_1 - m^2(x)f(x) \int_{\mathbb{R}} K^2(u) du \right| \leq \epsilon.$$

I now turn to  $T_2$ . Let  $J_{ij} = \frac{1}{h} \int_0^{\infty} K\left(\frac{t+x}{w_i}\right) K\left(\frac{t+x}{w_j}\right) f(t) dt$ , then  $hT_2 = \sigma^2(J_{00} + \sum_{i+j>0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} J_{ij})$ .

Using arguments similar to those for  $I_{00}$ , I have  $|J_{00} - f(x) \int_{\mathbb{R}} K^2(u) du| \leq \epsilon$ . Again, similar to the case of  $I_{ij}$ , I have  $J_{ij} \leq Cw_i \int_{\frac{x}{h}}^{\infty} K(u) du < \epsilon$  for sufficiently small  $h$ . Thus,

$$\left| hT_2 - \sigma^2 f(x) \int_{\mathbb{R}} K^2(u) du \right| \leq \epsilon. \quad (2.11)$$

Since  $Eu_1^2 = T_1 + T_2$ , I have

$$\left| hEu_1^2 - (m^2(x) + \sigma^2) f(x) \int_{\mathbb{R}} K^2(u) du \right| \leq \epsilon. \quad (2.12)$$

From Theorem 1,  $hE(u_1)^2 = o(1)$ , consequently

$$V(m_H(x)) = \frac{1}{nh} \left\{ \frac{m^2(x) + \sigma^2}{f(x)} \int_{\mathbb{R}} K^2(u) du + o(1) \right\}.$$

Case ( $x = 0$ ): First, I consider  $T_1$ .

$$I_{00} = \frac{1}{h} \int_0^{\infty} K^2\left(\frac{0-t}{h}\right) m^2(t) f(t) dt = \int_{-\infty}^0 K^2(u) m^2(-hu) f(-hu) du,$$

and

$$\begin{aligned}
\left| I_{00} - m^2(0)f(0) \int_{-\infty}^0 K^2(u)du \right| &= \left| \int_{-\infty}^0 K^2(u) \left[ m^2(-hu)f(-hu) - m^2(0)f(0) \right] du \right| \\
&= \left| \int_{-\infty}^{-C} K^2(u) \left[ m^2(-hu)f(-hu) - m^2(0)f(0) \right] du \right. \\
&\quad \left. + \int_{-C}^0 K^2(u) \left[ m^2(-hu)f(-hu) - m^2(0)f(0) \right] du \right| \\
&\leq C \int_{-\infty}^{-C} K^2(u)du + \bar{p}(Ch, 0) \int_0^C K^2(u)du.
\end{aligned}$$

For  $C$  sufficiently large and  $\epsilon, h$  sufficiently small I have

$$\left| I_{00} - m^2(0)f(0) \int_{-\infty}^0 K^2(u)du \right| \leq \epsilon. \tag{2.13}$$

Now I consider the case where  $i + j > 0$ . Note that  $I_{ij} = \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) m^2(hu)f(hu)du$ .

$$\begin{aligned}
\left| I_{ij} - m^2(0)f(0) \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| &= \left| \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) \left[ m^2(hu)f(hu) - m^2(0)f(0) \right] du \right| \\
&= \left| \int_0^C K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) \left[ m^2(hu)f(hu) - m^2(0)f(0) \right] du \right. \\
&\quad \left. + \int_C^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) \left[ m^2(hu)f(hu) - m^2(0)f(0) \right] du \right| \\
&\leq \bar{p}(hC, 0) \int_0^C \left| K\left(\frac{u}{w_i}\right) \right| \left| K\left(\frac{u}{w_j}\right) \right| du \\
&\quad + C \int_C^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du,
\end{aligned}$$

where for sufficiently large  $C$ , and for all  $\epsilon > 0$ ,  $\left| \int_C^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| < \epsilon$  and for sufficient small  $h$ ,  $\bar{p}(hC, 0) < \epsilon$ . Thus,

$$\left| I_{ij} - m^2(0)f(0) \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| < \epsilon. \tag{2.14}$$

Consequently, (2.13) and (2.14) give

$$\left| hT_1 - m^2(0)f(0) \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| < \epsilon.$$

Turning to the term  $T_2$ .

$$hT_2 = \frac{\sigma^2}{h} \int_0^\infty K^2\left(\frac{t}{h}\right) f(t) dt + \frac{\sigma^2}{h} \sum_{i+j>0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_{-\infty}^\infty K\left(\frac{t}{w_i h}\right) K\left(\frac{t}{w_j h}\right) f(t) dt.$$

From the first term,

$$\frac{\sigma^2}{h} \int_0^\infty K^2\left(\frac{t}{h}\right) f(t) dt = \sigma^2 \int_0^\infty K^2(u) f(hu) du = \sigma^2 f(0) \int_0^\infty K^2(u) du + o(1).$$

From the second term,

$$\frac{1}{h} \int_0^\infty K\left(\frac{t}{w_i h}\right) K\left(\frac{t}{w_j h}\right) f(t) dt = \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) f(hu) du.$$

Now,

$$\begin{aligned} & \left| \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) f(hu) du - f(0) \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| \\ & \leq \left| \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) [f(hu) - f(0)] du \right| < \epsilon \end{aligned}$$

by the continuity of  $f$  and the dominated convergence theorem. Thus,

$$\left| hT_2 - \sigma^2 f(0) \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| \leq \epsilon. \quad (2.15)$$

Since  $Eu_1^2 = T_1 + T_2$ , I have

$$\left| hEu_1^2 - (m^2(0) + \sigma^2)f(0) \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right| \leq \epsilon. \quad (2.16)$$

Thus,

$$\begin{aligned} V(m_H(0)) &= \frac{1}{nhf^2(0)} \left( hEu_1^2 - hE^2u_1 \right) \\ &= \frac{1}{nh} \left\{ \frac{m^2(0) + \sigma^2}{f(0)} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du + o(1) \right\}. \end{aligned}$$

Combining the two cases, I have

$$V(m_H(x)) = \begin{cases} \frac{1}{nh} \left\{ \frac{m^2(0) + \sigma^2}{f(0)} \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du + o(1) \right\}, & x = 0 \\ \frac{1}{nh} \left\{ \frac{m^2(x) + \sigma^2}{f(x)} \lambda_0 + o(1) \right\}, & x > 0. \end{cases}$$

□

The expressions for the variance of  $m_H(x)$  given in Theorem 2 are analogous to those obtained in [Mynbaev and Martins-Filho \(2018\)](#) (see their equations (10) and (11)). The following corollary to Theorem 2 gives an expression for the variance of the infeasible NW estimator.

**Corollary 2.** *Suppose  $E((Y - m(X))^2 | X = x) = \sigma^2$ ,  $m, f \in C_b^4([0, \infty))$  and  $K$  satisfies Assumption 1. Then, if  $nh \rightarrow \infty$  and  $x \geq 0$ , the variance of  $m_{NW}(x)$  is given by*

$$V(m_{NW}(x)) = \frac{1}{nh} \frac{m^2(x) + \sigma^2}{f(x)} \eta_{0,x/h} + o((nh)^{-1}).$$

Note that  $m_H$  and  $m_{NW}$  have the same variance at interior points, but different variances at the boundary point ( $x = 0$ ). With suitable choice of  $w_i$  it may be possible to have the leading term of the expression in  $V(m_H(0)) \leq V(m_{NW}(0))$ .

*Remark 1.* An optimal plug-in bandwidth  $h_{pi}$  for  $m_H(0)$  can be obtained by minimizing asymptotic mean squared error (AMSE) at the boundary  $x = 0$ . As such, consider the asymptotic mean squared error (AMSE) given by

$$AMSE(h) = \left\{ \frac{h^2}{2} \left[ m^{(2)}(0) + \frac{2m^{(1)}(0)f^{(1)}(0)}{f(0)} + \frac{f^{(2)}(0)m(0)}{f(0)} \right] \kappa_2 \right\}^2 + \frac{1}{nh} \left\{ \frac{m^2(0) + \sigma^2}{f(0)} \gamma \right\} + s.o.$$

where  $\gamma = \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du$  and  $s.o.$  denotes terms of smaller order. Routine optimization of the leading terms with respect to  $h$  gives

$$h_{pi} = n^{-\frac{1}{5}} \left\{ \frac{m^2(0) + \sigma^2}{f(0)} \gamma \right\}^{\frac{1}{5}} \left[ m^{(2)}(0) + \frac{2m^{(1)}(0)f^{(1)}(0)}{f(0)} + \frac{f^{(2)}(0)m(0)}{f(0)} \right]^{-\frac{2}{5}} \kappa_2^{-\frac{2}{5}}. \quad (2.17)$$

### 2.1.3 Feasible Hestenes Regression Estimators

The regression model that motivated the infeasible  $m_H$  had regressors taking values in  $[0, \infty)$ . It is apparent that a similar estimator can be defined for the case where regressors take values in, and the support of  $m$  is,  $(-\infty, 0]$ . When the regressor takes values in  $\mathbb{R}$  and there are potentially two regressions, one to the right and one to the left of a discontinuity at  $x = 0$ , two infeasible Hestenes estimators can be constructed. The first, for the regression to the right of the point of discontinuity,

$$m_H^+(x) \equiv \frac{1}{f(x)} \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] Y_i d_i \text{ for } x \geq 0, \quad (2.18)$$

where  $d_i = I_{X_i \geq 0}$ , and the second for the regression to the left of the point of discontinuity

$$m_H^-(x) \equiv \frac{1}{f(x)} \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] Y_i (1 - d_i) \text{ for } x < 0. \quad (2.19)$$

In a RD model where the point of discontinuity is  $x = 0$ , an estimator for the jump at 0, denoted by  $J_H(0)$ , is naturally given by

$$J_H(0) = m_H^+(0) - m_H^-(0). \quad (2.20)$$

Since  $f$  is unknown, these are infeasible estimators. I define the feasible estimators by replacing  $f$  with the estimated  $\hat{f}$ . The feasible estimators are denoted by  $\hat{m}_H^+(x)$ ,  $\hat{m}_H^-(x)$  and  $\hat{J}_H(0)$ .

RDD assume that the density  $f$  of the running variable is continuous at the point of discontinuity, so it is quite natural to use the whole sample to estimate  $f$  by the Rosenblatt-Parzen estimator

$\hat{f}_R(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)$ . I define

$$\hat{m}_H^+(x) \equiv \frac{\sum_{i=1}^n \left[ K\left(\frac{X_i-x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j}+x}{h}\right) \right] Y_i d_i}{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)} \text{ for } x \geq 0.$$

Similarly, I define  $\hat{m}_H^-(x)$  for  $x < 0$  by changing  $d_i$  to  $1 - d_i$ . Theorem 3 and 4 provide bias and variance of the estimators.

Another option to obtain  $\hat{f}$  is to use the Hestenes density estimator developed by [Mynbaev and Martins-Filho \(2018\)](#)  $\hat{f}_H(x) = \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i-x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j}+x}{h}\right) \right]$ , where the observations to the right of the discontinuous point are used to estimate  $f_H^+$  for  $x \geq 0$  and the observations to the left to estimate  $f_H^-$  for  $x < 0$ . Consequently, I define

$$\hat{m}_{HT}^+(x) \equiv \frac{\sum_{i=1}^n \left[ K\left(\frac{X_i-x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j}+x}{h}\right) \right] Y_i d_i}{\sum_{i=1}^n \left[ K\left(\frac{X_i-x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j}+x}{h}\right) \right] d_i} \text{ for } x \geq 0. \quad (2.21)$$

While  $\hat{f}_R(x)$  is consistent only at interior points  $x > 0$ ,  $\hat{f}_H^+(x)$  is consistent at both interior and boundary points  $x \geq 0$  as shown in Remark 3 of [Mynbaev and Martins-Filho \(2018\)](#). To obtain the asymptotic properties of the  $\hat{m}_H^+$  or  $\hat{m}_{HT}^+$ , I need  $\hat{f}_R \xrightarrow{p} f$  or  $\hat{f}_H^+ \xrightarrow{p} f$ . Since I do not have consistency of  $\hat{f}_R$  at the boundary, I provide the asymptotic properties of  $\hat{m}_{HT}^+$  at the boundary and interior points in Theorem 5. This property is useful in guiding the simulations described in the next section, where I construct the Hestenes estimator  $\hat{m}_{HT}^+$  rather than  $\hat{m}_H^+$  because  $\hat{m}_{HT}^+$  has better finite sample performance  $\hat{m}_H^+$ . Similarly, I define  $\hat{m}_{HT}^-(x)$  for  $x < 0$  by changing  $d_i$  to  $1 - d_i$ .

**Theorem 3.** *Suppose that  $K$  satisfies assumption 1,  $f \in C_b^4(\mathbb{R})$  and  $m^+ \in C_b^4([0, \infty))$ . Then, for  $x \geq 0$ , the bias of  $\hat{m}_H^+(x)$  is given by*

$$E(\hat{m}_H^+(x)) - m^+(x) = \left[ \frac{1}{2} m^{+(2)}(x) + \frac{m^{+(1)}(x) f^{(1)}(x)}{f(x)} \right] \kappa_2 h^2 + O(h^4 + (nh)^{-1}). \quad (2.22)$$

*Proof.* Given that  $K$  satisfies assumption 1 and  $f(x) > 0$ , for  $n$  sufficiently large  $\hat{f}(x) > 0$  and

$E(\hat{f}(X)) > 0$ . Thus, using Taylor's Theorem, I expand  $\hat{m}_H^+(x) \equiv q(\hat{g}(x), \hat{f}(x)) = \frac{\hat{g}(x)}{\hat{f}(x)}$  at the point  $(E(\hat{g}(x)), E(\hat{f}(x)))$  and obtain,

$$\begin{aligned} \hat{m}_H^+(x) &= \frac{E(\hat{g}(x))}{E(\hat{f}(x))} + \frac{1}{E(\hat{f}(x))}(\hat{g}(x) - E(\hat{g}(x))) - \frac{E(\hat{g}(x))}{E(\hat{f}(x))^2}(\hat{f}(x) - E(\hat{f}(x))) \\ &\quad - \left\{ \frac{1}{E(\hat{f}(x))^2} (\hat{f}(x) - E(\hat{f}(x))) (\hat{g}(x) - E(\hat{g}(x))) - \frac{E(\hat{g}(x))}{E(\hat{f}(x))^3} (\hat{f}(x) - E(\hat{f}(x)))^2 \right\} + Z_n, \end{aligned}$$

where  $\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{\frac{X_i}{w_j} + x}{h}\right) \right] Y_i d_i$  and

$$\begin{aligned} Z_n(x) &= 3(\hat{g}(x) - E\hat{g}(x))(\hat{f}(x) - E\hat{f}(x))^2 \int_0^1 (1-t)^2 \frac{1}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^3} dt \\ &\quad - 3(\hat{f}(x) - E\hat{f}(x))^3 \int_0^1 (1-t)^2 \frac{E\hat{g}(x) + t(\hat{g}(x) - E\hat{g}(x))}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt. \end{aligned}$$

Taking the expectations on both sides of the expression for  $\hat{m}_H^+(x)$  gives,

$$E(\hat{m}_H^+(x)) = \frac{E(\hat{g}(x))}{E(\hat{f}(x))} - \frac{1}{E^2(\hat{f}(x))} Cov(\hat{g}(x), \hat{f}(x)) + \frac{E(\hat{g}(x))}{E^3(\hat{f}(x))} V(\hat{f}(x)) + E(Z_n(x)).$$

Now, from Theorem 1, I have

$$\begin{aligned} E(\hat{g}(x)) &= f(x)E(m_H^+(x)) \\ &= f(x)m^+(x) + \frac{h^2}{2}\kappa_2 \left[ f(x)m^{+(2)}(x) + 2m^{(1)}(x)f^{(1)}(x) + f^{(2)}(x)m^+(x) \right] + O(h^4). \end{aligned}$$

From standard results for kernel density estimators (see [Li and Racine, 2007](#))

$$E(\hat{f}(x)) = f(x) + \frac{h^2}{2}\kappa_2 f^{(2)}(x) + O(h^4).$$

Thus,

$$\frac{E(\hat{g}(x))}{E(\hat{f}(x))} = m^+(x) + \frac{h^2}{2}\kappa_2 \left[ m^{+(2)}(x) + 2\frac{m^{+(1)}(x)f^{(1)}(x)}{f(x)} \right] + O(h^4).$$

Now, from Lemma 1,  $Cov(\hat{f}(x), \hat{g}(x)) = O(\frac{1}{nh})$ , and from standard results for kernel density estimators [Li and Racine \(2007\)](#),  $Var(\hat{f}(x)) = O(\frac{1}{nh})$ . Lastly, from Lemma 2,  $E(Z_n(x)) = O((nh)^{-3/2})$ .



Thus,

$$E(\hat{m}_H^+(x)) - m^+(x) = \left[ \frac{1}{2}m^{+(2)}(x) + \frac{m^{+(1)}(x)f^{(1)}(x)}{f(x)} \right] \kappa_2 h^2 + O(h^4 + (nh)^{-1}).$$

□

**Theorem 4.** *Suppose that  $K$  satisfies Assumption 1,  $f \in C_b^4(\mathbb{R})$  and  $m^+ : [0, \infty) \rightarrow \mathbb{R}$  is such that  $m^+ \in C_b^4([0, \infty))$ . Then, for  $x \geq 0$ , the variance of  $\hat{m}_H^+(x)$  is given by*

$$V(\hat{m}_H^+(x)) = \begin{cases} \frac{1}{nh} \left\{ \frac{m^{+(2)}(0)+\sigma^2}{f(0)} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du \right. \\ \left. + \frac{m^{+(2)}(0)}{f(0)} \int_{\mathbb{R}} K^2(u) du + \frac{2m^{+(2)}(0)}{f(0)} \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right\}, & \text{if } x = 0 \cdot \\ \frac{1}{nh} \left\{ \frac{4m^{+(2)}(x)+\sigma^2}{f(x)} \int_{\mathbb{R}} K^2(u) du + o(1) \right\}, & \text{if } x > 0 \end{cases}$$

*Proof.* Write  $a = E(\hat{g}(x))$  and  $b = E(\hat{f}(x))$ . Then,  $\hat{m}_H^+(x) = \frac{\hat{g}(x)}{\hat{f}(x)} = \frac{a}{b} + \frac{1}{b}(\hat{g}(x) - a) - \frac{a}{b^2}(\hat{f}(x) - b) + S_n(x)$  where  $S_n(x) = 2(\hat{g}(x) - a)(\hat{f}(x) - b) \int_0^1 (1-t)(-1) \frac{1}{[b+t(\hat{f}(x)-b)]^2} dt + (\hat{f}(x) - b)^2 \int_0^1 (1-t) \frac{2[a+t(\hat{g}(x)-a)]}{[b+t(\hat{f}(x)-b)]^3} dt$ . Then,  $E(\hat{m}_H^+(x)) = \frac{a}{b} + E(S_n)$  and  $V(\hat{m}_H^+(x)) = \frac{1}{b^2}V(\hat{g}(x)) + \frac{a^2}{b^4}V(\hat{f}(x)) - \frac{2a}{b^3}Cov(\hat{g}(x), \hat{f}(x)) + W_n(x)$  where  $W_n(x) = V(S_n) + \frac{2}{b}Cov(\hat{g}(x), S_n) - \frac{2a}{b^2}Cov(\hat{f}(x), S_n)$ . Now,

$$\frac{1}{b^2}V(\hat{g}(x)) = V(m_H(x)) = \begin{cases} \frac{1}{nh} \left\{ \frac{m^{+(2)}(0)+\sigma^2}{f(0)} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du + o(1) \right\}, & x = 0 \\ \frac{1}{nh} \left\{ \frac{m^{+(2)}(x)+\sigma^2}{f(x)} \lambda_0 + o(1) \right\}, & x > 0 \end{cases}.$$

From the properties of the Rosenblatt-Parzen estimator  $\hat{f}$ , I have  $\frac{a^2}{b^4}V(\hat{f}(x)) = \frac{1}{nh} \left\{ \frac{m^{+(2)}(x)}{f(x)} \lambda_0 + o(1) \right\}$ .

From Lemma 1,

$$Cov(\hat{g}(x), \hat{f}(x)) = \begin{cases} \frac{1}{nh} \left\{ m^+(0)f(0) \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right\}, & x = 0 \\ \frac{1}{nh} \left\{ m^+(x)f(x) \int_{\mathbb{R}} K^2(u) du + o(1) \right\}, & x > 0 \end{cases},$$

and consequently,

$$-\frac{2a}{b^3} \text{Cov}(\hat{g}(x), \hat{f}(x)) = \begin{cases} \frac{1}{nh} \left\{ \frac{2m^2(0)}{f(0)} \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right\}, & x = 0 \\ \frac{1}{nh} \left\{ \frac{2m^2(x)}{f(x)} \int_{\mathbb{R}} K^2(u) du + o(1) \right\}, & x > 0. \end{cases}$$

Finally, using Lemma 2 I obtain  $W_n(x) = O\left(\left(\frac{1}{nh}\right)^{\frac{3}{2}}\right)$ . Thus,

$$V(\hat{m}_H^+(x)) = \begin{cases} \frac{1}{nh} \left( \frac{m^{+2}(0) + \sigma^2}{f(0)} \sum_{i,j=0}^{s+1} \frac{k_i}{w_i} \frac{k_j}{w_j} \int_0^\infty K\left(\frac{u}{w_i}\right) K\left(\frac{u}{w_j}\right) du + \frac{m^{+2}(0)}{f(0)} \lambda_0 \right. \\ \left. + \frac{2m^2(0)}{f(0)} \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right), & x = 0. \\ \frac{1}{nh} \left( \frac{4m^2(x) + \sigma^2}{f(x)} \lambda_0 + o(1) \right), & x > 0 \end{cases} \quad (2.23)$$

□

As in the case of bias, a similar expression for the variance of  $\hat{m}_H^-(x)$  can be obtained with the only change being that  $x \leq 0$ .

The next theorem gives asymptotic normality of  $\hat{m}_{HT}^+(x)$  for  $x \geq 0$ .

**Theorem 5.** *Suppose that  $K$  satisfies Assumption 1,  $f \in C_b^4(\mathbb{R})$  and  $m^+ : [0, \infty) \rightarrow \mathbb{R}$  is such that  $m^+ \in C_b^4([0, \infty))$ . If  $E\left(|(Y_i - m^+(X_i))d_i|^{2+\delta} | X\right) < \infty$ . Then, for  $x \geq 0$ ,*

$$(nh)^{\frac{1}{2}} \left\{ \hat{m}_{HT}^+(x) - E(\hat{m}_{HT}^+(x) | X_1, \dots, X_n) \right\} \xrightarrow{d} N(0, c/f^2(x)) \quad (2.24)$$

$$\text{where } c = \begin{cases} \sigma^2 f(0) \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du, & \text{if } x = 0 \\ \sigma^2 f(x) \lambda_0, & \text{if } x > 0. \end{cases}$$

*Proof.* Let  $w_0 = -1$ ,  $k_0 = -1$ ,  $r_i(x) = \sum_{j=0}^{s+1} \frac{k_j}{w_j} K\left(\frac{X_i + x}{h}\right)$ . It is intuitive to think  $r_i$  as a reflection kernel – a kernel for reflection estimators. Then,  $\hat{m}_{HT}^+(x) = \frac{\hat{g}(x)}{\hat{f}_H(x)} = \frac{\sum_{i=1}^n r_i(x) Y_i d_i}{\sum_{i=1}^n u_i}$ .  $\hat{f}_H(x) \xrightarrow{p} f(x)$ , thus I am concerned with the convergence in distribution of  $\hat{g}(x)$ . Note that

$$E(\hat{g}(x) | X_1, \dots, X_n) = (nh)^{-1} \sum_{i=1}^n r_i(x) m^+(X_i) d_i$$

and

$$\hat{g}(x) - E(\hat{g}(x)|X_1, \dots, X_n) = (nh)^{-1} \sum_{i=1}^n r_i(x)(Y_i - m^+(X_i))d_i.$$

Let  $Z_{in} = \frac{r_i(x)(Y_i - m^+(X_i))d_i}{nh}$  and note that  $E(Z_{in}) = 0$  and

$$V(Z_{in}) = E(Z_{in}^2) = \frac{\sigma^2}{(nh)^2} E(r_i(x)^2 d_i) = \frac{\sigma^2}{(nh)^2} \int_0^\infty r_1^2(x) f(X_1) dX_1.$$

Now, let

$$S_n^2 = \sum_{i=1}^n E(Z_{in}^2) = \frac{1}{nh} \frac{\sigma^2}{h} \int_0^\infty r_1^2(x) f(X_1) dX_1$$

and

$$X_{in} = \frac{Z_{in}}{S_n} = \frac{r_i(x)(Y_i - m^+(X_i))d_i}{(nh)^{\frac{1}{2}} \left( \frac{\sigma^2}{h} \int_0^\infty r_1^2(x) f(X_1) dX_1 \right)^{\frac{1}{2}}}.$$

Consequently,  $\sum_{i=1}^n X_{in} = 1$  and by Liapounov's Central Limit Theorem  $\sum_{i=1}^n X_{in} \xrightarrow{d} N(0, 1)$  provided that  $\lim_{n \rightarrow \infty} \sum_{i=1}^n E(|X_{in}|^{2+\delta}) = 0$  for some  $\delta > 0$ . Note that

$$|X_{in}| = \frac{|r_i(x)(Y_j - m^+(X_j))d_i|}{(nh)^{\frac{1}{2}} c(n)^{\frac{1}{2}}}$$

with

$$c(n) = \frac{\sigma^2}{h} \int_0^\infty r_1^2(x) f(X_1) dX_1$$

and

$$|X_{in}|^{2+\delta} = \frac{|u_i(Y_j - m^+(X_j))d_i|^{2+\delta}}{(nh)^{\frac{2+\delta}{2}} c(n)^{\frac{2+\delta}{2}}}.$$

$c(n)$  is non-stochastic. Therefore,

$$E(|X_{in}|^{2+\delta}) = (nhc(n))^{-1-\frac{\delta}{2}} E(|r_i(x)|^{2+\delta} \left| [Y_i - m^+(X_i)] d_i \right|^{2+\delta})$$

and

$$\sum_{i=1}^n E(|X_{in}|^{2+\delta}) = (nhc(n))^{-1-\frac{\delta}{2}} \sum_{i=1}^n E(|r_i(x)|^{2+\delta} \left| [Y_i - m^+(X_i)] d_i \right|^{2+\delta}).$$

Now, if

$$E \left( \left| [Y_i - m^+(X_i)] d_i \right|^{2+\delta} \mid X_i \right) < C < \infty,$$

then

$$\begin{aligned} E(|r_i|^{2+\delta} \left| [Y_i - m^+(X_i)] d_i \right|^{2+\delta}) &= E \left[ |r_i d_i|^{2+\delta} E \left( \left| [Y_i - m^+(X_i)] \right|^{2+\delta} \mid X_i \right) \right] \\ &\leq C \int_0^\infty r_1^{2+\delta}(x) f(X_1) dX_1. \end{aligned}$$

Consequently,

$$\sum_{i=1}^n E(|X_{in}|^{2+\delta}) \leq (nh)^{-\frac{\delta}{2}} (c(n))^{-1-\frac{\delta}{2}} C \frac{1}{h} \int_0^\infty |r_1(x)|^{2+\delta} f(X_1) dX_1.$$

Note that  $c(n) = hT_2$  in Theorem 2, thus I have for  $x > 0$ ,  $c(n) \rightarrow \sigma^2 f(x) \int_{\mathbb{R}} K^2(u) du$  from (2.11).

For  $x = 0$ ,  $c(n) \rightarrow \sigma^2 f(0) \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K \left( \frac{u}{w_i} \right) \right]^2 du$  from (2.15). By the  $c_r$ -Inequality

$$\begin{aligned} E \left| \sum_{j=0}^{s+1} \frac{k_j}{w_j} K \left( \frac{\frac{\alpha}{w_j} + x}{h} \right) \right|^{2+\delta} &\leq (s+2)^{1+\delta} \sum_{j=0}^{s+1} E \left| \frac{k_j}{w_j} K \left( \frac{\frac{\alpha}{w_j} + x}{h} \right) \right|^{2+\delta} \\ &= (s+2)^{1+\delta} h \sum_{j=0}^{s+1} \frac{1}{h} E \left| \frac{k_j}{w_j} K \left( \frac{\frac{\alpha}{w_j} + x}{h} \right) \right|^{2+\delta} \\ &= (s+2)^{1+\delta} h \sum_{j=0}^{s+1} \frac{1}{h} \int_0^\infty \left| \frac{k_j}{w_j} K \left( \frac{\frac{\alpha}{w_j} + x}{h} \right) \right|^{2+\delta} f(\alpha) d\alpha. \end{aligned}$$

Changing variable by setting  $u = \frac{\frac{\alpha}{w_j} + x}{h}$ ,

$$\frac{1}{h} E \left| \frac{k_j}{w_j} K \left( \frac{\frac{\alpha}{w_j} + x}{h} \right) \right|^{2+\delta} = |k_j|^{2+\delta} \int_{\frac{x}{h}}^\infty |K(u)|^{2+\delta} f(w_j(hu - x)) du.$$

For  $x > 0$ , since  $f$  is bounded,  $K$  satisfies assumption 1.

$$\int_{\frac{x}{h}}^\infty |K(u)|^{2+\delta} f(w_j(hu - x)) du \leq C \int_{\frac{x}{h}}^\infty |K(u)|^{2+\delta} du \leq \epsilon,$$

for sufficiently small  $h$ . For  $x = 0$ , and  $C > 0$

$$\begin{aligned}
& \left| \int_0^\infty |K(u)|^{2+\delta} f(w_j hu) du - f(0) \int_0^\infty |K(u)|^{2+\delta} du \right| \\
&= \left| \int_0^C |K(u)|^{2+\delta} [f(w_j hu) - f(0)] du + \int_C^\infty |K(u)|^{2+\delta} [f(w_j hu) - f(0)] du \right| \\
&\leq \bar{p}(w_j hC, 0) \int_0^C |K(u)|^{2+\delta} du + 2 \sup(f) \int_C^\infty |K(u)|^{2+\delta} du \\
&\leq \epsilon,
\end{aligned}$$

for sufficiently small  $h$ . Now, given that

$$\int_0^\infty |K(u)|^{2+\delta} f(w_j hu) du \rightarrow f(0) \int_0^\infty |K(u)|^{2+\delta} du,$$

since  $nh_n \rightarrow \infty$  I have that  $\lim_{n \rightarrow \infty} \sum_{i=1}^n E(|X_{in}|^{2+\delta}) = 0$ . Hence,

$$\frac{(nh)^{\frac{1}{2}} (\hat{g}(x) - E(\hat{g}(x)|X_1, \dots, X_n))}{c(n)^{\frac{1}{2}}} \xrightarrow{d} N(0, 1),$$

which implies that

$$(nh_n)^{\frac{1}{2}} (\hat{g}(x) - E(\hat{g}(x)|X_1, \dots, X_n)) \xrightarrow{d} N(0, c),$$

where

$$c = \lim_{n \rightarrow \infty} c(n) = \begin{cases} \sigma^2 f(0) \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du, & \text{if } x = 0 \\ \sigma^2 f(x) \lambda_0, & \text{if } x > 0. \end{cases}$$

Thus,

$$(nh)^{\frac{1}{2}} \left\{ \hat{m}_{HT}^+(x) - E(\hat{m}_{HT}^+(x)|X_1, \dots, X_n) \right\} \xrightarrow{d} N(0, c/f^2(x)).$$

□

Next theorem gives the order of the conditional mean of  $\hat{m}_{HT}^+$ .

**Theorem 6.** *Suppose that  $K$  satisfies Assumption 1,  $f \in C_b^4([0, \infty))$  and  $m \in C_b^4([0, \infty))$ . Then,*

for  $x \geq 0$ ,

$$\begin{aligned} & E(\hat{m}_{HT}^+(x)|X_1, \dots, X_n) \\ &= m(x) + \left( \frac{1}{2}m^{(2)}(x) + \frac{m^{(1)}(x)f^{(1)}(x)}{f(x)} \right) \kappa_2 h^2 + o_p(h^2) + O_p\left(\frac{h}{(nh)^{1/2}} + h^4\right) \end{aligned}$$

*Proof.* First, note that

$$E(\hat{m}_{HT}^+(x)|X_1, \dots, X_n) = E\left(\frac{\hat{g}(x)}{\hat{f}(x)}|X_1, \dots, X_n\right) = \frac{\frac{1}{nh} \sum_{i=1}^n r_i(x)m(X_i)d_i}{\hat{f}(x)},$$

where  $\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n r_i(x)d_i$ . Since  $m(x) = \frac{\frac{1}{nh} \sum_{i=1}^n r_i(x)m(x)d_i}{\hat{f}(x)}$ ,

$$E(\hat{m}_{HT}^+(x)|X_1, \dots, X_n) - m(x) = \frac{1}{\hat{f}(x)} \frac{1}{nh} \sum_{i=1}^n r_i(x) (m(X_i) - m(x)) d_i$$

Let  $J_n = \frac{1}{nh} \sum_{i=1}^n r_i(x) (m(X_i) - m(x)) d_i$ . Then,  $E(\hat{m}_{HT}^+(x)|X_1, \dots, X_n) - m^+(x) = \frac{J_n}{\hat{f}(x)}$ , and

$$\begin{aligned} E(J_n(x)) &= \frac{1}{h} E[r_1(x) (m(X_1) - m(x)) d_1] \\ &= \int_{-\infty}^{\infty} K(\psi) \mu(x - h\psi) d\psi - m(x) \int_{-\infty}^{\infty} K(\psi) \delta(x - h\psi) d\psi, \end{aligned}$$

where  $\delta(x) \equiv \begin{cases} f(x), & x \geq 0 \\ \sum_{j=1}^{s+1} k_j f(-w_j x), & x < 0 \end{cases}$ . The first term comes equation (2.7), which is the integral

representation of the bias of  $m_h$  from Theorem 1. The second term comes from Equation (2.6) in [Mynbaev and Martins-Filho \(2018\)](#), which is the integral representation of the bias of  $\hat{f}_h$  from their theorem 2.1. Using Taylor expansion.

$$\begin{aligned}
E(J_n(x)) &= \int_{-\infty}^{\infty} K(\psi) \left( \mu(x) + \frac{1}{2}\mu^{(2)}(x)(-h\psi)^2 + \frac{1}{24}\mu^{(4)}(\bar{x})(-h\psi)^4 \right) d\psi \\
&\quad - m(x) \int_{-\infty}^{\infty} K(\psi) \left( \delta(x) + \frac{1}{2}\delta^{(2)}(x)(-h\psi)^2 + \frac{1}{24}\delta^{(4)}(\bar{x})(-h\psi)^4 \right) d\psi \\
&= \mu(x) + \frac{1}{2}\mu^{(2)}(x)\kappa_2h^2 + \frac{1}{24} \int_{-\infty}^{\infty} K(\psi)\psi^4\mu^{(4)}(\bar{x})d\psi \\
&\quad - m(x) \left( \delta(x) + \frac{1}{2}\delta^{(2)}(x)\kappa_2h^2 + \frac{1}{24} \int_{-\infty}^{\infty} K(\psi)\psi^4\delta^{(4)}(\bar{x})d\psi \right).
\end{aligned}$$

For  $x \geq 0$ ,  $\mu(x) = f(x)m(x)$  and  $\delta(x) = f(x)$ . Then,

$$\begin{aligned}
E(J_n(x)) &= f(x)m(x) + \frac{1}{2} \left[ f(x)m^{(2)}(x) + 2m^{(1)}(x)f^{(1)}(x) + f^{(2)}(x)m(x) \right] \kappa_2h^2 \\
&\quad + \frac{1}{24} \int_{-\infty}^{\infty} K(\psi)\psi^4\mu^{(4)}(\bar{x})d\psi \\
&\quad - m(x) \left( f(x) + \frac{1}{2}f^{(2)}(x)\kappa_2h^2 + \frac{1}{24} \int_{-\infty}^{\infty} K(\psi)\psi^4\delta^{(4)}(\bar{x})d\psi \right) \\
&= \frac{1}{2} \left[ f(x)m^{(2)}(x) + 2m^{(1)}(x)f^{(1)}(x) \right] \kappa_2h^2 + O(h^4).
\end{aligned}$$

Let  $B(x) = \frac{1}{2} \left[ f(x)m^{(2)}(x) + 2m^{(1)}(x)f^{(1)}(x) \right] \kappa_2$ . Then  $E(J_n(x)) = B(x)h^2 + O(h^4)$ .  $B(x)$  is the coefficient of the leading term of  $J_n$  of the order  $O(h^2)$ . I next obtain the order of the remainder term of  $J_n$ .

$J_n = \frac{1}{nh} \sum_{i=1}^n r_i(x) (m(X_i) - m(x)) d_i$ . Let  $v_i = \frac{1}{h} r_i(x) (m(X_i) - m(x)) d_i$ . Then  $J_n = \frac{1}{n} \sum_{i=1}^n v_i$  and  $V(J_n) = \frac{1}{n} V(v_1) = \frac{1}{n} (E(v_1^2) - E^2(v_1))$ .

$$\begin{aligned}
E(v_1^2) &= E \left[ \frac{1}{h} r_1 (m(X_1) - m(x)) d_1 \right]^2 \\
&= \frac{1}{h^2} E \left[ r_1^2 (m(X_1) - m(x))^2 d_1 \right] \\
&= \frac{1}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} E \left[ K \left( \frac{X_1 + x}{w_i h} \right) K \left( \frac{X_1 + x}{w_j h} \right) (m(X_1) - m(x))^2 d_1 \right] \\
&= \frac{1}{h^2} \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_0^\infty K \left( \frac{X_1 + x}{w_i h} \right) K \left( \frac{X_1 + x}{w_j h} \right) (m(X_1) - m(x))^2 f(X_1) dX_1 \\
&= \frac{1}{h} \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_{-\frac{x}{h}}^\infty K \left( \left(1 + \frac{1}{w_i}\right)x + u \right) K \left( \left(1 + \frac{1}{w_i}\right)x + u \right) \\
&\quad (m(x + hu) - m(x))^2 f(x + hu) du.
\end{aligned}$$

By the Mean Value Theorem, for  $\lambda \in (0, 1)$ ,  $m(x + hu) - m(x) = m^{(1)}(x + \lambda hu)hu$ . Hence,

$$\begin{aligned}
E(v_1^2) &= h \sum_{i,j=0}^{s+1} \frac{k_i k_j}{w_i w_j} \int_{-\frac{x}{h}}^\infty u^2 K \left( \left(1 + \frac{1}{w_i}\right)x + u \right) \\
&\quad K \left( \left(1 + \frac{1}{w_i}\right)x + u \right) \left( m^{(1)}(x + \lambda hu) \right)^2 f(x + hu) du
\end{aligned}$$

Given that  $m, f \in \mathcal{C}_b^4$ , by Lebesgue's Dominated Convergence Theorem, as  $h \rightarrow 0$ ,

$$\int_{-\frac{x}{h}}^\infty u^2 K \left( \left(1 + \frac{1}{w_i}\right)x + u \right) \left( m^{(1)}(x + \lambda hu) \right)^2 f(x + hu) du \rightarrow \left( m^{(1)}(x) \right)^2 f(x) \int u^2 K^2(u) du$$

and  $\frac{E(v_1^2)}{h} = O(1)$ . Since  $E(v_1) = E(J_n) = B(x)h^2 + O(h^4) = O(h^2)$ ,  $V(J_n(x)) = O\left(\frac{h}{n}\right)$ .

Let  $\mathcal{I}_n = J_n - h^2 B(x)$ . Then

$$\mathcal{I}_n^2 = \left( J_n - E(J_n) + E(J_n) - h^2 B(x) \right)^2,$$

and

$$\begin{aligned}
E(\mathcal{I}_n^2) &= V(J_n) - \left( E(J_n) - h^2 B(x) \right)^2 \\
&= O\left(\frac{h}{n}\right) + O(h^8).
\end{aligned}$$



Thus, by Markov's Inequality,

$$\mathcal{I}_n = O_p \left( \left( \frac{h}{n} + h^8 \right)^{1/2} \right) = O_p \left( \frac{h}{(nh)^{1/2}} + h^4 \right).$$

$$J_n = h^2 B(x) + \mathcal{I}_n.$$

$$\begin{aligned} E(\hat{m}_{HT}(x) | X_1, \dots, X_n) &= m(x) + \frac{1}{\hat{f}(x)} J_n = m(x) + \frac{J_n}{\hat{f}(x)} - \frac{J_n}{f(x)} + \frac{J_n}{f(x)} \\ &= m(x) + \left( \frac{1}{\hat{f}(x)} - \frac{1}{f(x)} \right) J_n + \frac{J_n}{f(x)} \\ &= m(x) + \frac{1}{f(x)} (1 + o_p(1)) J_n \\ &= m(x) + \frac{1}{f(x)} (1 + o_p(1)) (h^2 B(x) + \mathcal{I}_n) \\ &= m(x) + \frac{h^2 B(x)}{f(x)} + o_p(h^2) + (1 + o_p(1)) \frac{\mathcal{I}_n}{f(x)} \\ &= m(x) + \frac{h^2 B(x)}{f(x)} + o_p(h^2) + O_p \left( \frac{h}{(nh)^{1/2}} + h^4 \right). \end{aligned}$$

□

*Remark 2.* Note that

$$\begin{aligned} &\sqrt{nh} \left[ \hat{m}_{HT}^+(x) - E(\hat{m}_{HT}^+(x) | X_1, \dots, X_n) \right] \\ &= \sqrt{nh} \left[ \hat{m}_{HT}^+(x) - \left( m(x) + \frac{B(x)h^2}{f(x)} + o_p(h^2) \right) \right] + O_p(h + \sqrt{nh^9}) \end{aligned}$$

If  $\sqrt{nh^9} \rightarrow 0$ , then  $O_p(h + \sqrt{nh^9}) = o_p(1)$ . Thus,

$$\sqrt{nh} \left[ \hat{m}_{HT}^+(x) - \left( m(x) + \frac{B(x)h^2}{f(x)} + o_p(h^2) \right) \right] \xrightarrow{d} N(0, c/f^2(x)).$$

An optimal bandwidth can be obtained by minimizing asymptotic weighted mean integrated squared error (AWMISE). I will define an optimal  $h_{pi}$  for  $\hat{m}_H^+(0)$  which is obtained by minimizing

$$AWMISE = \int_0^\infty \left\{ [B(0)h^2]^2 + \frac{1}{nh} \frac{\sigma^2 \delta}{f(0)} \right\} dx + s.o.,$$

where  $\delta = \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du$  and *s.o.* denotes terms of smaller orders. Then, routine optimization gives

$$h_{pi} = \left( \frac{1}{4n} \right)^{\frac{1}{5}} \left\{ \frac{\int_0^\infty \frac{\sigma^2 \delta}{f(0)} dx}{\int_0^\infty B(0) dx} \right\}^{\frac{1}{5}} = n^{-\frac{1}{5}} C,$$

$$\text{where } C = \left\{ \frac{\int_0^\infty \sigma^2 \delta dx}{4 \int_0^\infty [f(x)m^{+(2)}(x) + 2m^{+(1)}(x)f^{(1)}(x)] dx} \right\}^{\frac{1}{5}}.$$

#### 2.1.4 A Comparison with Local Linear Estimators

Local linear (LL) estimators are the most commonly used estimators for nonparametric regression in RD models. Estimation is normally conducted by selecting a uniform kernel  $K$  and a bandwidth that in effect constrains the estimation to subsamples of  $\{Y_i, X_i\}_{i=1}^n$  to the right ( $X_i \geq 0$ ) and to the left ( $X_i < 0$ ) of the point of discontinuity  $x = 0$ . Hence, two local linear estimators are obtained  $\hat{m}_{LL}^+(x)$  and  $\hat{m}_{LL}^-(x)$ .

Letting  $Z_i(x) = \begin{pmatrix} 1 & X_i - x \end{pmatrix}$ ,  $Z(x)' = \begin{pmatrix} Z_1(x)' & \dots & Z_n(x)' \end{pmatrix}$ ,  $K_{ix} = K\left(\frac{X_i - x}{h}\right)$ ,  $K(x) = \text{diag}\{K_{ix}\}_{i=1}^n$ , and  $Y' = \begin{pmatrix} Y_1 & \dots & Y_n \end{pmatrix}$ , only the observations  $\{(X_i, Y_i) : X_i \geq 0\}_{i=1}^n$  are used to estimate  $m^+(x)$  for  $x \geq 0$ .

$$\hat{m}_{LL}^+(x) = \begin{pmatrix} 1 & 0 \end{pmatrix} \left( Z(x)' K(x) Z(x) \right)^{-1} Z(x)' K(x) Y, \quad (2.25)$$

where  $X_i$  takes values in  $[0, +\infty)$ . Similarly, observations  $\{(X_i, Y_i) : X_i < 0\}_{i=1}^n$  are used to estimate  $m^-(x)$  where  $x < 0$ .

$$\hat{m}_{LL}^-(x) = \begin{pmatrix} 1 & 0 \end{pmatrix} \left( Z(x)' K(x) Z(x) \right)^{-1} Z(x)' K(x) Y, \quad (2.26)$$

where  $X_i$  takes values in  $(-\infty, 0]$ . Expressions for the conditional bias and variance of  $\hat{m}_{LL}^+(x)$  at boundary points were obtained by [Fan and Gijbels \(1992\)](#) and are given by

$$E\left(\hat{m}_{LL}^+(x) - m^+(x) | X_1, \dots, X_n\right) = \frac{h^2}{2} m^{(2)}(x) \left( \frac{\mu_{2, \frac{x}{h}}^2 - \mu_{1, \frac{x}{h}} \mu_{3, \frac{x}{h}}}{\mu_{0, \frac{x}{h}} \mu_{2, \frac{x}{h}} - \mu_{1, \frac{x}{h}}^2} \right) + o_p(h^2) \quad (2.27)$$

and

$$V\left(\hat{m}_{LL}^+(x)|X_1, \dots, X_n\right) = \frac{\mu_{2, \frac{x}{h}}^2 \eta_{0, \frac{x}{h}} - 2\mu_{2, \frac{x}{h}} \mu_{1, \frac{x}{h}} \eta_{1, \frac{x}{h}} + \mu_{1, \frac{x}{h}}^2 \eta_{2, \frac{x}{h}}}{\left(\mu_{0, \frac{x}{h}} \mu_{2, \frac{x}{h}} - \mu_{1, \frac{x}{h}}^2\right)^2} \frac{\sigma^2}{nhf(x)} + o_p((nh)^{-1}). \quad (2.28)$$

Compared to the bias of LL estimators at interior points, given by  $E(\hat{m}_{LL}^+(x)|X_1, \dots, X_n) - m^+(x) = \frac{h^2}{2} m^{+(2)}(x) \kappa_2 + o_p(h^2)$ , the leading terms of the LL estimators at the boundary and interior points have the same order  $h^2$  but different magnitude.<sup>2</sup> One way to interpret this is that LL estimators adapt to the boundary by adjusting a regular kernel to an effective kernel, substituting  $\kappa_2$  with

$$\left(\frac{\mu_{2, \frac{x}{h}}^2 - \mu_{1, \frac{x}{h}} \mu_{3, \frac{x}{h}}}{\mu_{0, \frac{x}{h}} \mu_{2, \frac{x}{h}} - \mu_{1, \frac{x}{h}}^2}\right).$$

Compared to the variance of LL estimator at the boundary, the variance of the Hestenes estimators at the boundary in equation (2.24) has leading terms of the same order  $1/nh$  but different magnitudes. The coefficients of the variance of the LL estimator are impacted by the partial kernels at the boundary point whereas those of Hestenes-based estimators depend on the chosen coefficients of Hestenes' extensions.

An asymptotic approximation for the conditional MSE of  $\hat{m}_{LL}^+(x)$  can easily be obtained as is given by

$$\begin{aligned} MSE(\hat{m}_{LL}^+(x)|X_1, \dots, X_n) &= \frac{h^4}{4} \left(m^{+(2)}(x)\right)^2 \left(\frac{\mu_{2, \frac{x}{h}}^2 - \mu_{1, \frac{x}{h}} \mu_{3, \frac{x}{h}}}{\mu_{0, \frac{x}{h}} \mu_{2, \frac{x}{h}} - \mu_{1, \frac{x}{h}}^2}\right)^2 \\ &+ \frac{\mu_{2, \frac{x}{h}}^2 \eta_{0, \frac{x}{h}} - 2\mu_{2, \frac{x}{h}} \mu_{1, \frac{x}{h}} \eta_{1, \frac{x}{h}} + \mu_{1, \frac{x}{h}}^2 \eta_{2, \frac{x}{h}}}{\left(\mu_{0, \frac{x}{h}} \mu_{2, \frac{x}{h}} - \mu_{1, \frac{x}{h}}^2\right)^2} \frac{\sigma^2}{nhf(x)} + o_p\left(h^4 + \frac{1}{nh}\right), \end{aligned}$$

and an optimal bandwidth  $h_{pi}$  for  $\hat{m}_{LL}^+$  can be obtained by minimizing the leading terms in this expression,

$$h_{pi} = n^{-\frac{1}{5}} t_{1n}^{\frac{1}{5}} t_{1n}^{-\frac{2}{5}}, \quad (2.29)$$

<sup>2</sup>Expressions for the unconditional bias and variance of the LL estimator when the regressors take values in  $\mathbb{R}$  were given by Fan (1993). In particular, he finds that  $E(\hat{m}_{LL}(x)) - m(x) = \frac{h^2}{2} m^{(2)}(x) \kappa_2 + o(h^2)$  and  $V(\hat{m}_{LL}(x)) = \frac{\sigma^2}{nh} f^{-1}(x) \lambda_0 + o((nh)^{-1})$ .

where  $t_{1n} = \left(m^{(2)}\right)^2 \left(\frac{\mu_{2,\frac{x}{h}}^2 - \mu_{1,\frac{x}{h}} \mu_{3,\frac{x}{h}}}{\mu_{0,\frac{x}{h}} \mu_{2,\frac{x}{h}} - \mu_{1,\frac{x}{h}}^2}\right)^2$  and  $t_{2n} = \frac{\mu_{2,\frac{x}{h}}^2 \eta_{0,\frac{x}{h}} - 2\mu_{2,\frac{x}{h}} \mu_{1,\frac{x}{h}} \eta_{1,\frac{x}{h}} + \mu_{1,\frac{x}{h}}^2 \eta_{2,\frac{x}{h}}}{\left(\mu_{0,\frac{x}{h}} \mu_{2,\frac{x}{h}} - \mu_{1,\frac{x}{h}}^2\right)^2} \frac{\sigma^2}{f(x)}$ . Although a direct comparison the bias and variance expression for the Hestenes-based and LL estimators is made difficult by the complexity of these expressions, my simulations will provide additional evidence on their relative magnitudes.

## 2.2 Estimators for a Jump Discontinuity

The Hestenes-based estimators  $\hat{m}_{HT}^+(0)$  and  $\hat{m}_{HT}^-(0)$  can be used to estimate the jump at  $x = 0$ , denote by  $J(0)$  by  $\hat{J}_{HT}(0) = \hat{m}_{HT}^+(0) - \hat{m}_{HT}^-(0)$ . [Hahn et al. \(2001\)](#) establishes the identification of the RD model and uses the jump discontinuity of the expected outcome at that point to measure an average treatment effect. Let  $Y$  be the outcome variable and  $X$  be the running variable. When  $X \in \mathbb{R}$  is above a threshold  $x = 0$ , the individual gets the treatment and  $D = 1$ , otherwise the individual does not get the treatment and  $D = 0$ . The regression jump is

$$J(0) = \frac{\lim_{x \downarrow 0} m^+(x) - \lim_{x \uparrow 0} m^-(x)}{\lim_{x \downarrow 0} E(D|X = x) - \lim_{x \uparrow 0} E(D|X = x)}, \quad (2.30)$$

and, in particular, for a sharp RD design

$$J(0) = \lim_{x \downarrow 0} m^+(x) - \lim_{x \uparrow 0} m^-(x),$$

because  $\lim_{x \downarrow c} E(D|X = x) - \lim_{x \uparrow c} E(D|X = x) = 1$ . An estimator for  $J(0)$

$$\hat{J}_{HT}(0) = \lim_{x \downarrow 0} \hat{m}_{HT}^+(x) - \lim_{x \uparrow 0} \hat{m}_{HT}^-(x) = \hat{m}_{HT}^+(0) - \hat{m}_{HT}^-(0), \quad (2.31)$$

where  $\hat{m}_H^+(0) = \lim_{x \downarrow 0} \hat{m}_H^+(x)$  and  $\hat{m}_H^-(0) = \lim_{x \uparrow 0} \hat{m}_H^-(x)$ .

**Theorem 7.** *Suppose that  $K$  satisfies Assumption 1,  $f \in C_b^4(\mathbb{R})$ ,  $m^+ : [0, \infty) \rightarrow \mathbb{R}$  is such that  $m^+ \in C_b^4([0, \infty))$ , and  $m^- : (-\infty, 0] \rightarrow \mathbb{R}$  is such that  $m^- \in C_b^4((-\infty, 0])$ . Then,*

$$\sqrt{nh} \left[ \hat{J}_{HT}(0) - \left( J(0) + U(0) + o_p(h^2) \right) \right] \xrightarrow{d} N\left(0, 2cf^{-2}(0)\right), \quad (2.32)$$

where  $J(0) = m^+(x) - m^-(x)$ ,  $U(0) = B^+(0) - B^-(0)$ .  $B^+(0) = \frac{1}{2}m^{+(2)}(0) + \left(\frac{m^{+(1)}(0)f^{(1)}(0)}{f(0)}\right)\kappa_2$  and  $B^-(0) = \frac{1}{2}m^{-(2)}(0) + \left(\frac{m^{-(1)}(0)f^{(1)}(0)}{f(0)}\right)\kappa_2$  and

$$c = \begin{cases} \sigma^2 f(0) \int_0^\infty \left[ \sum_{i=0}^{s+1} \frac{k_i}{w_i} K\left(\frac{u}{w_i}\right) \right]^2 du, & \text{if } x = 0 \\ \sigma^2 f(x) \lambda_0, & \text{if } x > 0. \end{cases}$$

*Proof.* From the definition  $\hat{J}_{HT}(0) = \hat{m}_{HT}^+(0) - \hat{m}_{HT}^-(0)$ ,

$$\begin{aligned} E(\hat{J}_{HT}(0)|X) &= E(\hat{m}_{HT}^+(0)) - E(\hat{m}_{HT}^-(0)) \\ &= m^+(x) + B^+(0)\kappa_2 h^2 + o_p(h^2) + O_p\left(\frac{h}{(nh)^{1/2}} + h^4\right) \\ &\quad - \left[ m^-(x) + B^-(0)\kappa_2 h^2 + o_p(h^2) + O_p\left(\frac{h}{(nh)^{1/2}} + h^4\right) \right] \\ &= J(0) + U(0) + o_p(h^2) + O_p\left(\frac{h}{(nh)^{1/2}} + h^4\right) \end{aligned}$$

If  $\sqrt{nh^9} \rightarrow 0$ , then  $O_p(h + \sqrt{nh^9}) = o_p(1)$ .

$$\begin{aligned} &\sqrt{nh} \left( \hat{J}_{HT}(0) - E(\hat{J}_{HT}(0)|X) \right) \\ &= \sqrt{nh} \left\{ \hat{J}_{HT}(0) - \left[ J(0) + U(0) + o_p(h^2) + O_p\left(\frac{h}{(nh)^{1/2}} + h^4\right) \right] \right\} \\ &= \sqrt{nh} \left\{ \hat{m}_{HT}^+(0) - \left[ m^+(0) + B^+(0)h^2 + o_p(h^2) \right] \right\} \\ &\quad + \sqrt{nh} \left\{ \hat{m}_{HT}^-(0) - \left[ m^-(0) + B^-(0)h^2 + o_p(h^2) \right] \right\} \end{aligned}$$

Because the regression estimators to the right  $\hat{m}_{HT}^+(0)$  and to the left  $\hat{m}_{HT}^-(0)$  of the threshold are constructed using random samples from each side of the threshold separately, the two regression estimators are independent and the covariances between them vanish. By Theorem 5, I obtain the asymptotic distribution of  $\hat{J}_{HT}(0)$  as

$$\sqrt{nh} \left[ \hat{J}_{HT}(0) - \left( J(0) + U(0) + o_p(h^2) \right) \right] \xrightarrow{d} N\left(0, 2cf^{-2}(0)\right).$$

□

### 2.2.1 A Comparison with Local Linear Jump Estimators

It is instructive to compare the asymptotic properties of my jump estimator  $\hat{J}_{HT}$  with those of the local linear jump estimator from [Porter \(2003\)](#). Theorem 3(a) in his paper provides the asymptotic distribution of the local polynomial estimator under the similar assumptions as my setup, in which the derivatives of regressions on both sides are not required to be the same. When the order of the polynomial  $p = 1$ , the estimator is local linear jump estimator. His LL jump estimator  $\bar{\alpha}_p$  for the true jump  $\alpha$  has the following asymptotic distribution:

$$\sqrt{nh}(\bar{\alpha}_p - \alpha) \xrightarrow{d} N\left(B_a, \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{f_0(\bar{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1\right),$$

where  $h^2 \sqrt{nh} \rightarrow C_a < \infty$ ,

$$B_a = \frac{C_a}{2} \left[ m^{(2)+}(\bar{x}) - m^{(2)-}(\bar{x}) \right] e_1' \Gamma^{-1} \begin{pmatrix} \gamma_2 \\ \gamma_3 \end{pmatrix}, \quad (2.33)$$

$\Gamma = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_2 \end{bmatrix}$ ,  $\Delta = \begin{bmatrix} \delta_0 & \delta_1 \\ \delta_1 & \delta_2 \end{bmatrix}$ ,  $e_1 = (1, 0)'$ ,  $\gamma_j = \int_0^\infty k(u) u^j du$ , and  $\delta_j = \int_0^\infty k^2(u) u^j du$ , for  $j = 0, 1, 2$ .

Some simple algebra manipulations show that the bias of the LL jump estimator is the difference of the biases of the LL regression estimators from both sides. The coefficients only involve with kernels and are equivalent to the coefficient of the leading term  $h^2$  in equation (2.27).

$$\begin{aligned} e_1' \Gamma^{-1} \begin{pmatrix} \gamma_2 \\ \gamma_3 \end{pmatrix} &= (1 \ 0) \frac{1}{r_0 r_2 - r_1^2} \begin{bmatrix} \gamma_2 & -\gamma_1 \\ -\gamma_1 & \gamma_0 \end{bmatrix} \begin{pmatrix} \gamma_2 \\ \gamma_3 \end{pmatrix} \\ &= \frac{\gamma_2^2 - \gamma_1 \gamma_3}{r_0 r_2 - r_1^2} \end{aligned}$$

The variance of the LL jump estimators is the sum of the variances of the LL regression estimators from both sides. The coefficients only involve with kernels and are equivalent to the

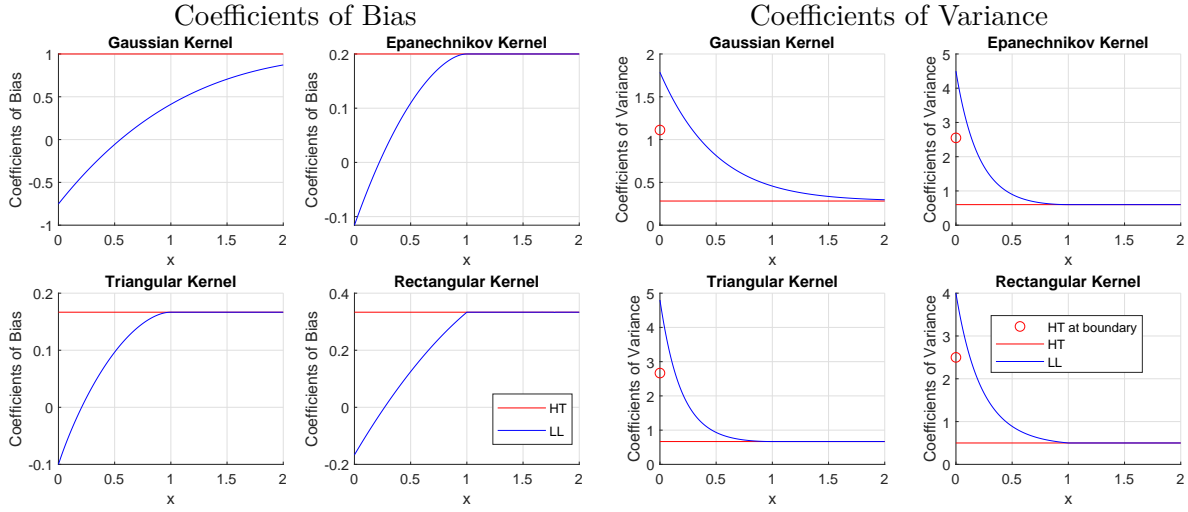
coefficient of the leading term  $\frac{1}{nh}$  in equation (2.28)

$$\begin{aligned}
& e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \\
&= (1 \ 0) \frac{1}{(r_0 r_2 - r_1^2)^2} \begin{bmatrix} \gamma_2 & -\gamma_1 \\ -\gamma_1 & \gamma_0 \end{bmatrix} \begin{bmatrix} \delta_0 & \delta_1 \\ \delta_1 & \delta_2 \end{bmatrix} \begin{bmatrix} \gamma_2 & -\gamma_1 \\ -\gamma_1 & \gamma_0 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
&= \frac{\gamma_2^2 \delta_0 - 2\gamma_1 \gamma_2 \delta_1 + \gamma_1^2 \delta_2}{(r_0 r_2 - r_1^2)^2}
\end{aligned}$$

Figure 2.2 shows the comparison on coefficients of the leading terms of asymptotic distribution of the Hestenes regression estimator from theorem 5 with those of the LL regression estimator from equation (2.27) and (2.28). The figure on the left shows the coefficients of the leading terms of bias of the asymptotic distribution for the HT estimator and the LL estimator with four different kernels: Gaussian, Epanechnikov, triangular, and rectangular. The adaptive feature of the LL estimators determines that at the boundary, the LL estimator has a smaller bias than that at interior points and gradually merges to be the same as the interior points. In contrast, at the boundary, the HT estimator has the same bias as the interior point. Although the LL jump estimator has smaller coefficients, its bias is not necessarily smaller than the HT jump estimator. The bias of HT jump estimator, shown in equation (2.32), is the sum of two differences between the leading terms from the regression estimators on both sides: one pair involves  $m^{(2)}$  and the other pair involves  $\frac{m^{(1)} f^{(1)}}{f}$ . The bias of the LL jump estimator, shown in equation (2.33), is the difference between the leading terms that only involve  $m^{(2)}$ .

The figure on the right shows the coefficients of the leading terms of variance of the asymptotic distribution for the HT and LL estimators for the four different kernels. The adaptive feature of the LL estimators determines that at the boundary, the LL estimator has a larger variance than interior points because of less observations and gradually merges to be the same as the interior points. At the boundary, the HT estimator has larger variances than the interior points but smaller variances than the LL estimator for all kernels. The table below the graph provides the coefficients at the boundary for the HT and LL estimators. It shows that the HT estimator have smaller variances

**Figure 2.2:** Coefficients of Leading Terms of Asymptotic Distribution



Coefficients of Variance at the Boundary

|    | Gaussian | Epanechnikov | Triangular | Rectangular |
|----|----------|--------------|------------|-------------|
| HT | 1.11     | 2.55         | 2.67       | 2.50        |
| LL | 1.79     | 4.50         | 4.80       | 4.00        |

than the LL estimators at the boundary for all four kernels. This implies that, more often than not, the HT jump estimators have smaller mean squared error than the LL jump estimators.

### 2.2.2 A Discussion about Bandwidth Selection

One critical aspect of bandwidth selection is that in RDD, the bandwidth should be locally rather than globally optimal in terms of some kind of criteria, such as MSE. This is different from traditional literature where the bandwidth minimizes some kind of integrated versions of MSE because I am less concerned about the performance of this estimator outside the point of discontinuity. I want to pick up an optimal  $h$  that leads the optimal estimator for the jump at the point where the function jumps. That is to say, I want to minimize some sort of local loss function rather than an integrated MSE. The locality here is for  $x$  be exactly at the point of discontinuity. That is why this bandwidth selection problem is different from the bandwidth selection problem we normally encounter in nonparametric regression.

This problem is recognized when LL estimators are used in RDD. [Imbens and Kalyanaraman](#)



(2012) provide an optimal data-adaptive bandwidth for LL estimators to minimize the mean-squared error. In the same vein, it is desirable to develop the optimal bandwidth for the Hestenes estimators. Since in RDD, I am only interested in the precise estimation of the jump at the discontinuity point, a locally optimal bandwidth suits my purpose better than a globally optimal bandwidth.

The choice of other parameters  $s$ ,  $w_i$ , and the kernel  $K$  might be intricately related to the bandwidth  $h$ . Because I want to choose  $s$  and  $w_i$  optimally, but the  $h$  I choose may be related to the choice of  $s$  and  $w_i$ . The optimal  $h$  for one element in the class might not be optimal for another. The kernel  $K$  is also determining some results, specifically, the structure of the variance. It is quite possible to have an optimal  $K$  for LL estimators that is different from the optimal  $K$  for my estimators.

In a regular nonparametric regression estimation, we often obtain optimal rates for  $h$  by minimizing the asymptotic mean integrated squared error (AMISE). Given that I have the asymptotic theorem, if I come up with some criterion to pick up  $h$ , that criterion should not be AMISE as in typical regression settings because I do not want to select bandwidth that minimizes global criterion and I do not care about estimates away from the point of discontinuity. If I am going to develop a theory about a bandwidth selection for fixed  $s$  and  $w_i$ , I should choose some loss function at the point of discontinuity that will lead to the optimal  $h$ .

This issue of the optimal rate of  $h$  is much more involved than just picking up a  $h$  that gives the asymptotic result. The choice of  $h$  in terms of asymptotic result is straightforward, because there is a great number of  $h$  that I can choose to give asymptotic results, namely  $h \rightarrow 0$  and  $nh \rightarrow \infty$ . Choice of  $h$  is more complicated for finite sample estimation. Furthermore, Choosing  $h$  by the conditional bias and variance could be different from choosing  $h$  by the unconditional bias and variance. It depends on the structure of both distributions and the expansion of the bias and variance. They may or may not give a different order of  $h$ . Once I derive the expressions for unconditional bias and variance of my estimator at the boundary, I can choose the optimal bandwidth for my estimators by minimizing the loss function associated with the unconditional bias and variance such as the mean squared error.

A related issue is optimal inference. In regular regression settings, suppose the regression has a continuous second order derivatives. When we calculate the optimal bandwidth for an estimation based on the minimization of AMISE, we get an bandwidth of the order of  $(1/n)^{1/5}c$ . If we use that for inference, the bias does not vanish. It is commonly said that the optimal bandwidth is not suitable for inference. This happens with LL estimators in RDD setting. [Calonico et al., 2014](#) try to solve the problem with the bias estimator to correct inference. This is because people who are dealing with estimating regression discontinuity are just using unmodified LL estimators, so the problem that occurs in normal settings reappears in RDD settings. Will my estimators have the same problem? if there is a loss function for optimal  $h$ , and if the optimal  $h$  based on this loss function indeed leads to an asymptotic distribution that has a bias and that bias does not go away when I use the optimal  $h$ , then I will have the same problem. But for now I cannot reach the conclusion. I will only know what is happening once a loss function is defined and an optimal  $h$  is chosen.

### 2.3 Monte Carlo Studies

I conduct simulations to compare the finite performance of the Hestenes (HT) estimators with the classical and popular regression estimators – Nadaraya-Watson (NW) and local linear (LL) estimators. I design eight groups of Monte Carlo experiments. Each group consists of four experiments that generate random samples according to the four regression functions that show different types of jumps. I examine the performance of these estimators under the following scenarios: 1) different bandwidth selections: the plug-in, cross-validation, and Jackknife cross-validation method; 2) choice of kernel functions: the Gaussian, Epanechnikov, Triangular, and Rectangular kernels; 3) different sample sizes:  $n = 500$  and  $n = 1000$ ; and 4) density of the regressor  $X$ : a standard Gaussian and a normal distribution with  $\mu = 0.5$  and  $\sigma^2 = 1$ . I run each Monte Carlo experiment with 2,000 repetitions.

The primary focus of my comparisons is the estimates at the point of discontinuity, so I measure performances by bias, sample variance (VAR), and root mean squared error (RMSE) of estimators at the discontinuity point. Although estimates at interior points are not the main focus, I check overall

estimation performance. I measure root average squared error (RASE) for all points including the boundary point and interior points.

Let  $x = 0$  be the discontinuity point,  $\hat{m}(x)$  be an estimate for the regression function  $m(x)$  at  $x$ , and  $\hat{J}(0)$  be an estimate for the jump  $J(0)$ . I generate  $M$  samples and calculate the following statistics for  $\hat{m}(x)$  and  $\hat{J}(0)$ :

$$\text{Bias}(\hat{\theta}) = \frac{\sum_{m=1}^M (\hat{\theta}_m - \theta)}{M}, \text{VAR}(\hat{\theta}) = \frac{\sum_{m=1}^M (\hat{\theta}_m - \bar{\hat{\theta}})^2}{M-1}, \text{RMSE}(\hat{\theta}) = \sqrt{\frac{\sum_{m=1}^M (\hat{\theta}_m - \theta)^2}{M}},$$

where  $\bar{\hat{\theta}} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m$ ,  $\theta$  denotes  $J(0)$ ,  $m^+(0)$ , or  $m^-(0)$ , and  $\hat{\theta}$  denotes  $\hat{J}(0)$ ,  $\hat{m}^+(0)$  or  $\hat{m}^-(0)$ . I also calculate

$$\text{Rel\_R}(\hat{\theta}) = \frac{\text{RMSE}(\hat{\theta})}{\min_{\hat{\theta}} \text{RMSE}(\hat{\theta})}.$$

I calculate these statistics for  $\hat{m}(x)$  at all points of evaluation. For each sample and estimator, I calculate a root average squared error (RASE) over  $K$  evaluation points,

$$\text{RASE}_m = \sqrt{\frac{\sum_{k=1}^K (\hat{m}(x_k) - m(x_k))^2}{K}} \text{ for the sample } m.$$

Then I calculate an average of RASE across all generated samples,

$$\text{AMSE} = \frac{\sum_{m=1}^M \text{RASE}_m}{M}.$$

I also calculate

$$\text{Rel\_A}(\hat{\theta}) = \frac{\text{AMSE}(\hat{\theta})}{\min_{\hat{\theta}} \text{AMSE}(\hat{\theta})}.$$

### 2.3.1 Estimators Under the Study

I compare three types of estimators: NW estimators, LL estimators, and Hestenes estimators. For all three estimators, I split samples using  $X_i \geq 0$  and  $X_i < 0$ . I use observations to the right of the discontinuity point to estimate the regression function on the right and use observations to

the left of the discontinuity point to estimate the regression function on the left, and then calculate the jump at the discontinuity point.

I construct NW estimators as

$$\hat{m}_{NW}^+(x) = \frac{(nh)^{-1} \sum_{i=1}^n K\left(\frac{X_i-x}{h}\right) Y_i d_i}{(nh)^{-1} \sum_{i=1}^n K\left(\frac{X_i-x}{h}\right) d_i} \text{ for } x \geq 0.$$

Similarly, I define  $\hat{m}_{NW}^-(x)$  for  $x < 0$  by changing  $d_i$  to  $1 - d_i$ .

For LL estimators, I use the modified version of the LL estimator by [Fan \(1992\)](#) to avoid singularities:<sup>3</sup>

$$\hat{m}_{LL}^+(x) = \sum_{j=1}^n w_j Y_j d_j / \left[ \sum_{j=1}^n w_j d_j + \left( \sum_{j=1}^n d_j \right)^{-2} \right],$$

where  $w_j \equiv K\left(\frac{x-X_j}{h}\right)[s_{n,2} - (x - X_j)s_{n,1}]$  and  $s_{n,l} = \sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)(x - X_j)^l, l = 1, 2$ . Similarly, I define  $\hat{m}_{LL}^-(x)$  for  $x < 0$  by changing  $d_j$  to  $1 - d_j$ .

I construct HT estimators  $\hat{m}_{HT}$  as in equation (2.21). Hestenes estimators are denoted as Hsb, such as H00, H10, H11, or H21, where the first digit stands for the degree of smoothness of the composite function  $\mu(x)$ ,  $s = 0, 1, 2$ , and the second digit denotes which sequence of  $w_i$  is used:  $b = 0$  means the sequence  $w_i = 1/i$  is used while  $b = 1$  means the sequence  $w_i = i$  is used, with  $i = 1, 2, \dots, n$ .

Subsequently, I obtain the jump estimator  $\hat{J}_d(0) = \hat{m}_d^+(x) - \hat{m}_d^-(x)$ , where  $d$  could be the NW, LL, or HT estimator.  $\hat{m}_d^-(x)$  is the regression estimator on the left and  $\hat{m}_d^+(x)$  is the regression estimator on the right of the discontinuity point.

### 2.3.2 The Data Generating Process (DGP)

For each experiment, I generate data using the following model:

$$Y = m(X) + \epsilon \quad \text{where } \epsilon \sim N(0, 1).$$

<sup>3</sup>As proved by [Fan, 1992](#), this estimators has the same asymptotic properties as the regular LL estimator described in equations (2.25) and (2.26), so I can use properties derived from the latter estimators for comparison.

The regressor  $X$  is generated with a standard normal or a normal with  $\mu = 0.5$  and  $\sigma^2 = 1$ . I design the four regression functions  $m(X)$  that have different types of jumps. Table 2.1 shows these regression models, the changes in  $m$ , and the sign changes in the first derivative  $m^{(1)}$  and second derivatives  $m^{(2)}$  at the point of discontinuity.

**Table 2.1:** Regression Models

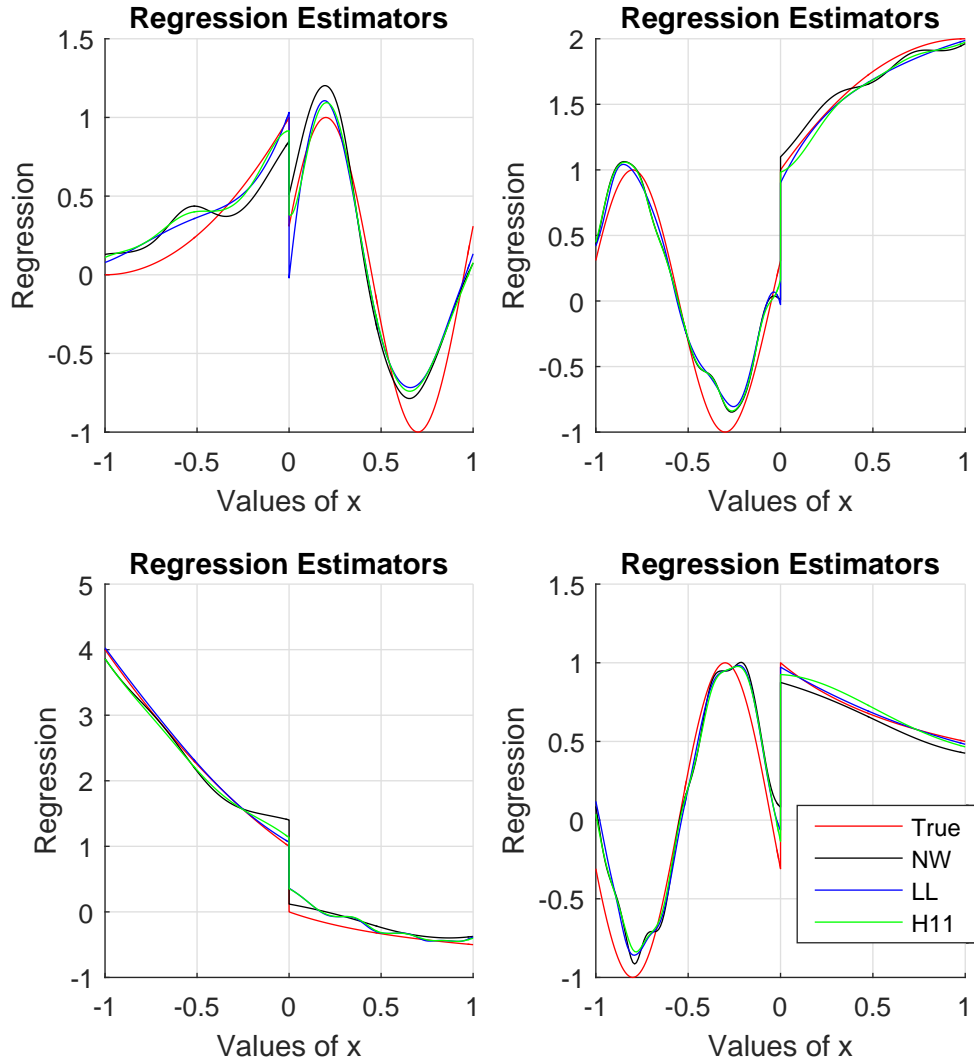
$\Delta m$ : Changes in  $m(x)$  at  $x = 0$ , .

$m^{-(1)}$ ,  $m^{+(1)}$ ,  $m^{-(2)}$ , and  $m^{+(2)}$ : the signs of the first and second order derivatives of  $m(x)$  to the left and right of  $x = 0$ .

|       | Regression  | $\Delta m$ | $m^{-(1)}$ | $m^{+(1)}$ | $m^{-(2)}$ | $m^{+(2)}$ |
|-------|---|------------|------------|------------|------------|------------|
| Reg 1 | $m(x) = \begin{cases} (x+1)^2, & x < 0 \\ \sin(2\pi x + 0.1\pi), & x \geq 0 \end{cases}$            | drop       | +          | +          | +          | -          |
| Reg 2 | $m(x) = \begin{cases} \sin(2\pi x + 0.1\pi), & x < 0 \\ -(x-1)^2 + 2, & x \geq 0 \end{cases}$       | jump       | +          | +          | -          | -          |
| Reg 3 | $m(x) = \begin{cases} (x-1)^2, & x < 0 \\ \frac{1}{x+1} - 1, & x \geq 0 \end{cases}$                | drop       | -          | -          | +          | +          |
| Reg 4 | $m(x) = \begin{cases} -\sin(2\pi x + 0.1\pi), & x < 0 \\ \frac{1}{x+1} - 1, & x \geq 0 \end{cases}$ | jump       | -          | -          | +          | +          |

Figure 2.3 shows the true regression function and the estimates from the NW, LL, and HT estimators. As expected, the estimators differ in their behavior the most in the vicinity of the discontinuity point. NW estimators have a significant bias while the LL and HT estimators stay close to each other and are close to the true regression functions at the boundary.

**Figure 2.3:** Regression Functions Estimated by Three Estimators



True regression (red),  $\hat{m}_{NW}$  (black),  $\hat{m}_{LL}$  (blue),  $\hat{m}_{Hsb}$  (green), where  $s$  the degree of smoothness,  $b = 0, w_i = \frac{1}{i}$  and  $b = 1, w_i = i$ .

Graphics can give us an intuitive impression of each estimator's behavior, but to evaluate performance across estimators, we rely on large sample simulations to show the distribution of the estimators.

### 2.3.3 Simulation Results

The simulation results of the eight group experiments are shown on the eight tables 2.2 to 2.9. For each table, Panel A shows the results of the jump estimator and Panel B shows the results of the

regression estimators to the left and right of the discontinuity point. The bias of the jump estimator is the difference of the biases of the regression estimators from both sides while the variance of the jump estimator is the sum of the variances of the regression estimators from both sides. AMSE measures the error over the whole domain, which is not my main concern since I am interested in the estimates of the jump. As expected, all three types of estimators have similar values for their AMSE. Across the tables, I observe that the NW estimator has the largest bias at the discontinuity point in all cases.

The first group of experiments has the default setup: the bandwidths are chosen through the cross-validation method; the kernel is the Gaussian kernel; and the sample size is 500; The regressor  $X$  is generated with a standard normal. Table 2.2 shows the results of the jump and regression estimation. The HT estimators have smaller (3 out of the 4 jumps and 7 out of the 8 regressions) or similar biases as the LL estimators. The HT estimators have the smallest RMSE in 3 out of the 4 jumps and 5 out of the 8 regressions.

Table 2.3 shows the results of the jump and regression estimation as the sample size increases to 1,000. Again, the HT estimators have smaller (3 out of the 4 jumps and 7 out of the 8 regressions) or similar biases as the LL estimators. The HT estimators have the smallest RMSE in 4 out of the 4 jumps and 6 out of the 8 regressions. Compared to Table 2.2, as the sample size increases, the bias and variance of the LL and HT estimators become smaller while the variances of the NW estimator become smaller but their biases have almost no change. This accords with the theoretical result that the biases of LL and HT estimators decrease in the order of  $h^2$  while the biases of NW estimators decrease in the order of  $h$ .

The simulations conducted above use a standard normal distribution to generate  $X$  with the peak of the density occurring at the discontinuity point. This splits observations almost evenly between the two sides of the discontinuity point, which in reality is not always the case. For example, in the case of using students' grade as the running variable to decide whether a third-year student can be promoted to the next grade or will be held back for one more year, the cutoff grade (let us say 60 out of 100) is often located away from the peak of the grade distribution (say 80 out of 100). Therefore, it is useful to check how estimators detect the jump when the cutoff point is

**Table 2.2:** Simulation Results for the Default Setting

| Panel A: Jump Estimators       |              |              |                 |        |              |        |              |              |                 |        |              |        |
|--------------------------------|--------------|--------------|-----------------|--------|--------------|--------|--------------|--------------|-----------------|--------|--------------|--------|
|                                | NW           | LL           | HT00            | HT10   | HT11         | HT21   | NW           | LL           | HT00            | HT10   | HT11         | HT21   |
|                                |              |              | DGP 1           |        |              |        |              |              | DGP 2           |        |              |        |
| Bias                           | 0.529        | 0.207        | 0.725           | 0.063  | 0.644        | 0.145  | 0.473        | -0.003       | 0.576           | -0.035 | 0.089        | -0.252 |
| Variance                       | 0.060        | 0.152        | 0.037           | 0.237  | 0.079        | 0.282  | 0.086        | 0.318        | 0.091           | 0.485  | 0.207        | 0.389  |
| RMSE                           | 0.583        | 0.442        | 0.750           | 0.491  | 0.702        | 0.551  | 0.556        | 0.564        | 0.650           | 0.697  | 0.464        | 0.672  |
| REL_R                          | 1.319        | <b>1.000</b> | 1.698           | 1.111  | 1.589        | 1.246  | 1.200        | 1.217        | 1.401           | 1.504  | <b>1.000</b> | 1.450  |
| AMSE                           | 0.162        | 0.178        | 0.200           | 0.175  | 0.205        | 0.182  | 0.178        | 0.181        | 0.204           | 0.182  | 0.182        | 0.184  |
| REL_A                          | <b>1.000</b> | 1.097        | 1.231           | 1.076  | 1.259        | 1.121  | <b>1.000</b> | 1.016        | 1.149           | 1.024  | 1.026        | 1.033  |
|                                |              |              | DGP 3           |        |              |        |              |              | DGP 4           |        |              |        |
| Bias                           | -0.293       | -0.007       | -0.251          | -0.002 | -0.046       | 0.124  | -0.525       | -0.048       | -0.361          | 0.012  | -0.125       | 0.188  |
| Variance                       | 0.046        | 0.155        | 0.053           | 0.261  | 0.060        | 0.163  | 0.086        | 0.218        | 0.117           | 1.214  | 0.185        | 0.907  |
| RMSE                           | 0.363        | 0.393        | 0.341           | 0.511  | 0.250        | 0.422  | 0.602        | 0.470        | 0.497           | 1.102  | 0.448        | 0.970  |
| REL_R                          | 1.455        | 1.576        | 1.364           | 2.048  | <b>1.000</b> | 1.693  | 1.343        | 1.048        | 1.109           | 2.459  | <b>1.000</b> | 2.166  |
| AMSE                           | 0.125        | 0.133        | 0.129           | 0.149  | 0.111        | 0.134  | 0.192        | 0.164        | 0.191           | 0.188  | 0.175        | 0.193  |
| REL_A                          | 1.126        | 1.198        | 1.163           | 1.338  | <b>1.000</b> | 1.202  | 1.169        | <b>1.000</b> | 1.164           | 1.146  | 1.064        | 1.178  |
| Panel B: Regression Estimators |              |              |                 |        |              |        |              |              |                 |        |              |        |
|                                | NW           | LL           | HT00            | HT10   | HT11         | HT21   | NW           | LL           | HT00            | HT10   | HT11         | HT21   |
|                                |              |              | DGP 1- Negative |        |              |        |              |              | DGP 1- Positive |        |              |        |
| Bias                           | -0.164       | -0.009       | -0.300          | 0.004  | -0.084       | 0.046  | 0.365        | 0.198        | 0.425           | 0.067  | 0.559        | 0.192  |
| Variance                       | 0.027        | 0.077        | 0.014           | 0.118  | 0.032        | 0.078  | 0.031        | 0.074        | 0.022           | 0.120  | 0.044        | 0.194  |
| RMSE                           | 0.232        | 0.278        | 0.322           | 0.343  | 0.198        | 0.283  | 0.405        | 0.336        | 0.450           | 0.353  | 0.598        | 0.480  |
| REL_R                          | 1.173        | 1.403        | 1.629           | 1.733  | <b>1.000</b> | 1.431  | 1.206        | <b>1.000</b> | 1.339           | 1.050  | 1.777        | 1.427  |
|                                |              |              | DGP 2- Negative |        |              |        |              |              | DGP 2- Positive |        |              |        |
| Bias                           | -0.378       | 0.004        | -0.502          | 0.023  | -0.089       | 0.222  | 0.095        | 0.001        | 0.073           | -0.012 | 0.000        | -0.030 |
| Variance                       | 0.039        | 0.188        | 0.029           | 0.214  | 0.083        | 0.194  | 0.047        | 0.127        | 0.062           | 0.261  | 0.124        | 0.190  |
| RMSE                           | 0.426        | 0.434        | 0.531           | 0.463  | 0.301        | 0.493  | 0.237        | 0.356        | 0.260           | 0.510  | 0.352        | 0.437  |
| REL_R                          | 1.417        | 1.443        | 1.765           | 1.539  | <b>1.000</b> | 1.641  | <b>1.000</b> | 1.500        | 1.098           | 2.153  | 1.484        | 1.843  |
|                                |              |              | DGP 3- Negative |        |              |        |              |              | DGP 3- Positive |        |              |        |
| Bias                           | 0.199        | 0.000        | 0.153           | 0.002  | -0.002       | -0.110 | -0.093       | -0.007       | -0.098          | 0.000  | -0.049       | 0.014  |
| Variance                       | 0.026        | 0.089        | 0.034           | 0.145  | 0.034        | 0.069  | 0.020        | 0.068        | 0.019           | 0.129  | 0.027        | 0.096  |
| RMSE                           | 0.257        | 0.299        | 0.239           | 0.380  | 0.183        | 0.284  | 0.170        | 0.261        | 0.170           | 0.359  | 0.171        | 0.310  |
| REL_R                          | 1.398        | 1.629        | 1.305           | 2.073  | <b>1.000</b> | 1.550  | 1.003        | 1.535        | <b>1.000</b>    | 2.110  | 1.004        | 1.824  |
|                                |              |              | DGP 4- Negative |        |              |        |              |              | DGP 4- Positive |        |              |        |
| Bias                           | 0.470        | 0.029        | 0.312           | -0.013 | 0.107        | -0.189 | -0.055       | -0.019       | -0.048          | -0.001 | -0.018       | -0.001 |
| Variance                       | 0.033        | 0.116        | 0.051           | 0.674  | 0.086        | 0.186  | 0.053        | 0.100        | 0.065           | 0.527  | 0.097        | 0.707  |
| RMSE                           | 0.504        | 0.341        | 0.385           | 0.821  | 0.312        | 0.470  | 0.237        | 0.317        | 0.260           | 0.725  | 0.312        | 0.841  |
| REL_R                          | 1.619        | 1.096        | 1.237           | 2.634  | <b>1.000</b> | 1.509  | <b>1.000</b> | 1.339        | 1.098           | 3.067  | 1.319        | 3.553  |



**Table 2.3:** Varying Sample Size

| Panel A: Jump Estimators |        |              |              |        |              |        |        |        |              |        |              |        |
|--------------------------|--------|--------------|--------------|--------|--------------|--------|--------|--------|--------------|--------|--------------|--------|
|                          | NW     | LL           | HT00         | HT10   | HT11         | HT21   | NW     | LL     | HT00         | HT10   | HT11         | HT21   |
|                          |        |              | DGP 1        |        |              |        |        |        | DGP 2        |        |              |        |
| Bias                     | 0.709  | 0.037        | 0.455        | -0.012 | 0.073        | -0.094 | 0.605  | 0.050  | 0.525        | -0.010 | 0.071        | -0.168 |
| Variance                 | 0.024  | 0.127        | 0.038        | 0.257  | 0.108        | 0.249  | 0.033  | 0.118  | 0.036        | 0.244  | 0.100        | 0.240  |
| RMSE                     | 0.725  | 0.358        | 0.495        | 0.506  | 0.336        | 0.507  | 0.631  | 0.347  | 0.558        | 0.494  | 0.324        | 0.518  |
| REL_R                    | 2.156  | 1.065        | 1.471        | 1.506  | <b>1.000</b> | 1.509  | 1.952  | 1.073  | 1.724        | 1.526  | <b>1.000</b> | 1.600  |
| AMSE                     | 0.149  | 0.119        | 0.129        | 0.137  | 0.128        | 0.140  | 0.149  | 0.141  | 0.129        | 0.142  | 0.133        | 0.145  |
| REL_A                    | 1.260  | <b>1.000</b> | 1.091        | 1.157  | 1.082        | 1.178  | 1.159  | 1.092  | <b>1.000</b> | 1.102  | 1.033        | 1.125  |
|                          |        |              | DGP 3        |        |              |        |        |        | DGP 4        |        |              |        |
| Bias                     | -0.621 | 0.002        | -0.198       | 0.007  | 0.004        | 0.006  | -0.535 | -0.038 | -0.454       | 0.020  | -0.063       | 0.169  |
| Variance                 | 0.022  | 0.118        | 0.034        | 0.244  | 0.098        | 0.257  | 0.036  | 0.122  | 0.038        | 0.257  | 0.108        | 0.261  |
| RMSE                     | 0.639  | 0.343        | 0.270        | 0.494  | 0.314        | 0.507  | 0.568  | 0.352  | 0.494        | 0.508  | 0.335        | 0.538  |
| REL_R                    | 2.367  | 1.272        | <b>1.000</b> | 1.829  | 1.162        | 1.879  | 1.693  | 1.049  | 1.474        | 1.514  | <b>1.000</b> | 1.606  |
| AMSE                     | 0.199  | 0.113        | 0.105        | 0.135  | 0.126        | 0.138  | 0.148  | 0.140  | 0.128        | 0.142  | 0.133        | 0.145  |
| REL_A                    | 1.891  | 1.071        | <b>1.000</b> | 1.286  | 1.193        | 1.313  | 1.159  | 1.097  | <b>1.000</b> | 1.111  | 1.041        | 1.133  |

| Panel B: Regression Estimators |        |              |                 |        |              |        |              |        |                 |        |              |        |
|--------------------------------|--------|--------------|-----------------|--------|--------------|--------|--------------|--------|-----------------|--------|--------------|--------|
|                                | NW     | LL           | HT00            | HT10   | HT11         | HT21   | NW           | LL     | HT00            | HT10   | HT11         | HT21   |
|                                |        |              | DGP 1- Negative |        |              |        |              |        | DGP 1- Positive |        |              |        |
| Bias                           | -0.384 | -0.004       | -0.120          | 0.006  | -0.008       | 0.016  | 0.324        | 0.032  | 0.335           | -0.005 | 0.064        | -0.077 |
| Variance                       | 0.005  | 0.048        | 0.019           | 0.088  | 0.043        | 0.082  | 0.018        | 0.072  | 0.017           | 0.160  | 0.059        | 0.155  |
| RMSE                           | 0.391  | 0.218        | 0.183           | 0.297  | 0.208        | 0.288  | 0.351        | 0.271  | 0.359           | 0.399  | 0.250        | 0.401  |
| REL_R                          | 2.139  | 1.193        | <b>1.000</b>    | 1.627  | 1.139        | 1.574  | 1.402        | 1.081  | 1.435           | 1.595  | <b>1.000</b> | 1.603  |
|                                |        |              | DGP 2- Negative |        |              |        |              |        | DGP 2- Positive |        |              |        |
| Bias                           | -0.474 | -0.039       | -0.388          | 0.010  | -0.060       | 0.161  | 0.131        | 0.010  | 0.136           | -0.001 | 0.011        | -0.007 |
| Variance                       | 0.016  | 0.047        | 0.020           | 0.094  | 0.045        | 0.111  | 0.017        | 0.070  | 0.016           | 0.145  | 0.054        | 0.134  |
| RMSE                           | 0.490  | 0.221        | 0.413           | 0.306  | 0.221        | 0.370  | 0.184        | 0.264  | 0.186           | 0.380  | 0.233        | 0.366  |
| REL_R                          | 2.217  | 1.000        | 1.866           | 1.384  | <b>1.000</b> | 1.673  | <b>1.000</b> | 1.437  | 1.011           | 2.071  | 1.271        | 1.992  |
|                                |        |              | DGP 3- Negative |        |              |        |              |        | DGP 3- Positive |        |              |        |
| Bias                           | 0.560  | -0.006       | 0.135           | -0.003 | -0.009       | -0.008 | -0.061       | -0.003 | -0.063          | 0.004  | -0.005       | -0.002 |
| Variance                       | 0.006  | 0.046        | 0.018           | 0.091  | 0.042        | 0.082  | 0.017        | 0.072  | 0.016           | 0.151  | 0.057        | 0.173  |
| RMSE                           | 0.565  | 0.214        | 0.191           | 0.301  | 0.204        | 0.286  | 0.143        | 0.269  | 0.142           | 0.389  | 0.239        | 0.415  |
| REL_R                          | 2.956  | 1.120        | <b>1.000</b>    | 1.576  | 1.069        | 1.494  | 1.011        | 1.900  | <b>1.000</b>    | 2.746  | 1.688        | 2.933  |
|                                |        |              | DGP 4- Negative |        |              |        |              |        | DGP 4- Positive |        |              |        |
| Bias                           | 0.479  | 0.048        | 0.395           | 0.000  | 0.071        | -0.149 | -0.056       | 0.010  | -0.059          | 0.020  | 0.008        | 0.020  |
| Variance                       | 0.016  | 0.047        | 0.019           | 0.096  | 0.046        | 0.110  | 0.018        | 0.074  | 0.017           | 0.165  | 0.060        | 0.153  |
| RMSE                           | 0.496  | 0.221        | 0.419           | 0.310  | 0.225        | 0.363  | 0.145        | 0.273  | 0.143           | 0.406  | 0.244        | 0.391  |
| REL_R                          | 2.242  | <b>1.000</b> | 1.895           | 1.400  | 1.019        | 1.644  | 1.012        | 1.904  | <b>1.000</b>    | 2.838  | 1.705        | 2.733  |

away from the peak of the distribution of the running variable. Table 2.4 shows the performance of these estimators when the peak of the density function  $f$  is shifted away from the discontinuity point. The regressor  $X$  is generated with a normal distribution with  $\mu = 0.5$  and  $\sigma^2 = 1$ . The HT estimators have smaller (2 out of the 4 jumps and 5 out of the 8 regressions) or similar biases as the LL estimators. The HT estimators have the smallest RMSE in 4 out of the 4 jumps and 4 out of the 8 regressions. Comparing to table 2.2, the default setting, there are no fundamental changes in performance of estimators when the peak of the density function  $f$  is shifted away from the cutoff point.

Bandwidth selection is critical in nonparametric estimation. When I compare performance of different estimators, I want to use a method that is impartial to all estimators when choosing bandwidth. In the default setting, I choose bandwidth by the cross-validation method. That is, I choose a globally optimal bandwidth by minimizing the sum of the squared deviations between the observed outcomes and the true regression function. The calculation depends solely on the construction of the estimators and the true regression functions.

I add another two methods, the plug-in and jackknife cross-validation method. For the plug-in method, I choose a locally optimal bandwidth by optimizing the leading terms of the mean squared error of the regression estimator at the boundary point. This method uses the asymptotic properties of the estimators at the boundary. Again, I take advantage of the fact that I know the true density and regression function. For the NW estimator, I use results from Corollary 1 and 2 to get

$$\begin{aligned}
 AMSE(h) = & \left\{ -\frac{1}{2}m(0) - h \left( \frac{m(0)f^{(1)}(0)}{f(0)} + m^{(1)}(0) \right) \mu_{1,0} \right. \\
 & \left. + \frac{h^2}{2} \left( m^{(2)}(0) + \frac{2m^{(1)}(0)f^{(1)}(0)}{f(0)} + \frac{f^{(2)}(0)m(0)}{f(0)} \right) \mu_{2,0} \right\}^2 \\
 & + \frac{1}{nh} \left\{ \frac{m^2(0) + \sigma^2}{f(0)} \eta_{0,0} \right\} + s.o.
 \end{aligned}$$

and then obtain the optimal bandwidth by numerically solve the problem of minimizing  $AMSE$ . For the LL estimator, I obtain the optimal bandwidth using equation (2.29) that minimizes conditional mean squared error. Lastly, for the HT estimator, I calculate the optimal bandwidth using equation

**Table 2.4:** Different Density Function of the Regressor  $X$

| Panel A: Jump Estimators |              |       |              |              |              |       |        |        |              |        |              |        |
|--------------------------|--------------|-------|--------------|--------------|--------------|-------|--------|--------|--------------|--------|--------------|--------|
|                          | NW           | LL    | HT00         | HT10         | HT11         | HT21  | NW     | LL     | HT00         | HT10   | HT11         | HT21   |
|                          | DGP 1        |       |              |              |              |       |        |        |              |        |              |        |
| Bias                     | 0.719        | 0.452 | 0.711        | 0.313        | 0.819        | 1.136 | 0.442  | 0.051  | 0.469        | -0.147 | 0.179        | -0.212 |
| Variance                 | 0.047        | 0.113 | 0.027        | 0.132        | 0.080        | 0.242 | 0.096  | 0.273  | 0.097        | 18.122 | 0.193        | 1.477  |
| RMSE                     | 0.751        | 0.563 | 0.730        | 0.479        | 0.866        | 1.238 | 0.539  | 0.525  | 0.563        | 4.258  | 0.474        | 1.233  |
| REL_R                    | 1.566        | 1.175 | 1.522        | <b>1.000</b> | 1.807        | 2.582 | 1.137  | 1.107  | 1.187        | 8.978  | <b>1.000</b> | 2.600  |
| AMSE                     | 0.176        | 0.237 | 0.312        | 0.253        | 0.233        | 0.318 | 0.184  | 0.189  | 0.184        | 0.204  | 0.185        | 0.212  |
| REL_A                    | <b>1.000</b> | 1.351 | 1.774        | 1.441        | 1.323        | 1.810 | 1.001  | 1.029  | <b>1.000</b> | 1.111  | 1.008        | 1.153  |
|                          | DGP 2        |       |              |              |              |       |        |        |              |        |              |        |
| Bias                     | -0.337       | 0.008 | -0.303       | 0.006        | -0.011       | 0.069 | -0.420 | -0.010 | -0.416       | 0.184  | -0.021       | 0.190  |
| Variance                 | 0.052        | 0.233 | 0.049        | 0.243        | 0.122        | 0.605 | 0.083  | 0.313  | 0.083        | 45.019 | 0.292        | 3.523  |
| RMSE                     | 0.407        | 0.483 | 0.375        | 0.493        | 0.350        | 0.781 | 0.509  | 0.559  | 0.506        | 6.710  | 0.541        | 1.886  |
| REL_R                    | 1.162        | 1.380 | 1.071        | 1.410        | <b>1.000</b> | 2.232 | 1.005  | 1.104  | <b>1.000</b> | 13.253 | 1.068        | 3.725  |
| AMSE                     | 0.141        | 0.156 | 0.134        | 0.150        | 0.147        | 0.156 | 0.178  | 0.187  | 0.177        | 0.243  | 0.201        | 0.207  |
| REL_A                    | 1.053        | 1.167 | <b>1.000</b> | 1.118        | 1.093        | 1.160 | 1.002  | 1.053  | <b>1.000</b> | 1.371  | 1.136        | 1.166  |
|                          | DGP 3        |       |              |              |              |       |        |        |              |        |              |        |
| Bias                     | -0.337       | 0.008 | -0.303       | 0.006        | -0.011       | 0.069 | -0.420 | -0.010 | -0.416       | 0.184  | -0.021       | 0.190  |
| Variance                 | 0.052        | 0.233 | 0.049        | 0.243        | 0.122        | 0.605 | 0.083  | 0.313  | 0.083        | 45.019 | 0.292        | 3.523  |
| RMSE                     | 0.407        | 0.483 | 0.375        | 0.493        | 0.350        | 0.781 | 0.509  | 0.559  | 0.506        | 6.710  | 0.541        | 1.886  |
| REL_R                    | 1.162        | 1.380 | 1.071        | 1.410        | <b>1.000</b> | 2.232 | 1.005  | 1.104  | <b>1.000</b> | 13.253 | 1.068        | 3.725  |
| AMSE                     | 0.141        | 0.156 | 0.134        | 0.150        | 0.147        | 0.156 | 0.178  | 0.187  | 0.177        | 0.243  | 0.201        | 0.207  |
| REL_A                    | 1.053        | 1.167 | <b>1.000</b> | 1.118        | 1.093        | 1.160 | 1.002  | 1.053  | <b>1.000</b> | 1.371  | 1.136        | 1.166  |
|                          | DGP 4        |       |              |              |              |       |        |        |              |        |              |        |
| Bias                     | -0.337       | 0.008 | -0.303       | 0.006        | -0.011       | 0.069 | -0.420 | -0.010 | -0.416       | 0.184  | -0.021       | 0.190  |
| Variance                 | 0.052        | 0.233 | 0.049        | 0.243        | 0.122        | 0.605 | 0.083  | 0.313  | 0.083        | 45.019 | 0.292        | 3.523  |
| RMSE                     | 0.407        | 0.483 | 0.375        | 0.493        | 0.350        | 0.781 | 0.509  | 0.559  | 0.506        | 6.710  | 0.541        | 1.886  |
| REL_R                    | 1.162        | 1.380 | 1.071        | 1.410        | <b>1.000</b> | 2.232 | 1.005  | 1.104  | <b>1.000</b> | 13.253 | 1.068        | 3.725  |
| AMSE                     | 0.141        | 0.156 | 0.134        | 0.150        | 0.147        | 0.156 | 0.178  | 0.187  | 0.177        | 0.243  | 0.201        | 0.207  |
| REL_A                    | 1.053        | 1.167 | <b>1.000</b> | 1.118        | 1.093        | 1.160 | 1.002  | 1.053  | <b>1.000</b> | 1.371  | 1.136        | 1.166  |
|                          | DGP 4        |       |              |              |              |       |        |        |              |        |              |        |
| Bias                     | -0.337       | 0.008 | -0.303       | 0.006        | -0.011       | 0.069 | -0.420 | -0.010 | -0.416       | 0.184  | -0.021       | 0.190  |
| Variance                 | 0.052        | 0.233 | 0.049        | 0.243        | 0.122        | 0.605 | 0.083  | 0.313  | 0.083        | 45.019 | 0.292        | 3.523  |
| RMSE                     | 0.407        | 0.483 | 0.375        | 0.493        | 0.350        | 0.781 | 0.509  | 0.559  | 0.506        | 6.710  | 0.541        | 1.886  |
| REL_R                    | 1.162        | 1.380 | 1.071        | 1.410        | <b>1.000</b> | 2.232 | 1.005  | 1.104  | <b>1.000</b> | 13.253 | 1.068        | 3.725  |
| AMSE                     | 0.141        | 0.156 | 0.134        | 0.150        | 0.147        | 0.156 | 0.178  | 0.187  | 0.177        | 0.243  | 0.201        | 0.207  |
| REL_A                    | 1.053        | 1.167 | <b>1.000</b> | 1.118        | 1.093        | 1.160 | 1.002  | 1.053  | <b>1.000</b> | 1.371  | 1.136        | 1.166  |
|                          | DGP 4        |       |              |              |              |       |        |        |              |        |              |        |
| Bias                     | -0.337       | 0.008 | -0.303       | 0.006        | -0.011       | 0.069 | -0.420 | -0.010 | -0.416       | 0.184  | -0.021       | 0.190  |
| Variance                 | 0.052        | 0.233 | 0.049        | 0.243        | 0.122        | 0.605 | 0.083  | 0.313  | 0.083        | 45.019 | 0.292        | 3.523  |
| RMSE                     | 0.407        | 0.483 | 0.375        | 0.493        | 0.350        | 0.781 | 0.509  | 0.559  | 0.506        | 6.710  | 0.541        | 1.886  |
| REL_R                    | 1.162        | 1.380 | 1.071        | 1.410        | <b>1.000</b> | 2.232 | 1.005  | 1.104  | <b>1.000</b> | 13.253 | 1.068        | 3.725  |
| AMSE                     | 0.141        | 0.156 | 0.134        | 0.150        | 0.147        | 0.156 | 0.178  | 0.187  | 0.177        | 0.243  | 0.201        | 0.207  |
| REL_A                    | 1.053        | 1.167 | <b>1.000</b> | 1.118        | 1.093        | 1.160 | 1.002  | 1.053  | <b>1.000</b> | 1.371  | 1.136        | 1.166  |

| Panel B: Regression Estimators |                 |              |        |        |              |        |              |        |              |        |        |        |
|--------------------------------|-----------------|--------------|--------|--------|--------------|--------|--------------|--------|--------------|--------|--------|--------|
|                                | NW              | LL           | HT00   | HT10   | HT11         | HT21   | NW           | LL     | HT00         | HT10   | HT11   | HT21   |
|                                | DGP 1- Negative |              |        |        |              |        |              |        |              |        |        |        |
| Bias                           | -0.360          | -0.020       | -0.347 | -0.048 | -0.132       | 0.033  | 0.359        | 0.432  | 0.364        | 0.265  | 0.687  | 1.169  |
| Variance                       | 0.013           | 0.058        | 0.014  | 0.053  | 0.029        | 0.091  | 0.035        | 0.058  | 0.014        | 0.082  | 0.053  | 0.154  |
| RMSE                           | 0.378           | 0.242        | 0.366  | 0.235  | 0.216        | 0.304  | 0.405        | 0.494  | 0.383        | 0.391  | 0.724  | 1.233  |
| REL_R                          | 1.747           | 1.119        | 1.694  | 1.086  | <b>1.000</b> | 1.407  | 1.059        | 1.290  | <b>1.000</b> | 1.021  | 1.893  | 3.223  |
|                                | DGP 2- Positive |              |        |        |              |        |              |        |              |        |        |        |
| Bias                           | -0.288          | -0.043       | -0.349 | 0.146  | -0.180       | 0.181  | 0.153        | 0.007  | 0.120        | 0.000  | -0.001 | -0.031 |
| Variance                       | 0.062           | 0.109        | 0.052  | 17.779 | 0.081        | 0.413  | 0.032        | 0.162  | 0.043        | 0.315  | 0.110  | 1.069  |
| RMSE                           | 0.380           | 0.333        | 0.417  | 4.218  | 0.336        | 0.668  | 0.236        | 0.403  | 0.240        | 0.561  | 0.332  | 1.034  |
| REL_R                          | 1.144           | <b>1.000</b> | 1.253  | 12.681 | 1.010        | 2.007  | <b>1.000</b> | 1.710  | 1.017        | 2.381  | 1.409  | 4.389  |
|                                | DGP 3- Positive |              |        |        |              |        |              |        |              |        |        |        |
| Bias                           | 0.252           | -0.014       | 0.180  | -0.008 | -0.003       | -0.050 | -0.085       | -0.006 | -0.123       | -0.002 | -0.014 | 0.019  |
| Variance                       | 0.025           | 0.101        | 0.032  | 0.098  | 0.064        | 0.527  | 0.026        | 0.124  | 0.016        | 0.138  | 0.057  | 0.081  |
| RMSE                           | 0.297           | 0.319        | 0.254  | 0.312  | 0.253        | 0.728  | 0.182        | 0.353  | 0.177        | 0.371  | 0.239  | 0.285  |
| REL_R                          | 1.174           | 1.259        | 1.004  | 1.233  | <b>1.000</b> | 2.874  | 1.032        | 1.994  | <b>1.000</b> | 2.100  | 1.353  | 1.615  |
|                                | DGP 4- Positive |              |        |        |              |        |              |        |              |        |        |        |
| Bias                           | 0.351           | 0.000        | 0.347  | -0.219 | 0.016        | -0.203 | -0.069       | -0.011 | -0.068       | -0.034 | -0.005 | -0.013 |
| Variance                       | 0.049           | 0.146        | 0.049  | 43.786 | 0.160        | 1.309  | 0.034        | 0.161  | 0.035        | 1.406  | 0.126  | 2.170  |
| RMSE                           | 0.414           | 0.382        | 0.412  | 6.619  | 0.400        | 1.162  | 0.197        | 0.401  | 0.198        | 1.186  | 0.354  | 1.473  |
| REL_R                          | 1.084           | <b>1.000</b> | 1.078  | 17.319 | 1.046        | 3.039  | <b>1.000</b> | 2.031  | 1.003        | 6.006  | 1.795  | 7.460  |

(2.17).

For the jackknife cross-validation method, I choose a globally optimal bandwidth by minimizing the sum of squared deviations between the observed outcomes and the regression estimates, which is obtained by applying respective kernel estimators, the LL, NW, or HT estimators, to leave-one-out samples. As described in equation 2.23 of [Li and Racine, 2007](#), the  $h$  is chosen by minimizing the objective function,

$$CV(h) = n^{-1} \sum_{i=1}^n (Y_i - \hat{m}_{-i}(X_i))^2$$

where  $\hat{m}_{-i}^{NW}(x) = \frac{\sum_{l \neq i}^n K(\frac{X_l - X_i}{h}) Y_l}{\sum_{l \neq i}^n K(\frac{X_l - X_i}{h})}$  is the leave-one-out NW estimator of  $m(X_i)$ . Similarly, I can get the leave-one-out LL and HT estimators by replacing the NW estimator with the LL and HT estimators. This is a fully data-driven method that can be easily applied in empirical research. Among all three method, the jackknife cross-validation method is the most time-consuming, which requires  $n^2$  estimations, followed by the cross-validation method using the true regression, which requires  $n$  estimations. The plug-in method uses the least computational time because it does not involve any estimation.

Table 2.5 shows the simulation estimation results of using the plug-in method. The HT estimators have smaller (2 out of the 4 jumps and 5 out of the 8 regressions) or similar biases to the LL estimators. The HT estimators have the smallest RMSE in 4 out of 4 jumps and 6 out of 8 regressions. Table 2.6 shows the simulation results of the jackknife cross-validation method. The HT estimators have smaller (3 out of the 4 jumps and 7 out of the 8 regressions) or similar biases to the LL estimators. The HT estimators have the smallest RMSE in 4 out of the 4 jumps and 6 out of the 8 regressions. Across the three tables, 2.2,2.5,2.6, the results are qualitatively similar.

Kernel plays an important role in nonparametric estimation. In the default setting, I use the Gaussian kernel, which has support on the whole real line. I add another three kernels with compact support: the Epanechnikov kernel, the triangular Kernel, and the rectangular Kernel. The Gaussian kernel assigns weights to all observations and has convenient mathematical properties. The Epanechnikov kernel is optimal in a mean squared error sense for continuous regressions ([Epanechnikov, 1969](#)). [Cheng et al. \(1997\)](#) show that the triangular kernel is the minimax optimal boundary

**Table 2.5:** Select Bandwidth by a Plug-in Method

| Panel A: Jump Estimators |        |       |              |              |       |       |        |       |              |        |              |        |
|--------------------------|--------|-------|--------------|--------------|-------|-------|--------|-------|--------------|--------|--------------|--------|
|                          | NW     | LL    | HT00         | HT10         | HT11  | HT21  | NW     | LL    | HT00         | HT10   | HT11         | HT21   |
|                          |        |       | DGP 1        |              |       |       |        |       | DGP 2        |        |              |        |
| Bias                     | 0.805  | 0.365 | 0.805        | 0.306        | 0.829 | 0.702 | 0.668  | 0.003 | 0.490        | 0.143  | 0.002        | -0.313 |
| Variance                 | 0.035  | 0.085 | 0.035        | 0.102        | 0.056 | 0.089 | 0.059  | 0.269 | 0.074        | 31.806 | 0.238        | 12.181 |
| RMSE                     | 0.827  | 0.467 | 0.827        | 0.442        | 0.862 | 0.763 | 0.711  | 0.519 | 0.561        | 5.640  | 0.487        | 3.503  |
| REL_R                    | 1.872  | 1.058 | 1.872        | <b>1.000</b> | 1.953 | 1.727 | 1.459  | 1.065 | 1.151        | 11.573 | <b>1.000</b> | 7.188  |
| AMSE                     | 0.193  | 0.202 | 0.193        | 0.248        | 0.214 | 0.243 | 0.190  | 0.173 | 0.163        | 0.217  | 0.182        | 0.191  |
| REL_A                    | 1.001  | 1.049 | <b>1.000</b> | 1.284        | 1.112 | 1.260 | 1.169  | 1.060 | <b>1.000</b> | 1.331  | 1.120        | 1.175  |
|                          |        |       | DGP 3        |              |       |       |        |       | DGP 4        |        |              |        |
| Bias                     | -0.717 | 0.006 | -0.193       | -0.030       | 0.015 | 0.065 | -0.592 | 0.010 | -0.417       | 0.151  | 0.013        | 0.313  |
| Variance                 | 0.042  | 0.259 | 0.072        | 16.539       | 0.227 | 1.158 | 0.059  | 0.272 | 0.074        | 7.165  | 0.244        | 3.197  |
| RMSE                     | 0.746  | 0.508 | 0.331        | 4.066        | 0.476 | 1.078 | 0.640  | 0.522 | 0.497        | 2.680  | 0.494        | 1.815  |
| REL_R                    | 2.252  | 1.536 | <b>1.000</b> | 12.280       | 1.438 | 3.255 | 1.295  | 1.056 | 1.007        | 5.423  | <b>1.000</b> | 3.672  |
| AMSE                     | 0.248  | 0.164 | 0.149        | 0.208        | 0.180 | 0.179 | 0.189  | 0.173 | 0.162        | 0.208  | 0.183        | 0.192  |
| REL_A                    | 1.668  | 1.101 | <b>1.000</b> | 1.394        | 1.205 | 1.203 | 1.163  | 1.065 | <b>1.000</b> | 1.283  | 1.126        | 1.185  |

| Panel B: Regression Estimators |        |              |                |              |        |        |              |        |                |        |        |              |
|--------------------------------|--------|--------------|----------------|--------------|--------|--------|--------------|--------|----------------|--------|--------|--------------|
|                                | NW     | LL           | HT00           | HT10         | HT11   | HT21   | NW           | LL     | HT00           | HT10   | HT11   | HT21         |
|                                |        |              | DGP 1-Negative |              |        |        |              |        | DGP 1-Positive |        |        |              |
| Bias                           | -0.421 | -0.071       | -0.421         | -0.145       | -0.350 | -0.246 | 0.844        | 0.269  | 0.844          | 0.143  | 0.588  | 0.127        |
| Variance                       | 0.009  | 0.029        | 0.009          | 0.023        | 0.012  | 0.017  | 0.041        | 0.089  | 0.041          | 0.130  | 0.066  | 0.124        |
| RMSE                           | 0.432  | 0.185        | 0.432          | 0.210        | 0.366  | 0.278  | 0.867        | 0.402  | 0.867          | 0.387  | 0.642  | 0.375        |
| REL_R                          | 2.329  | <b>1.000</b> | 2.329          | 1.130        | 1.976  | 1.499  | 2.314        | 1.072  | 2.314          | 1.033  | 1.713  | <b>1.000</b> |
|                                |        |              | DGP 2-Negative |              |        |        |              |        | DGP 3-Positive |        |        |              |
| 0.215                          | -0.536 | -0.194       | -0.536         | -0.081       | -0.404 | -0.102 | 0.308        | 0.075  | 0.308          | 0.062  | 0.184  | 0.025        |
| Variance                       | 0.028  | 0.060        | 0.028          | 0.089        | 0.047  | 0.092  | 0.013        | 0.029  | 0.013          | 0.038  | 0.019  | 0.033        |
| RMSE                           | 0.561  | 0.312        | 0.561          | 0.309        | 0.458  | 0.320  | 0.328        | 0.185  | 0.328          | 0.204  | 0.231  | 0.183        |
| REL_R                          | 1.816  | 1.011        | 1.816          | <b>1.000</b> | 1.484  | 1.035  | 1.793        | 1.012  | 1.793          | 1.116  | 1.261  | <b>1.000</b> |
|                                |        |              | DGP 3-Negative |              |        |        |              |        | DGP 4-Positive |        |        |              |
| Bias                           | 0.651  | -0.002       | 0.128          | 0.070        | -0.010 | -0.029 | -0.066       | 0.005  | -0.065         | 0.040  | 0.005  | 0.037        |
| Variance                       | 0.011  | 0.125        | 0.040          | 15.198       | 0.122  | 0.291  | 0.031        | 0.136  | 0.031          | 1.243  | 0.106  | 0.874        |
| RMSE                           | 0.659  | 0.353        | 0.238          | 3.898        | 0.349  | 0.540  | 0.188        | 0.368  | 0.188          | 1.115  | 0.326  | 0.935        |
| REL_R                          | 2.770  | 1.483        | <b>1.000</b>   | 16.375       | 1.467  | 2.267  | <b>1.000</b> | 1.964  | 1.002          | 5.945  | 1.736  | 4.986        |
|                                |        |              | DGP 4-Negative |              |        |        |              |        | DGP 4-Positive |        |        |              |
| Bias                           | 0.535  | 0.196        | 0.535          | 0.084        | 0.405  | 0.104  | -0.192       | -0.035 | -0.192         | -0.067 | -0.147 | -0.092       |
| Variance                       | 0.029  | 0.061        | 0.029          | 0.086        | 0.048  | 0.111  | 0.008        | 0.028  | 0.008          | 0.021  | 0.011  | 0.015        |
| RMSE                           | 0.561  | 0.316        | 0.561          | 0.305        | 0.461  | 0.348  | 0.213        | 0.170  | 0.213          | 0.160  | 0.180  | 0.154        |
| REL_R                          | 1.843  | 1.036        | 1.843          | <b>1.000</b> | 1.512  | 1.144  | 1.380        | 1.103  | 1.380          | 1.036  | 1.169  | <b>1.000</b> |

**Table 2.6:** Select Bandwidth by a Jackknife Cross-Validation Method

| Panel A: Jump Estimators       |              |              |                 |        |              |              |              |        |                 |        |              |              |
|--------------------------------|--------------|--------------|-----------------|--------|--------------|--------------|--------------|--------|-----------------|--------|--------------|--------------|
|                                | NW           | LL           | HT00            | HT10   | HT11         | HT21         | NW           | LL     | HT00            | HT10   | HT11         | HT21         |
|                                |              |              | DGP 1           |        |              |              |              |        | DGP 2           |        |              |              |
| Bias                           | 0.755        | 0.682        | 0.566           | 0.376  | 0.704        | 0.275        | 0.507        | 0.000  | 0.522           | -0.041 | 0.103        | -0.263       |
| Variance                       | 0.032        | 0.082        | 0.053           | 0.112  | 0.066        | 0.121        | 0.110        | 0.473  | 0.080           | 1.282  | 0.160        | 0.930        |
| RMSE                           | 0.776        | 0.740        | 0.611           | 0.503  | 0.749        | 0.443        | 0.606        | 0.688  | 0.593           | 1.133  | 0.413        | 1.000        |
| REL_R                          | 1.750        | 1.669        | 1.379           | 1.134  | 1.691        | <b>1.000</b> | 1.465        | 1.664  | 1.436           | 2.740  | <b>1.000</b> | 2.418        |
| AMSE                           | 0.215        | 0.311        | 0.153           | 0.286  | 0.308        | 0.190        | 0.202        | 0.222  | 0.171           | 0.183  | 0.162        | 0.191        |
| REL_A                          | 1.408        | 2.036        | <b>1.000</b>    | 1.875  | 2.016        | 1.245        | 1.249        | 1.369  | 1.060           | 1.133  | <b>1.000</b> | 1.180        |
|                                |              |              | DGP 3           |        |              |              |              |        | DGP 4           |        |              |              |
| Bias                           | -0.231       | 0.010        | -0.307          | -0.007 | -0.015       | 0.046        | -0.448       | -0.017 | -0.449          | 0.006  | -0.123       | 0.225        |
| Variance                       | 0.061        | 0.137        | 0.046           | 0.575  | 0.109        | 0.276        | 0.072        | 0.260  | 0.083           | 0.312  | 0.172        | 0.447        |
| RMSE                           | 0.339        | 0.370        | 0.374           | 0.758  | 0.331        | 0.527        | 0.522        | 0.510  | 0.534           | 0.559  | 0.433        | 0.705        |
| REL_R                          | 1.024        | 1.119        | 1.130           | 2.290  | <b>1.000</b> | 1.593        | 1.207        | 1.179  | 1.233           | 1.291  | <b>1.000</b> | 1.630        |
| AMSE                           | 0.139        | 0.125        | 0.126           | 0.167  | 0.135        | 0.153        | 0.163        | 0.168  | 0.173           | 0.174  | 0.169        | 0.190        |
| REL_A                          | 1.111        | <b>1.000</b> | 1.006           | 1.331  | 1.078        | 1.217        | <b>1.000</b> | 1.029  | 1.060           | 1.066  | 1.033        | 1.161        |
| Panel B: Regression Estimators |              |              |                 |        |              |              |              |        |                 |        |              |              |
|                                | NW           | LL           | HT00            | HT10   | HT11         | HT21         | NW           | LL     | HT00            | HT10   | HT11         | HT21         |
|                                |              |              | DGP 1- Negative |        |              |              |              |        | DGP 1- Positive |        |              |              |
| Bias                           | -0.331       | -0.042       | -0.239          | -0.033 | -0.053       | -0.060       | 0.424        | 0.640  | 0.327           | 0.342  | 0.651        | 0.215        |
| Variance                       | 0.012        | 0.041        | 0.018           | 0.052  | 0.041        | 0.030        | 0.020        | 0.041  | 0.035           | 0.057  | 0.025        | 0.090        |
| RMSE                           | 0.349        | 0.206        | 0.273           | 0.231  | 0.209        | 0.182        | 0.447        | 0.672  | 0.377           | 0.418  | 0.670        | 0.369        |
| REL_R                          | 1.917        | 1.130        | 1.503           | 1.268  | 1.151        | <b>1.000</b> | 1.211        | 1.819  | 1.022           | 1.131  | 1.813        | <b>1.000</b> |
|                                |              |              | DGP 2- Negative |        |              |              |              |        | DGP 2- Positive |        |              |              |
| Bias                           | -0.438       | -0.020       | -0.403          | 0.035  | -0.079       | 0.241        | 0.068        | -0.021 | 0.119           | -0.006 | 0.024        | -0.021       |
| Variance                       | 0.035        | 0.134        | 0.039           | 1.074  | 0.092        | 0.617        | 0.070        | 0.333  | 0.038           | 0.193  | 0.064        | 0.301        |
| RMSE                           | 0.477        | 0.367        | 0.449           | 1.037  | 0.313        | 0.821        | 0.273        | 0.577  | 0.229           | 0.439  | 0.253        | 0.549        |
| REL_R                          | 1.524        | 1.171        | 1.433           | 3.312  | <b>1.000</b> | 2.625        | 1.194        | 2.522  | <b>1.000</b>    | 1.919  | 1.108        | 2.399        |
|                                |              |              | DGP 3- Negative |        |              |              |              |        | DGP 3- Positive |        |              |              |
| Bias                           | 0.160        | -0.025       | 0.214           | -0.023 | -0.023       | -0.039       | -0.072       | -0.015 | -0.093          | -0.030 | -0.038       | 0.008        |
| Variance                       | 0.030        | 0.063        | 0.023           | 0.084  | 0.073        | 0.152        | 0.032        | 0.072  | 0.023           | 0.498  | 0.037        | 0.119        |
| RMSE                           | 0.236        | 0.253        | 0.263           | 0.291  | 0.271        | 0.392        | 0.192        | 0.269  | 0.178           | 0.706  | 0.197        | 0.344        |
| REL_R                          | <b>1.000</b> | 1.074        | 1.115           | 1.235  | 1.148        | 1.665        | 1.078        | 1.509  | <b>1.000</b>    | 3.962  | 1.104        | 1.932        |
|                                |              |              | DGP 4- Negative |        |              |              |              |        | DGP 4- Positive |        |              |              |
| Bias                           | 0.374        | 0.009        | 0.393           | -0.009 | 0.110        | -0.218       | -0.074       | -0.008 | -0.056          | -0.003 | -0.014       | 0.006        |
| Variance                       | 0.040        | 0.132        | 0.038           | 0.158  | 0.082        | 0.205        | 0.032        | 0.117  | 0.044           | 0.138  | 0.082        | 0.209        |
| RMSE                           | 0.424        | 0.363        | 0.438           | 0.397  | 0.306        | 0.502        | 0.192        | 0.342  | 0.216           | 0.371  | 0.286        | 0.457        |
| REL_R                          | 1.386        | 1.186        | 1.433           | 1.297  | <b>1.000</b> | 1.641        | <b>1.000</b> | 1.780  | 1.124           | 1.930  | 1.488        | 2.377        |

kernel. That is, over all regression functions of the same smoothness and among all possible choices of kernels for linear estimators, the triangular kernel minimizes the asymptotic mean-squared error. A typical approach to implement RDD in empirical research is to conduct a local linear or local polynomial estimation through the OLS procedure, which, in essence, uses the rectangular kernel.

Table 2.7 shows the simulation results of using the Epanechnikov kernel. The HT estimators have smaller (1 out of the 4 jumps and 4 out of the 8 regressions) or similar biases to the LL estimators. The HT estimators have the smallest RMSE in 3 out of the 4 jumps and 7 out of the 8 regressions. Table 2.8 shows the simulation results of using the triangular kernel. The HT estimators have smaller (3 out of the 4 jumps and 7 out of the 8 regressions) or similar biases to the LL estimators. The HT estimators have the smallest RMSE in 4 out of the 4 jumps and 8 out of the 8 regressions. Table 2.9 shows the simulation results of using the rectangular kernel. The HT estimators have similar biases as the LL estimators. The HT estimators have the smallest RMSE in 4 out of the 4 jumps and 8 out of the 8 regressions.

Table 2.7: Choose a Epanechnikov Kernel

Panel A: Jump Estimators

|          | NW     | LL    | HT00   | HT10  | HT11   | HT21  | NW     | LL    | HT00   | HT10   | HT11   | HT21   |  |
|----------|--------|-------|--------|-------|--------|-------|--------|-------|--------|--------|--------|--------|--|
|          |        |       | DGP 1  |       |        |       |        |       |        | DGP 2  |        |        |  |
| Bias     | 0.797  | 0.338 | 0.736  | 0.274 | 0.892  | 0.658 | 0.448  | 0.017 | 0.700  | -0.212 | 0.337  | -0.295 |  |
| Variance | 0.031  | 0.150 | 0.028  | 0.194 | 0.052  | 0.093 | 0.083  | 0.285 | 0.055  | 5.626  | 0.097  | 0.643  |  |
| RMSE     | 0.816  | 0.514 | 0.755  | 0.519 | 0.920  | 0.725 | 0.533  | 0.534 | 0.739  | 2.381  | 0.459  | 0.854  |  |
| REL_R    | 1.589  | 1.000 | 1.469  | 1.009 | 1.791  | 1.412 | 1.160  | 1.163 | 1.609  | 5.186  | 1.000  | 1.860  |  |
| AMSE     | 0.230  | 0.247 | 0.184  | 0.180 | 0.189  | 0.274 | 0.167  | 0.165 | 0.178  | 0.200  | 0.170  | 0.193  |  |
| REL_A    | 1.276  | 1.370 | 1.019  | 1.000 | 1.048  | 1.522 | 1.016  | 1.000 | 1.079  | 1.212  | 1.032  | 1.170  |  |
|          |        |       | DGP 3  |       |        |       |        |       |        | DGP 4  |        |        |  |
| Bias     | -0.296 | 0.020 | -0.444 | 0.136 | -0.069 | 0.399 | -0.319 | 0.017 | -0.429 | 0.023  | -0.244 | 0.397  |  |
| Variance | 0.045  | 0.214 | 0.031  | 0.105 | 0.040  | 0.203 | 0.109  | 0.430 | 0.071  | 0.216  | 0.140  | 0.568  |  |
| RMSE     | 0.365  | 0.463 | 0.477  | 0.351 | 0.212  | 0.602 | 0.459  | 0.656 | 0.505  | 0.465  | 0.447  | 0.851  |  |
| REL_R    | 1.717  | 2.182 | 2.247  | 1.653 | 1.000  | 2.835 | 1.028  | 1.469 | 1.131  | 1.042  | 1.000  | 1.907  |  |
| AMSE     | 0.128  | 0.141 | 0.130  | 0.141 | 0.134  | 0.190 | 0.186  | 0.196 | 0.179  | 0.190  | 0.192  | 0.199  |  |
| REL_A    | 1.000  | 1.095 | 1.016  | 1.100 | 1.041  | 1.479 | 1.037  | 1.094 | 1.000  | 1.062  | 1.071  | 1.112  |  |

Panel B: Regression Estimators

|          | NW     | LL     | HT00            | HT10   | HT11   | HT21   | NW     | LL     | HT00   | HT10            | HT11   | HT21   |  |
|----------|--------|--------|-----------------|--------|--------|--------|--------|--------|--------|-----------------|--------|--------|--|
|          |        |        | DGP 1- Negative |        |        |        |        |        |        | DGP 1- Positive |        |        |  |
| Bias     | -0.326 | -0.022 | -0.353          | 0.037  | -0.216 | 0.088  | 0.471  | 0.316  | 0.382  | 0.310           | 0.676  | 0.746  |  |
| Variance | 0.013  | 0.091  | 0.012           | 0.110  | 0.018  | 0.060  | 0.019  | 0.060  | 0.017  | 0.084           | 0.034  | 0.033  |  |
| RMSE     | 0.346  | 0.303  | 0.369           | 0.333  | 0.255  | 0.260  | 0.490  | 0.399  | 0.404  | 0.424           | 0.700  | 0.768  |  |
| REL_R    | 1.354  | 1.187  | 1.446           | 1.305  | 1.000  | 1.017  | 1.227  | 1.000  | 1.013  | 1.063           | 1.752  | 1.922  |  |
|          |        |        | DGP 2- Negative |        |        |        |        |        |        | DGP 2- Positive |        |        |  |
| Bias     | -0.314 | -0.004 | -0.521          | 0.169  | -0.288 | 0.195  | 0.134  | 0.013  | 0.179  | -0.043          | 0.049  | -0.100 |  |
| Variance | 0.048  | 0.169  | 0.029           | 5.501  | 0.061  | 0.554  | 0.035  | 0.119  | 0.026  | 0.125           | 0.037  | 0.096  |  |
| RMSE     | 0.383  | 0.411  | 0.548           | 2.351  | 0.379  | 0.769  | 0.230  | 0.346  | 0.241  | 0.356           | 0.199  | 0.326  |  |
| REL_R    | 1.010  | 1.084  | 1.445           | 6.198  | 1.000  | 2.028  | 1.161  | 1.741  | 1.215  | 1.794           | 1.000  | 1.643  |  |
|          |        |        | DGP 3- Negative |        |        |        |        |        |        | DGP 3- Positive |        |        |  |
| Bias     | 0.212  | -0.028 | 0.309           | -0.153 | -0.021 | -0.359 | -0.085 | -0.008 | -0.135 | -0.017          | -0.090 | 0.040  |  |
| Variance | 0.023  | 0.144  | 0.018           | 0.072  | 0.024  | 0.069  | 0.023  | 0.067  | 0.014  | 0.033           | 0.017  | 0.135  |  |
| RMSE     | 0.260  | 0.381  | 0.337           | 0.308  | 0.158  | 0.445  | 0.174  | 0.258  | 0.179  | 0.182           | 0.159  | 0.369  |  |
| REL_R    | 1.653  | 2.416  | 2.135           | 1.956  | 1.000  | 2.820  | 1.095  | 1.625  | 1.126  | 1.146           | 1.000  | 2.326  |  |
|          |        |        | DGP 4- Negative |        |        |        |        |        |        | DGP 4- Positive |        |        |  |
| Bias     | 0.279  | -0.005 | 0.344           | -0.005 | 0.248  | -0.356 | -0.040 | 0.012  | -0.085 | 0.018           | 0.004  | 0.040  |  |
| Variance | 0.053  | 0.173  | 0.045           | 0.119  | 0.064  | 0.494  | 0.053  | 0.255  | 0.025  | 0.094           | 0.074  | 0.077  |  |
| RMSE     | 0.362  | 0.416  | 0.404           | 0.346  | 0.354  | 0.788  | 0.234  | 0.505  | 0.178  | 0.306           | 0.271  | 0.281  |  |
| REL_R    | 1.047  | 1.205  | 1.169           | 1.000  | 1.026  | 2.279  | 1.313  | 2.830  | 1.000  | 1.718           | 1.522  | 1.575  |  |



**Table 2.8:** Choose a Triangular Kernel

Panel A: Jump Estimators

|          | NW     | LL           | HT00   | HT10  | HT11         | HT21  | NW           | LL           | HT00         | HT10   | HT11         | HT21  |
|----------|--------|--------------|--------|-------|--------------|-------|--------------|--------------|--------------|--------|--------------|-------|
|          |        |              | DGP 1  |       |              |       |              |              | DGP 2        |        |              |       |
| Bias     | 0.645  | 0.314        | 0.613  | 0.664 | 0.816        | 0.528 | 0.484        | -0.012       | 0.783        | -0.110 | 0.376        | 0.427 |
| Variance | 0.034  | 0.170        | 0.027  | 0.098 | 0.053        | 0.079 | 0.077        | 0.431        | 0.048        | 0.237  | 0.097        | 0.774 |
| RMSE     | 0.671  | 0.519        | 0.634  | 0.734 | 0.848        | 0.598 | 0.558        | 0.656        | 0.813        | 0.499  | 0.487        | 0.978 |
| REL_R    | 1.293  | <b>1.000</b> | 1.223  | 1.416 | 1.635        | 1.154 | 1.146        | 1.347        | 1.667        | 1.024  | <b>1.000</b> | 2.007 |
| AMSE     | 0.290  | 0.230        | 0.239  | 0.394 | 0.186        | 0.282 | 0.165        | 0.194        | 0.181        | 0.187  | 0.174        | 0.225 |
| REL_A    | 1.557  | 1.235        | 1.286  | 2.120 | <b>1.000</b> | 1.516 | <b>1.000</b> | 1.174        | 1.094        | 1.128  | 1.050        | 1.360 |
|          |        |              | DGP 3  |       |              |       |              |              | DGP 4        |        |              |       |
| Bias     | -0.292 | 0.006        | -0.378 | 0.211 | -0.054       | 0.434 | -0.304       | -0.006       | -0.617       | 0.118  | -0.132       | 0.077 |
| Variance | 0.045  | 0.143        | 0.037  | 0.120 | 0.040        | 0.122 | 0.114        | 0.285        | 0.052        | 0.217  | 0.143        | 0.229 |
| RMSE     | 0.361  | 0.379        | 0.424  | 0.406 | 0.208        | 0.557 | 0.454        | 0.534        | 0.657        | 0.481  | 0.400        | 0.485 |
| REL_R    | 1.736  | 1.821        | 2.040  | 1.950 | <b>1.000</b> | 2.680 | 1.133        | 1.333        | 1.641        | 1.200  | <b>1.000</b> | 1.210 |
| AMSE     | 0.130  | 0.126        | 0.141  | 0.171 | 0.131        | 0.227 | 0.188        | 0.175        | 0.175        | 0.178  | 0.193        | 0.195 |
| REL_A    | 1.026  | <b>1.000</b> | 1.114  | 1.356 | 1.041        | 1.799 | 1.076        | <b>1.000</b> | <b>1.000</b> | 1.014  | 1.105        | 1.116 |

Panel B: Regression Estimators

|          | NW     | LL     | HT00            | HT10         | HT11         | HT21   | NW     | LL     | HT00            | HT10   | HT11         | HT21   |
|----------|--------|--------|-----------------|--------------|--------------|--------|--------|--------|-----------------|--------|--------------|--------|
|          |        |        | DGP 1- Negative |              |              |        |        |        | DGP 2- Positive |        |              |        |
| Bias     | -0.227 | -0.016 | -0.317          | 0.035        | -0.166       | 0.075  | 0.418  | 0.299  | 0.296           | 0.699  | 0.649        | 0.604  |
| Variance | 0.020  | 0.109  | 0.014           | 0.072        | 0.022        | 0.052  | 0.014  | 0.063  | 0.013           | 0.027  | 0.031        | 0.026  |
| RMSE     | 0.267  | 0.331  | 0.338           | 0.271        | 0.224        | 0.240  | 0.434  | 0.389  | 0.317           | 0.718  | 0.673        | 0.625  |
| REL_R    | 1.193  | 1.479  | 1.510           | 1.209        | <b>1.000</b> | 1.072  | 1.372  | 1.230  | <b>1.000</b>    | 2.266  | 2.124        | 1.973  |
|          |        |        | DGP 2- Negative |              |              |        |        |        | DGP 3- Positive |        |              |        |
| Bias     | -0.344 | 0.006  | -0.575          | 0.060        | -0.324       | -0.503 | 0.141  | -0.006 | 0.208           | -0.051 | 0.051        | -0.076 |
| Variance | 0.044  | 0.241  | 0.025           | 0.140        | 0.059        | 0.077  | 0.033  | 0.187  | 0.022           | 0.093  | 0.037        | 0.688  |
| RMSE     | 0.403  | 0.491  | 0.596           | 0.378        | 0.405        | 0.574  | 0.230  | 0.433  | 0.256           | 0.309  | 0.198        | 0.833  |
| REL_R    | 1.064  | 1.297  | 1.575           | <b>1.000</b> | 1.071        | 1.518  | 1.160  | 2.185  | 1.292           | 1.561  | <b>1.000</b> | 4.206  |
|          |        |        | DGP 3- Negative |              |              |        |        |        | DGP 4- Positive |        |              |        |
| Bias     | 0.213  | -0.001 | 0.275           | -0.182       | -0.057       | -0.379 | -0.079 | 0.004  | -0.103          | 0.030  | -0.110       | 0.055  |
| Variance | 0.024  | 0.075  | 0.021           | 0.049        | 0.027        | 0.048  | 0.021  | 0.073  | 0.017           | 0.071  | 0.013        | 0.075  |
| RMSE     | 0.262  | 0.274  | 0.310           | 0.286        | 0.174        | 0.438  | 0.166  | 0.270  | 0.165           | 0.268  | 0.159        | 0.280  |
| REL_R    | 1.507  | 1.573  | 1.782           | 1.639        | <b>1.000</b> | 2.513  | 1.042  | 1.691  | 1.035           | 1.684  | <b>1.000</b> | 1.755  |
|          |        |        | DGP 4- Negative |              |              |        |        |        | DGP 4- Positive |        |              |        |
| Bias     | 0.263  | 0.002  | 0.526           | -0.105       | 0.132        | -0.035 | -0.041 | -0.004 | -0.091          | 0.013  | 0.000        | 0.041  |
| Variance | 0.054  | 0.124  | 0.027           | 0.150        | 0.077        | 0.130  | 0.059  | 0.168  | 0.023           | 0.061  | 0.064        | 0.097  |
| RMSE     | 0.352  | 0.352  | 0.551           | 0.401        | 0.307        | 0.363  | 0.246  | 0.410  | 0.178           | 0.248  | 0.253        | 0.315  |
| REL_R    | 1.145  | 1.146  | 1.795           | 1.306        | <b>1.000</b> | 1.182  | 1.381  | 2.303  | <b>1.000</b>    | 1.393  | 1.422        | 1.769  |



Over all eight groups of experiments, the HT jump estimators have the smallest RMSE in 29 out of the 32 jump estimation experiments and the HT jump estimators have the smallest RMSE in 50 out of the 64 regression estimation experiments. In summary, the HT estimators perform better than the NW estimators in all cases. Compared to the LL estimators, the HT estimators have similar or, in some cases, smaller bias size. In most cases, the HT estimators have smaller variance and smaller mean squared error. These findings about the finite samples distributions of these estimators – bias, variance and RMSE – agree with my theoretical predictions. In addition, they are not affected by types of regression functions, density functions of regressors, methods of obtaining optimal bandwidth, or choice of kernel functions.

## 2.4 Empirical Illustration

To illustrate the applicability of my estimators in empirical settings, I use data collected by [Litschig and Morrison \(2013\)](#) who use a regression discontinuity model to examine the impact of intergovernmental transfer programs on education and poverty reduction outcomes. I begin with a discussion of the assumptions underlying RD designs and their implications for empirical modeling. I then describe how a typical empirical RD model is estimated to verify these assumptions and address some implementation issues.

### 2.4.1 Assumptions on RDD and Their Implications for Empirical Studies

Identification of RDD depends on several assumptions, which have important implications in empirical studies. The most important assumption is on the regression function associated with the outcome variable. Instead of assuming a specific functional form for the regression function, identification of RD models assumes the existence of a smooth regression at the vicinity of the discontinuity point. This is in contrast with the difference-in-difference (DID) method, where a more stringent condition “equal trend” is required: regression functions before and after an intervention must be the same. The reason for this difference is that RD designs assess a local average treatment effect (LATE).

The second assumption relates to how a treatment is assigned in association with a running

variable. Around the discontinuity point, individuals are similar but receive different treatments based on whether their associated running variable values are above or below the threshold, which determines the treatment group and control group. I assume that the jump in the regression of outcomes is actually caused by a treatment variable rather than other covariates and the running variable has a continuous density. To verify the empirical validity of these assumptions, researchers check that all other covariates across the discontinuity point are continuous to ensure that the running variable, rather than other covariates, is the true impetus for the treatment effect. Researchers normally follow the recommendation from [McCrary \(2008\)](#) to check that the density of the running variable around the discontinuity point is continuous to ensure that no individual endogenously manipulates the running variable. This also trivially satisfies an assumption that sample data exist on both sides of the discontinuity point. This is an important theoretical assumption for regression estimators that use one-sided data (NW or LL) to ensure that the denominator, which consists of a kernel density estimator, is not equal to zero.

Lastly, an important assumption for identification is knowledge of the point of discontinuity – that is, the discontinuity of treatment status when the running variable crosses the discontinuity point is known to the econometrician: in sharp RDD, the jump in the probability of receiving the treatment is equal to one while in fuzzy RDD this jump in probability is between 0 and 1. Without this underlying assumption, the jump in the expected outcome will not be assigned to any treatment. Moreover, as shown in equation (2.30), the assumption ensures the denominator of the estimator is not zero.

One common practice of implementing empirical RDD is that, instead of using all data to perform a nonparametric estimation, researchers often use a compactly supported kernel by arbitrarily restricting a running variable to a small range of values around the threshold. They also obtain estimation with different bandwidths to ensure the estimation of the parameter of interest is not sensitive to different bandwidths. However, this procedure has several drawbacks. First, they do not know which bandwidth is optimal, and therefore, which estimate is the best. Second, when there is a high volume of data in a small range of discrete running variable values (running variables are often discrete, such as age, test scores, or the number of employees of firms), and the variance

of the estimate is large. Third, data away from the threshold have information that can influence the estimated regression and, therefore, affect the jump size estimate. By contrast, my estimation procedure allows the bandwidth to be chosen optimally and any kernel function including kernel without a compact support such as Gaussian kernel to be used.

### 2.4.2 A Typical Empirical RD Model Estimation Procedure

The running variable  $X$  is not an object of direct interest, but it is of interest insofar as the expected value of the outcome variable  $Y$  has a jump at a particular value of  $X$ . Normally, economists are interested in regression slopes, but not in this case, where the primary focus is on the regression jump. At the discontinuity point, another variable  $D$  – the treatment variable – experiences a jump, and the jump in the expectation of  $Y$  is thought of as the average treatment effect (ATE) of the treatment  $D$  on the outcome  $Y$  under the assumptions discussed in the previous subsection. Empirical work using RDD often involve the following procedures:

1. Use a scatter plot to visually check if there is a jump in the regression function of the outcome variable and treatment variable with respect to the running variable at the discontinuity point.
2. Perform regression estimation on the outcome and treatment variable.
3. Conduct a robustness test to ensure that no other covariates have a jump at the discontinuity point.
4. Conduct a test suggested by [McCrary \(2008\)](#) for checking that the density of the running variable is continuous at the discontinuity point.
5. Repeat estimation with a truncated sample to ensure that points far away from the discontinuity point do not exert undue influence.

Abstracting from specific empirical context or designs – sharp or fuzzy – estimations in RD models can be simply categorized into regression and density discontinuity estimation at the discontinuity point. The regression discontinuity estimation includes estimation of the regression discontinuity of the outcome variables, the treatment variable, and, sometimes the pre-treatment variables, with

respect to the running variable, while the density estimation includes the discontinuity in the density of the running variable. In many studies (for example, [Litschig and Morrison, 2013](#), [Matsudaira, 2008](#)), a local least squares method is used for regression estimation and a histogram or McCrary's procedure is used to check for a discontinuity in the density.

### 2.4.3 An Empirical Example

[Litschig and Morrison \(2013\)](#) exploit an opportunity provided by the passing of Decree 188181, a federal funds transfer plan in Brazil, which stipulates that federal funds – FPM (the federal *Fundo de Participação dos Municípios*) – must be distributed to local communities according to municipal population. Per capita spending on intergovernmental transfers thus undergo a jump at several population thresholds, which constitutes a sharp RD design with multiple thresholds. The treatment variable is per capita spending and the running variable is the population of the municipality. The impacts of the treatment are measured by outcome variables such as years of schooling, literacy rate, poverty rate and political party reelection rate. The study estimates the jump in the conditional means of the outcome variables using least squares procedure and the jump in the density of the running variable using the method recommended by [McCrary \(2008\)](#). To show the robustness of their results, Litschig and Morrison vary the choice of bandwidth, use various truncation of their sample with respect to the percentage of population away from the discontinuity point, and try different functional forms, including linear, quadratic, cubic and quartic. Compared to their approach, my approach is more flexible. Instead of arbitrarily choosing the bandwidth, I choose an optimal bandwidth based on the sample; instead of trying different functional forms, I do not specify any functional form.

A direct comparison between my estimation results with Litschig and Morrison's is not possible because I choose different parameters, such as bandwidth. Nevertheless, I try to produce comparable results using my estimators. Since they have estimates with and without additional covariates, I estimate the jump in the regression of the running variable with and without extra covariates.

For estimation without extra covariates, I consider

$$Y = \begin{cases} m^+(X) + \epsilon, & X \geq 0 \\ m^-(X) + \epsilon, & X < 0 \end{cases}, \quad (2.34)$$

where  $X$  is the running variable: the population of a municipality and  $Y$  represents either a treatment or the outcome variable. The jump size is estimated by  $\hat{J}(0) = \lim_{x \downarrow 0} \hat{m}^+(x) - \lim_{x \uparrow 0} \hat{m}^-(x)$ .

For estimation that includes additional continuous covariates, I propose the following additive model,

$$Y = m_1(X_1) + m_2(X_2) + \dots + m_p(X_p) + \epsilon, \quad (2.35)$$

where  $m_1(x) = \begin{cases} m_1^+(x) & \text{if } x \geq 0 \\ m_1^-(x) & \text{if } x < 0 \end{cases}$  and  $m_2, \dots, m_p$  are sufficiently smooth. To estimate the jump,

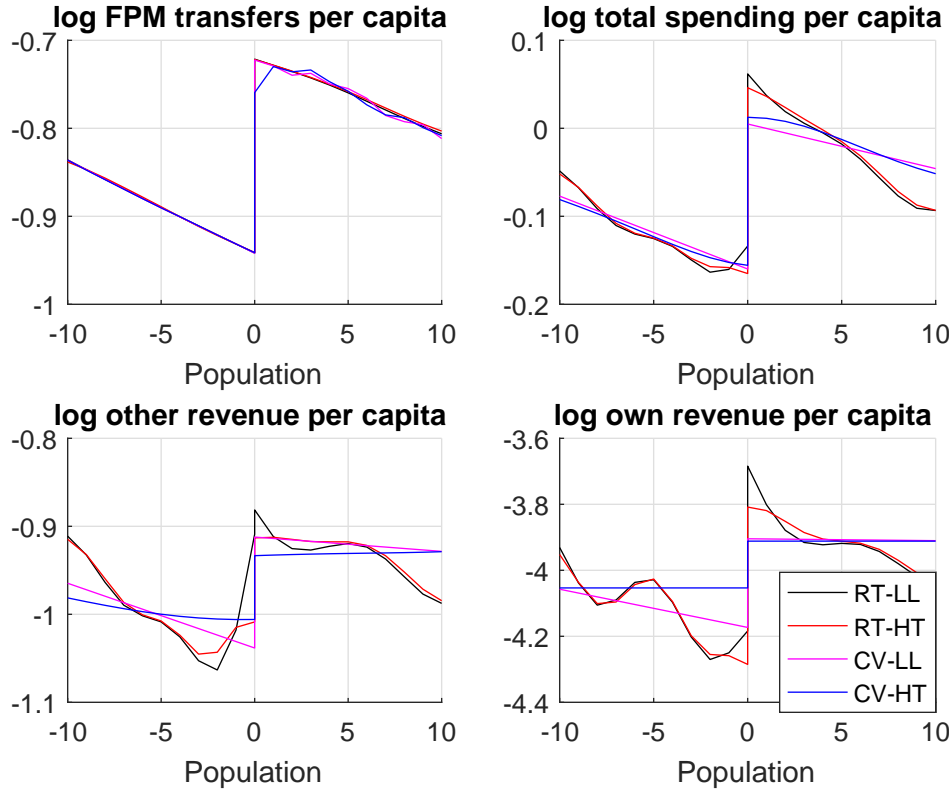
I compare the results from the local linear estimators by [Calonico et al., 2018](#) with the two-stage estimators I proposed – that uses the marginal integration estimator as the first-stage estimator and the Hestenes estimator as the second-stage estimator. The details of these two estimators are in Chapter 3.

I use two methods to obtain an optimal bandwidth  $h$ : a plug-in rule of thumb and a jackknife cross-validation method. For the rule-of-thumb method,  $h = n^{-1/5} \text{std}(X)$ . For the jackknife cross-validation method, I choose  $h$  to minimize the sum of squared leave-one-out residuals using observations to the right and to the left of the threshold respectively.

The estimation procedure is straightforward. I use Hestenes estimators to estimate the jump in the regressions of the treatment variables and outcome variables, and I use Hestenes density estimators proposed by [Mynbaev and Martins-Filho \(2018\)](#) to estimate the jump in the density of the running variable.

Figure 2.4 shows the estimates of the treatment variable from the model 2.34. It depicts four scenarios: RT-LL, the local linear estimator with the bandwidth obtained by the rule of thumb method; RT-HT, the Hestenes estimator with the bandwidth obtained by the rule of thumb method; CV-LL, the local linear estimator with the bandwidth obtained by the jackknife cross-validation

**Figure 2.4:** Impacts on Total Spending, Other Revenue, and Own Revenue



RT-LL: the local linear estimator with the bandwidth obtained by the Rule of Thumb method; RT-HT: the Hestenes estimator with the bandwidth obtained by the Rule of Thumb method; CV-LL: the local linear estimator with the bandwidth obtained by the cross-validation method. CV-HT: the Hestenes estimator with the bandwidth obtained by the cross-validation method.

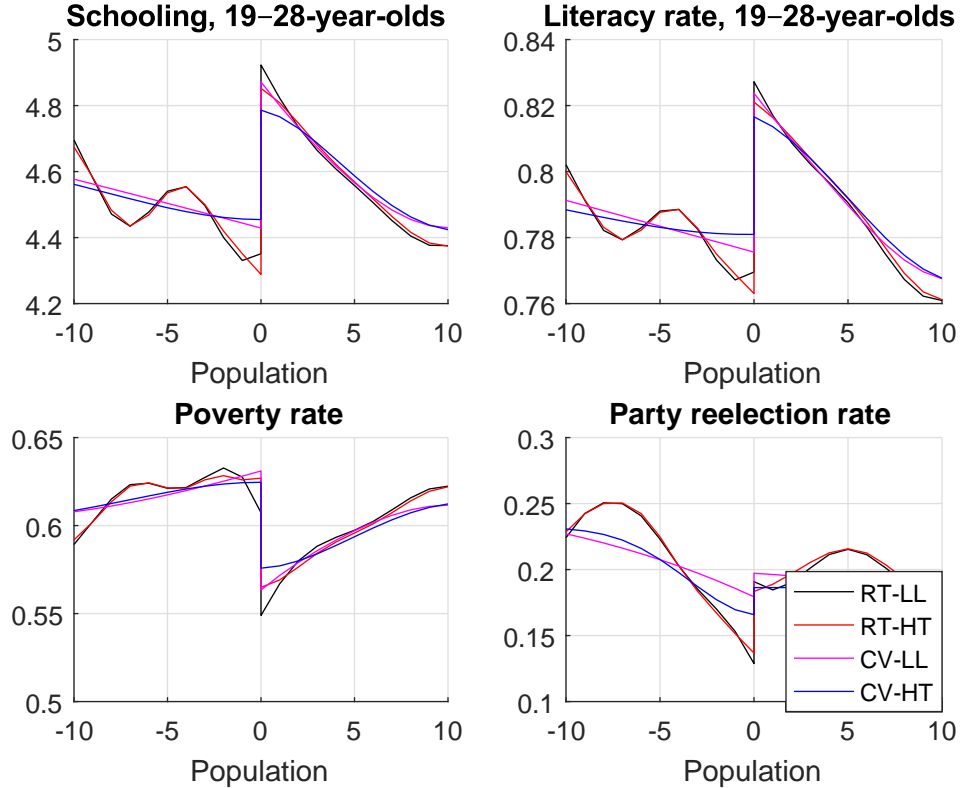
method. CV-HT, the Hestenes estimator with the bandwidth obtained by the cross-validation method. The results shows the treatment is assigned: there is a jump in total spending per capita. This agrees with the authors' explanation that the federal transfer caused an increase in per capita public spending without crowding out other types of spending.

Figure 2.5 shows the estimates of the outcome variables from the model 2.34. It depicts the same four scenarios as Figure 2.4. The results show that there is an increase in years of schooling for the 19-28 age group, a decrease in the illiteracy rate and poverty rates, and an increase in the reelection rate of the incumbent party. These agree with the authors' findings.

Table 2.10 displays the jump estimates of the treatment and outcome variables from the univariate regression model 2.34. To compare with their results, I show their corresponding estimates.



**Figure 2.5:** Impacts on Schooling, Literacy, Poverty, and Party Reelection



RT-LL: local linear estimators with the bandwidth obtained by the Rule of Thumb method; RT-HT: Hestenes estimators with the bandwidth obtained by the Rule of Thumb method; CV-LL: local linear estimators with the bandwidth obtained by the cross-validation method. CV-HT: Hestenes estimators with the bandwidth obtained by the cross-validation method.

The obvious difference is that I choose the bandwidth optimally by the rule-of-thumb or jackknife cross-validation method while they choose bandwidth arbitrarily. Given the estimator and the bandwidth selection method, I know optimal estimates while they do not know which estimates are optimal. The estimates are qualitatively similar but quantitatively different. The estimates from LL and HT differ less than the estimates from different bandwidth selection methods. Since the rule-of-thumb method only uses the sample data of the running variable to determine the bandwidth whereas the cross-validation method uses the sample data and estimation methods to determine the bandwidth, the latter, in general, provides better bandwidth selection. As demonstrated from the Monte Carlo studies in last section, HT estimators have smaller MSE than LL estimators, which means that HT estimators can provide more precise estimates and powerful tests than LL estimators. In this specific example, the estimates provided by the HT estimators using the cross-validation seem to have smaller effects compared to the estimates provided by Litschig and Morrison.

Table 2.11 displays the jump estimates of the treatment and outcome variables from the multivariate regression model 2.35. Compared the estimates from the univariate regression model, the effects become smaller. However, the HT estimates from the multivariate regression model (column 2 in Table 2.11) are close to the estimates of HT estimators from the univariate regression model (column 4 in Table 2.10) than the LL estimates.

Finally, I check the continuity of the density of the population running variable. Figure 2.6, which corresponds to their online appendix Figure 1, graphically presents estimates from the LL density estimators by [Cheng \(1994\)](#) and the Hestenes density estimators by [Mynbaev and Martins-Filho \(2018\)](#). The two estimates almost completely overlap. Since the jump density estimators by [McCrary \(2008\)](#) are based on the LL density estimators by [Cheng \(1994\)](#), it is not surprising that Figure 2.6 looks very similar to Litschig and Morrison's analogous Figure (online appendix Figure 1). Table 2.12 compares my density estimates with theirs. The estimates on the jump from the three estimators are slightly different, but they all lead to the failing to reject the null hypothesis that there are no discontinuities at any of the six cutoff points. The discrepancy between the estimates from the LL and McCrary estimators could be explained by differences in the bin size

**Table 2.10:** Univariate Regression Without Additional Covariates

| Estimates on Treatment and Outcome Variables |               |       |        |       |                  |          |          |       |                     |       |        |
|--|---------------|-------|--------|-------|------------------|----------|----------|-------|---------------------|-------|--------|
| Estimators                                   | LL            |       | HT     |       | LL               |          | HT       |       | Litschig & Morrison |       |        |
| Bandwidth                                    | Rule-of-thumb |       |        |       | Cross-validation |          |          |       | Arbitrary           |       |        |
| h  | L             | R     | L      | R     | L                | R        | L        | R     |                     |       |        |
| <b>Treatment</b>                             |               |       |        |       |                  |          |          |       |                     |       |        |
| h  | 0.195         | 0.214 | 0.195  | 0.214 | $\infty$         | 0.077    | $\infty$ | 0.074 | 2                   | 3     | 4      |
| Spending                                     | 0.196         |       | 0.211  |       | 0.165            |          | 0.168    |       | 0.158               | 0.161 | 0.197  |
| <b>Outcomes</b>                              |               |       |        |       |                  |          |          |       |                     |       |        |
| h  | 0.195         | 0.214 | 0.195  | 0.214 | $\infty$         | $\infty$ | 5.461    | 2.794 | 2                   | 3     | 4      |
| Schooling                                    | 0.572         |       | 0.564  |       | 0.443            |          | 0.332    |       | 0.322               | 0.516 | 0.528  |
| h  | 0.195         | 0.214 | 0.195  | 0.214 | $\infty$         | $\infty$ | 3.230    | 2.418 | 2                   | 3     | 4      |
| Literacy                                     | 0.058         |       | 0.058  |       | 0.048            |          | 0.036    |       | 0.057               | 0.063 | 0.059  |
| h  | 0.195         | 0.214 | 0.195  | 0.214 | $\infty$         | $\infty$ | 1.201    | 4.196 | 2                   | 3     | 4      |
| Poverty                                      | -0.058        |       | -0.062 |       | -0.068           |          | -0.049   |       | -0.037              | -0.06 | -0.054 |
| h  | 0.195         | 0.214 | 0.195  | 0.214 | $\infty$         | 0.29     | 3.124    | 0.253 | 2                   | 4     | 10     |
| Reelection                                   | 0.062         |       | 0.047  |       | 0.018            |          | 0.021    |       | 0.119               | 0.086 | 0.106  |

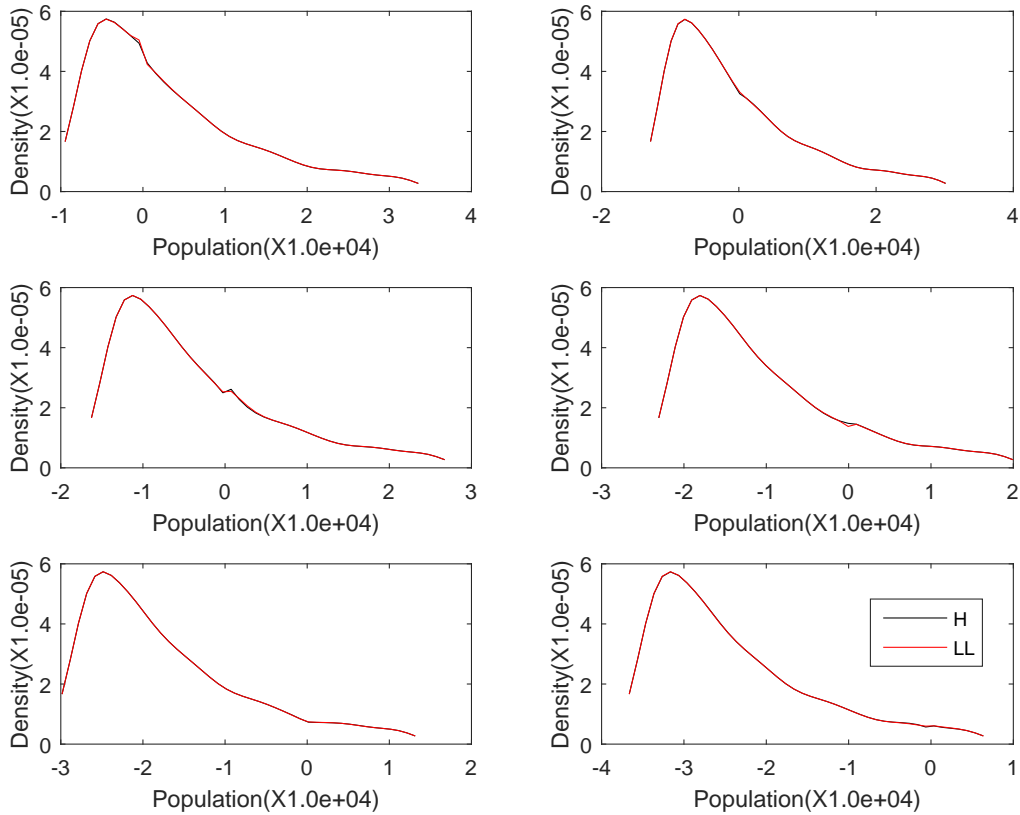
Note: Estimates of Litschig & Morrison are from [Litschig and Morrison \(2013\)](#): Spending from Table 4 row 1, Schooling from Table 5 row 1, Literacy from Table 7 row 1, Poverty from Table 8 row 1, Reelection from Table 9 row 1. The running variable, pscore, is the population of municipality at the neighborhood of the threshold measured in percentage. The rows indicated by h are the bandwidth values. The value  $\infty$  is resulted from the Cross-validation method, which chooses a large values beyond the range of the running valuable. L indicates the bandwidth obtained for the regression to the left of the threshold and R indicates the bandwidth obtained for the regression to the right of the threshold.

**Table 2.11:** Multivariate Regression With Additional Covariates

| Estimates on Treatment and Outcome Variables |               |        |                     |        |        |
|--|---------------|--------|---------------------|--------|--------|
| Estimators                                   | LL            | HT     | Litschig & Morrison |        |        |
| Bandwidth                                    | Rule-of-thumb |        | Arbitrary           |        |        |
| <b>Treatment</b>                             |               |        |                     |        |        |
| h  | 3.257         | 3.257  | 2                   | 3      | 4      |
| Spending                                     | 0.124         | 0.197  | 0.191               | 0.145  | 0.167  |
| <b>Outcomes</b>                              |               |        |                     |        |        |
| h  | 3.257         | 3.257  | 2                   | 3      | 4      |
| Schooling                                    | 0.200         | 0.314  | 0.225               | 0.301  | 0.275  |
| h  | 3.257         | 3.257  | 2                   | 3      | 4      |
| Literacy                                     | 0.024         | 0.039  | 0.047               | 0.049  | 0.041  |
| h  | 3.257         | 3.257  | 2                   | 3      | 4      |
| Poverty                                      | -0.017        | -0.048 | -0.064              | -0.051 | -0.037 |
| h  | 3.257         | 3.257  | 2                   | 4      | 10     |
| Reelection                                   | 0.026         | 0.063  | 0.186               | 0.106  | 0.103  |

Note: Estimates of Litschig & Morrison are from [Litschig and Morrison \(2013\)](#): Spending from Table 4 row 1, Schooling from Table 5 row 1, Literacy from Table 7 row 1, Poverty from Table 8 row 1, Reelection from Table 9 row 1. The running variable, pscore, is the population of municipality at the neighborhood of the threshold measured in percentage. The rows indicated by h are the bandwidth values.

**Figure 2.6:** Population Density Estimated by LL Estimator and Hestenes Estimator



The population is centered at the thresholds.

and choices of bandwidth.

In summary, the empirical example shows that the common practice of using OLS procedures with arbitrarily bandwidth selection to implement RDD provides good preliminary results that are similar to the estimates from the LL and HT estimators. However, researchers need to be aware the limitations of this method: the rectangular kernels and arbitrary bandwidth selection. To avoid this shortcomings and obtain more precise estimates, researchers should consider to use the HT estimators since they have smaller MSE than the LL estimators in most of cases. Including additional covariates, estimates from the HT estimators are similar to its univariate counterparts while the LL estimators give quite different estimates. This prescient finding will be further confirmed in Chapter 3 that the HT estimators with additional covariates provides the smallest MSE, followed by the univariate HT estimators, the LL estimators with additional covariates, and the univariate LL estimators. This gives another reason to use the HT estimators especially when covariates (or

**Table 2.12:** Density Estimates on Population

| The Cutoffs | Local Linear | Hestenes | McCrary | (SE of McCrary) |
|-------------|--------------|----------|---------|-----------------|
| 10188       | -0.0288      | 0.0456   | -0.0720 | 0.095           |
| 13584       | 0.0331       | 0.0291   | 0.0110  | 0.111           |
| 16980       | -0.4656      | -0.3222  | 0.1800  | 0.136           |
| 23772       | 0.1955       | 0.0393   | 0.0540  | 0.174           |
| 30564       | 0.1336       | -0.1626  | -0.0110 | 0.269           |
| 37356       | -0.6823      | -0.5325  | 0.3500  | 0.357           |

pre-treatment variables) are not available.

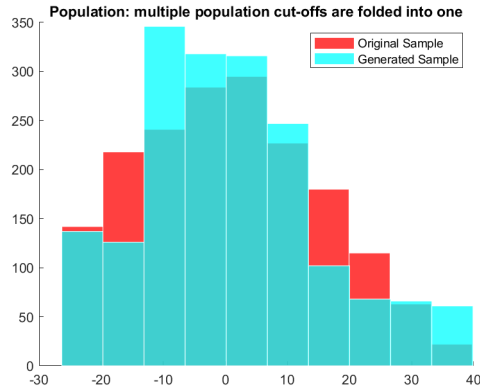
## 2.5 Empirically Motivated Simulations

Empirical exercises are useful as indicators to show the applicability of an estimator and to compare qualitative/quantitative results between estimators, but they are not useful as indicators to compare relative estimation performance. We can not verify whether one estimator performs better than the other empirically except we know something about truth. RMSE in Monte Carlo simulations is better metric and comparing asymptotic biases and distributions is also desirable. Here, to relieve the concern that regression functions are arbitrarily chosen in Monte Carlo studies, I conduct empirically motivated simulations to verify the findings.

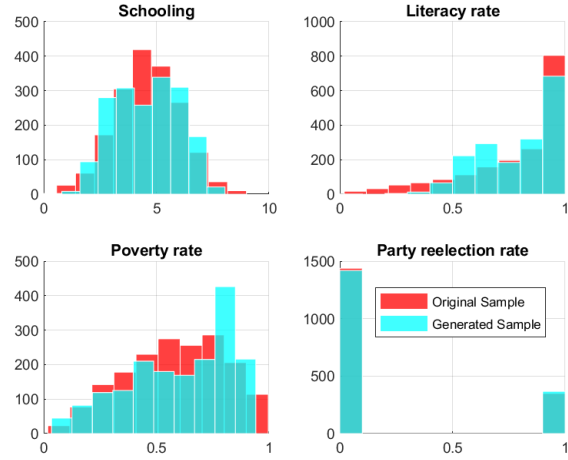
I use the estimated regression from the empirical example of [Litschig and Morrison \(2013\)](#) as the true regression to generate samples. At the same time, since the running variable and the outcome variables have different marginal distribution on the different intervals of the real line, I make sure the generated sample have the similar marginal distributions as their counterparts from the empirical sample. Using the generated samples, I compare the finite sample performance of NW, LL and the Hestenes Estimators.

First, I generate samples of the running variable. The running variable, *pscore*, is the population of a county or municipality, which measures in percentage how far away a county's population is over/under the cutoff and folds multiple population cut-offs into one. cut-off. The distribution of *pscore* looks more like a truncated normal, so I use a truncated normal to fit the marginal distribution of the running variable.

**Figure 2.7:** The Histogram of Original Samples and the Generated Samples  
The running variable



The Outcome Variables



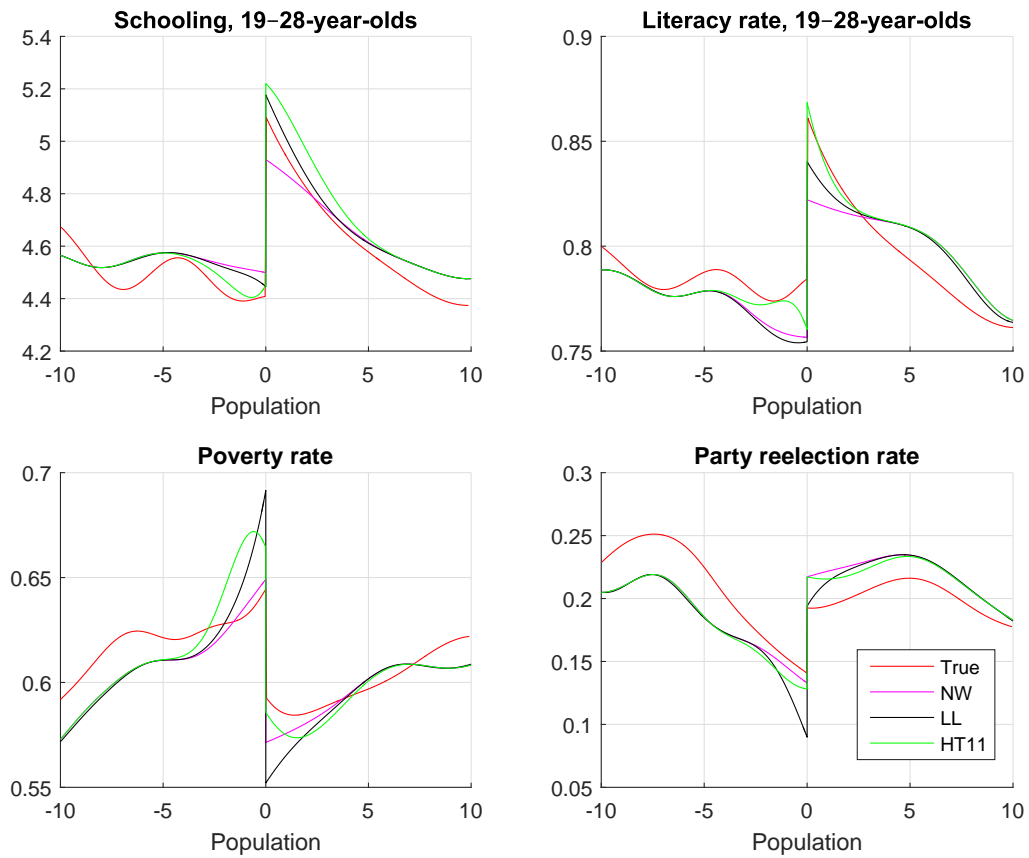
Second, I generate samples of the outcome variables that use the estimated regression as true regression. Years of schooling are rational numbers between 0 and 9, so I use truncated normal to generate a sample of years of schooling. The literacy rate is between 0 and 1 and the poverty rate is between 0 and 1. Following the paper by [Ferrari and Cribari-Neto \(2004\)](#), I take parameters from the distribution of the literacy rate and the poverty rate, and use Beta distribution to generate samples of these two outcome variables respectively. Finally, the party reelected rate is either 0 or 1. I use binomial distribution to fit the party reelected rate. Figure 2.7 shows that the generated running variable has almost the same marginal distribution as the real running variable, and the generated outcome variables have similar marginal distributions as the real outcome variables.

Figure 2.8 shows the estimated jump in the regression of outcome variables using NW, LL and the Hestenes estimators.

Table 2.13 displays the simulation results. In all cases, NW estimators have the largest biases. While the Hestenes estimators have smaller or similar biases to the LL estimators, their RMSE is only half of that of LL estimators. These results agree with the theoretic prediction and are in line with the results of Monte Carlo studies from section 2.3. Thus, in this empirical motivated simulation, my Hestenes estimators outperform the LL estimators in terms of mean squared error.

Through Monte Carlo simulations, both theoretically and empirically motivated, I have verified

**Figure 2.8:** The Estimated Jump in the Empirically Motivated Simulation





**Table 2.13:** Results from the Empirical Motivated Simulation

| Sample Size = 1,787, Repetitions = 2,000                  |                    |       |        |               |        |        |              |       |        |                       |        |       |
|---|--------------------|-------|--------|---------------|--------|--------|--------------|-------|--------|-----------------------|--------|-------|
| Sample Bias and Mean Squared Error of the Jump Estimators |                    |       |        |               |        |        |              |       |        |                       |        |       |
|   | Years of Schooling |       |        | Literacy Rate |        |        | Poverty Rate |       |        | Party Reelection Rate |        |       |
|   | NW                 | LL    | HT11   | NW            | LL     | HT11   | NW           | LL    | HT11   | NW                    | LL     | HT11  |
| Bias  | -0.164             | 0.027 | -0.022 | -0.017        | -0.002 | -0.002 | 0.010        | 0.002 | -0.001 | -0.010                | -0.002 | 0.006 |
| Variance  | 0.029              | 0.273 | 0.059  | 0.001         | 0.006  | 0.001  | 0.001        | 0.005 | 0.001  | 0.002                 | 0.017  | 0.004 |
| RMSE  | 0.237              | 0.524 | 0.243  | 0.033         | 0.080  | 0.039  | 0.027        | 0.074 | 0.035  | 0.047                 | 0.131  | 0.063 |
| Rel_R   | 1.000              | 2.211 | 1.028  | 1.000         | 2.446  | 1.187  | 1.000        | 2.696 | 1.287  | 1.000                 | 2.814  | 1.356 |

that in a number of setting, the MSE of my estimators is smaller than that of commonly used LL estimators. Consequently, there are DGP out there that my estimators outperform the commonly used LL estimators. In such situations, I certainly recommend people to use my estimators because smaller MSE is desirable. Reply on asymptotic results, I have compared my estimators to the properties of LL estimators provided by [Porter, 2003](#). My estimators have theoretically small variance than the LL estimators, so it is possible in many settings my estimators will be preferable to the LL estimators. Methodologically, LL and polynomial estimators reduce bias by taking into account of the shape (derivatives) of the regression when estimating its level, my estimators reduce bias by taking into account of the shape (smoothness) of the regression and completing the partial integrated kernels, so researchers have more control over the tools such as choose higher order smoothness through  $s$ , a suitable sequence of  $w_i$ , or the higher-order bias-reducing kernels.

## 2.6 Conclusions and Future Studies

In this Chapter, I provided a new class of nonparametric estimators for regression discontinuity based on the extension proposed by [Hestenes \(1941\)](#). Compared to the NW estimators, my estimators restore the bias at the boundary points to be the same as that of the interior points. A theoretical comparison between my estimators and the popular local linear approach shows that these two types of estimators have the same unconditional bias order of  $O(h^2)$  and variance order  $O(\frac{1}{nh})$ , but my estimators have smaller variances of the asymptotic distribution at the boundary, which in many cases leads to a smaller MSE. Through Monte Carlo studies, I verify that my estimators perform better than NW estimators in all cases and outperform the LL estimators in terms of smaller MSE in most cases. By applying my estimators to an empirical study by [Litschig and Morrison \(2013\)](#), I show my estimators are easy to use and provide more flexibility than their OLS estimation approach. Empirically motivated simulations reinforce my findings about the finite sample performance of my estimators.

Since the general reflection method I propose in this chapter to construct the jump estimator is new and fundamental to RD research, many topics are left unexplored. One area I have not given a thorough discussion is the bandwidth selection. Not only this relates to an issue that in RDD

I am only interested in the estimation of the jump at the discontinuity point, a locally optimal bandwidth suits my purpose better than a globally optimal bandwidth, but also the choice of  $h$  relates to the choice of other parameters,  $s$ ,  $w_i$  and  $K$ .

What I have defined is a class of estimators. That is, every time, I change the sequence of  $w_i$  and degree of smoothness  $s$ , I change the way I extend the regression and construct the estimator. I have explored two sequences of  $w_i$  in this chapter and found they have a smaller variance of asymptotic distribution than LL. I may find other sequences of  $w_i$  that provide better bias and trade-offs with different DGP and outperform LL estimators or some other currently used estimators. In fact, there are a number of ways to extend smooth functions and the theory I provide here are still valid for all sequences of  $w_i$ . On another notion, when  $s$  gets larger, the trade-off between bias and variance can become disadvantageous for some estimators in the class. From simulation, I can give some practical guidance on how to choose them. For example,  $s$  cannot be too large and  $w = i$  is often better than  $w = 1/i$ , but it is unsatisfying that I do not have a framework to guide me to choose these parameters. It would be helpful to work out a theoretical framework to choose  $s$  and  $w_i$  optimally based on some criteria.

For the foreseeable future, I want to investigate the optimal bandwidth and inference and discuss the choice of  $s$ ,  $w_i$  and  $K$  so I can give a complete road map for how to do estimation of regression discontinuity with my estimators. Other extensions for my estimators include applications in fuzzy RDD, with discreet running variables, or categorical outcome variables. It will be valuable to explore if my estimators perform better than currently used estimators in these settings.

## Chapter 3

### Estimation of a Regression Jump with Additional Covariates in a Nonparametric Additive Model

I consider an additive multivariate regression where one of the component functions is discontinuous, whereas all others are continuous. My goal is to estimate the jump discontinuity of this component function. I assume additivity to overcome the “curse of dimensionality,” which exists in multivariate nonparametric estimation, where the convergence rates are inversely related to the number of regressors  $D$ . [Stone \(1985\)](#) proves that there exist estimators for the component functions of this model that converge at the univariate nonparametric rate, i.e.,  $\sqrt{nh}$  ( $n$  is the sample size and  $h$  is the bandwidth) which is not dependent on  $D$ . An important application of this model is to regression discontinuity designs with additional covariates. The reasons for including additional covariates are two-fold: 1) to guard against misspecification and 2) to reduce the variance of the estimators so that the precision of estimation increases. [Calonico et al., 2018](#) propose an additive separable linear model to match the common practice of empirical RDD studies:

$$Y = m_1(X_1) + Z_2\gamma_2 + \cdots + Z_D\gamma_D + \epsilon, \quad (3.1)$$

where  $E(\epsilon|X, Z) = 0$ . Without loss of generality,  $m_1(x_1)$  is discontinuous at  $x_1 = 0$  and  $x_1 \in \mathbb{R}$ .

Their method does not suffer from the “curse of dimensionality” because only the function of running variable is estimated locally while other parameters are estimated globally. However, this model may be misspecified because covariates come in a linear-in-parameters form. To expand upon their model, I allow for more flexibility by not specifying a functional form of any regression component, either continuous or discontinuous. Thus, I consider

$$Y = m_1(X_1) + m_2(Z_2) + \cdots + m_D(Z_D) + \epsilon. \quad (3.2)$$

If  $m_2(Z_2), \dots, m_D(Z_D)$  were known, I could write (3.2) as

$$Y^* \equiv Y - m_2(Z_2) - \cdots - m_D(Z_D) = m_1(X_1) + \epsilon, \quad (3.3)$$

and consider the estimation of  $m_1$ , using  $Y^*$  as the dependent variable and  $X_1$  as the running variable. This brings us to a jump discontinuity problem of a univariate regression. I can then use my Hestenes-based method developed in Chapter 2 to estimate the jump of  $m_1$ .

Unfortunately, I do not know  $m_2, \dots, m_D$ , which need to be estimated. A collection of estimators have been proposed for continuous additive models: marginal integration (MI), instrument variable (IV), backfitting (BF), and B-splines (SP) (see [Li and Racine, 2007](#), [Martins-Filho and Yang, 2007](#)). My strategy is to use one of these methods to estimate the additive functions  $m_2, \dots, m_D$  at the first stage; then, I use my Hestenes-based method to estimate the jump discontinuity of  $m_1(x_1)$  at the second stage. The challenge is to account for discontinuity at both stages. Furthermore, it is difficult to get asymptotic properties for my Hestenes-based estimators when the regressand is generated: the properties of  $\hat{m}_1(x_1)$  at the second stage depend on the estimates from the first stage,  $\hat{m}_2, \dots, \hat{m}_D$ .

In this chapter, I use several estimators for  $m_1(x_1)$ , and then I conduct a Monte Carlo study to compare and contrast the finite sample performance of these estimators. I set out to answer these questions: 1) For a data generation process that has additive components as part of the regression discontinuity design, does taking into account these extra variables help in estimating the jump? 2) Among available nonparametric estimation alternatives – MI, IV, BF, SP – will there some estimators that outperform the others in finite samples? 3) Suppose I have a data generation process where the additional components of the regression function are nonlinear. Does ignoring the nonlinearity have consequence in the estimation of the jump? That is, will there be differences in estimates between [Calonico et al., 2018](#)'s method using linear additional components and my method using flexible additional components?

### 3.1 Additional covariates in RDD

I am primarily concerned with a regression model that has multiple regressors, in which the regressors  $X$  has a discontinuous regression components and the other regressors  $Z = (Z_2, \dots, Z_D)$  have continuous regression components, as described in 3.2. It does not need any justification to include continuous component functions of  $Z$  when estimating the jump discontinuity in  $m_1(X_1)$  if the additional variables in  $Z$  are in the specification of the conditional mean of the regressand  $Y$ . Their omission leads to a misspecified model.

As a special case, when I cast the problem in a RDD setting, I find that it is very common for researchers to include additional covariates besides the running variable. In fact, there is an on-going debate on what roles additional covariates play in estimating the treatment effects – the jump discontinuity. The question is explored from two perspectives: identification and estimation of the treatment effect. That is, 1) if assumptions of RDD are not met and the treatment effect is not identified, will including additional covariates in regression help the identification of the treatment effect? 2) Will including additional covariates improve the precision of estimation?

Consider the model (3.2). Let  $Y$  be an outcome variable,  $X$  be the running variable that determines if an individual receives the treatment, and  $Z$  be a vector of additional covariates. When  $X \in \mathbb{R}$  is above a threshold  $c \in \mathbb{R}$  and without loss of generality,  $c = 0$ , the individual receives the treatment and  $T = 1$ ; otherwise, the individual does not receive the treatment and  $T = 0$ . That is,  $T = I(X \geq 0)$ , where  $I(A)$  is the indicator function of event  $A$ . Thus,  $Y(T)$  are potential outcomes.

First, I consider identification. In the potential outcome framework, the identification of the treatment effect depends on the conditional independence assumption (CIA) – conditional on  $X$ ,  $Y$  is independent from  $T$ ,  $Y \perp T|X$ . In RDD, I assume that the conditional expectation of potential outcomes  $E(Y(T)|X)$  is continuous at the vicinity of  $c = 0$ , i.e.,  $E(Y(T)|X = -\epsilon) = E(Y(T)|X = \epsilon)$ . Then  $E(Y(T)|T = 0) = E(Y(T)|T = 1)$ , which means  $E(Y(T))$  does not depend on  $T$ . Thus, the continuity of  $E(Y(T)|X)$  implies conditional independence at the vicinity of the threshold. CIA is ensured by the continuity of  $E(Y(T)|X)$ .

Frölich (2007) argues that the model should include additional covariates  $Z$  when the exogeneity of  $X$  is not so credible in some situations unless it is conditional on  $Z$ . However, this argument is debatable. As long as individuals can not completely manipulate the running variable, the conditions for identification— the continuity of the regression functions of the potential outcomes — are still met (Lee and Lemieux (2010)). Furthermore, Calonico et al., 2018 argue that it is unlikely that additional covariates improve the credibility of the identification of the treatment effect. If  $E(Y(T)|X)$  is discontinuous, then  $E(Y(T)|X, Z)$  will not make  $E(Y(T)|X)$  continuous. By the law of iterated expectation, if  $E(Y(T)|X, Z)$  is continuous,  $E(Y(T)|X) = E(E(Y(T)|X, Z)|X)$  is also continuous. Thus, if the regression function of potential outcomes is discontinuous to begin with, additional covariates will not improve identification.

On a side note, Imbens and Lemieux (2008) recommend checking the continuity of conditional expectation  $E(Z|X)$ . Although verifying the continuity of  $E(Z|X)$  says nothing about the continuity of  $E(Y(T)|X)$ , it relieves our concern that we might erroneously attribute the jump of the regression with respect to the running variable,  $\alpha = \lim_{\epsilon \rightarrow 0} E(Y|X = \epsilon) - E(Y|X = -\epsilon)$ , to the average treatment effect while other covariates have discontinuities at the threshold; therefore, they could also contribute to the discontinuity of  $E(Y|X)$ .

Second, I turn to estimation. In what way do additional covariates affect estimation? Using the separable, additive and linear-in-parameter specification, Calonico et al., 2018 show that additional covariates improve precision by reducing the variance of the estimator for the regression discontinuity. In addition, they explore from which channels covariates affect efficiency most and find that additional covariates improve precision the most when the additional covariates  $Z$  correlates with the potential outcome  $Y(T)$ . This is analogous to the notion in OLS that adding a control variable that is uncorrelated with the regressor of interest (here the running variable) but that explains part of the variation in  $Y$  can actually reduce the standard errors by making the residuals smaller. Imbens and Lemieux (2008) argue that including additional covariates can reduce bias when observations away from the threshold are used. It is easy to see that if  $Z$  correlates with  $X$  and  $Y$ , and we do not include  $Z$  in regression, the estimates of the regression on  $X$  are biased. This is similar to the omit variable bias in OLS.

In a word, the literature of RDD, in general, agrees that additional covariates in RDD may not lead to identification of the treatment effect, but they reduce the variance of the jump estimator and, therefore, improve the precision of the estimation.

### 3.2 Condition for Identification of the Model

Consider a D-dimensional nonparametric additive model:

$$E(Y|X_1, \dots, X_D) = m_1(X_1) + m_2(X_2) + \dots + m_D(X_D), \quad (3.4)$$

where  $m_1(x) = \begin{cases} m_1^+(x) & \text{if } x \geq 0 \\ m_1^-(x) & \text{if } x < 0 \end{cases}$  and  $m_2, \dots, m_D$  are sufficiently smooth. In contrast with typical additive models, one of regression component,  $m_1$ , is discontinuous. Therefore, unlike the additive model with all continuous components, the model does not have an intercept term. If I do not impose any condition, the jump is not identified. The model admits infinite representations.

$$E(Y|X_1, \dots, X_D) = \mu_1(X_1) + \mu_2(X_2) + \dots + \mu_D(X_D),$$

where  $\mu_1(X_1) = m_1(X_1) + c$ ,  $\mu_2(X_2) = m_2(X_2) - c$ , and  $\mu_d = m_d$  for  $d = 3, \dots, D$ . To make the jump identifiable, it is necessary to anchor all continuous components of the regression by assuming each of them has a fixed mean. Without lost generality, I assume the simplest form: they all have zero means, i.e.,  $E(m_i(X_i)) = 0$ , for  $i = 2, \dots, D$ . This model is identified because it is not possible to find another set of functions of  $\{\mu_1(X_1), \mu_2(X_2), \dots, \mu_D(X_D) : E(\mu_d(X_d)) = 0, \text{ for } d = 2, \dots, D\}$  that generates the same  $E(Y|X_1 = x_1, \dots, X_D = x_D)$ . Thus, the jump discontinuity at  $x_1 = 0$  is identified,  $\alpha = m_1^+(0) - m_1^-(0)$ .

### 3.3 Estimators Under the Study

I define four different two-stage estimators: in the first stage, I use MI, IV, BF, and SP estimators for component functions of additional covariates; then, in the second stage, I use my Hestenes-based estimators for the regression discontinuity of the running variable. Although the



four first-stage estimators are readily available for continuous regression estimation, the novelty here is to use each in combination with the Hestenes-based estimators to estimate the jump discontinuity.

[Martins-Filho and Yang \(2007\)](#) evaluate the finite sample performance of different estimation methods for additive regression models with continuous components. They examine two types of backfitting estimators, the marginal integration estimator, and two versions of instrumental variable estimators under a common bandwidth selection procedure and conclude that the classic backfitting estimator has the best finite sample performance based on estimators' average squared error. My focus is the estimation of the discontinuous component of the regression rather than the additive model and the first-stage estimators, both of which provide a necessary structure for me to estimate the jump. In terms of choice of estimators for the additive mode, I add spline estimators in the list because of its computational efficiency.

I consider the model:

$$Y = m_1(X_1) + \sum_{d=2}^D m_d(Z_d) + \epsilon$$

$$E(Y|X_1 = x_1, Z_2 = z_2, \dots, Z_D = z_D) = m_1(x_1) + \sum_{d=2}^D m_d(z_d), \quad (3.5)$$

where  $m_1$  is discontinuous and the rest of the components are continuous.  $E(m_d(Z)) = 0$  for  $d = 2, \dots, D$ .

Let  $\{Y_i, X_{i1}, Z_{i2}, \dots, Z_{iD}\}_{i=1}^n$  be a random sample. At the first stage, I estimate  $m_d(Z_{id})$  for  $d = 2, \dots, D$  at  $n$  evaluation points, so I do not lose observations for the second stage. Write  $Y_i^* = Y_i - \hat{m}_2(Z_{i2}) - \dots - \hat{m}_D(Z_{iD})$ . At the second stage, I regress  $Y_i^*$  on  $X_{i1}$  using my Hestenes-based estimators to estimate the jump of discontinuity at  $m_1(0)$ .

### 3.3.1 Marginal Integration Estimation

Marginal integration estimators are proposed by [Tjøstheim and Auestad \(1994\)](#), [Newey \(1994\)](#), and [Linton and Nielsen \(1995\)](#). The idea of Marginal Integration estimation comes from the simple notion that taking conditional expectation with respect to regressor  $Z_d$  on both sides of the model (3.5) does not give the regression function  $m_d(Z_d)$ , because the regressors  $(X_1, Z_2, \dots, Z_D)$  could be correlated. Instead, let  $f_{-d}(x_{-d}) = \int f_X(x_1, \dots, x_D) dx_d$  and note  $\int f_{-d}(x_{-d}) dx_{-d} = 1$ . For

$d = 2, \dots, D,$

$$\begin{aligned}
& \int E(Y|X_1, Z_2, \dots, Z_D) f_{-d}(X_1, Z_{-d}) dX_1 dZ_{-d} \\
&= m_d(Z_d) + \int m_1(X_1) f_1(X_1) dX_1 + \sum_{i=2, i \neq d}^D \int m_i(Z_i) f_i(Z_i) dZ_i \\
&= m_d(Z_d) + E(m_1(X_1)) \quad \text{by assumption} \quad E(m_d(Z)) = 0.
\end{aligned}$$

Thus,  $m_d(Z_d) = \alpha_d(Z_d) - E(m_1(x_1))$ , where

$$\alpha_d(Z_d) = \int E(Y|X_1, Z_2, \dots, Z_D) f_{-d}(X_1, Z_{-d}) dX_1 dZ_{-d}.$$

Since  $E(m_1(X_1)) = E(Y)$ , I estimate  $E(m_1(X_1))$  by  $\hat{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ . To estimate  $\alpha_d(Z_d)$ , I split the sample by  $X_1 \geq 0$  or  $X_1 < 0$  because of the discontinuity in  $m_1$ . I use the multivariate Nadaraya-Watson estimator to estimate  $E(Y|X_1, Z_2, \dots, Z_D)$  at  $(X_{i1}, z_d, Z_{i,-d})$  for  $i = 1, \dots, n$ . Note that among these evaluation points, only one of them,  $(X_{i1}, Z_{id}, Z_{i,-d})$  where  $Z_{id} = z_d$ , is observed; all others are not observed. Since I do not know the marginal density  $f_{-d}(X_1, Z_{-d})$ , to estimate integration, I use empirical density  $1/n^1$ . After performing  $n$  multivariate regression estimations on  $m(X_{i1}, \dots, z_d, \dots, Z_{iD})$  for  $i = 1, \dots, n$ , and taking an average, I get  $\hat{\alpha}(z_d) = \frac{1}{n} \sum_{i=1}^n m(X_{i1}, \dots, z_d, \dots, Z_{iD})$ . I then repeat the process  $n$  times to get  $\hat{\alpha}(z_{jd})$  for  $j = 1, \dots, n$ . Lastly, I calculate  $\hat{m}_d(Z_{jd}) = \hat{\alpha}(Z_{jd}) - \frac{1}{n} \sum_{j=1}^n \hat{\alpha}(Z_{jd})$  to ensure  $\hat{m}_d(Z_{jd}) = 0$ .

The MI estimators are oracle efficient and have desirable asymptotic properties, but the method is computationally intensive due to a large number of multivariate regression estimations. This can be seen from a simple calculation. For a sample of size  $n$  with  $D$  regressors, to evaluate each regression function at  $n$  points, the number of multivariate regressions is  $n^2$ . For  $D$  regression functions, the number of multivariate regressions is  $D * n^2$ . In a special case, when  $D = 2$ , evaluation points for the two regressions are the sample points, so I can save calculations by using each bivariate regression twice, and the number of multivariate regressions is  $n^2$  rather than

---

<sup>1</sup>If I want this joint density to be estimated, I need to know how to estimate a multivariate density with one dimension, such as  $X_1$ , bounded. Here I rely on the assumption of RDD that all regressors, including  $X_1$ , have continuous densities.

$2 * n^2$ . From the estimation procedure, it is clear that additivity does not come into play when the estimation is performed; it only plays a role later when obtaining the regression component functions.

### 3.3.2 Instrument Variable Estimation

Marginal integration estimators are computationally expensive. [Kim et al. \(1999\)](#) propose a two-stage estimator that is oracle efficient and reduces the computation to the order of  $n$  univariate estimations. As I mentioned in the last subsection, conditional expectation with respect to regressor  $X_d$  in general does not give us the regression component  $m_d(X_d)$ , because regressors are correlated. Only when regressors are uncorrelated, the conditional expectation and the regression component function are equal. For example, if  $X_1$  and  $X_2$  are independent,  $f_{1,2}(x_1, x_2) = f_1(x_1)f_2(x_2)$ , then  $m_1(x_1) = E(y|x_1)$ . Therefore, the idea of IV estimation is to remove the correlation among regressors by using instrument variables.

Consider the following D-dimensional additive model:

$$Y = m_1(X_1) + m_2(Z_2) + \dots + m_D(Z_D) + u \quad (3.6)$$

where  $E(u|X_1, Z_2, \dots, Z_D) = 0$ ,  $m_X$  is discontinuous at  $x = 0$  and  $m_d$  is continuous for  $d = 2, \dots, D$ . The marginal density functions of all regressors are continuous, and  $E(m_d(Z)) = 0$ .

Let  $f_Z(Z) = f_Z(z_1, \dots, z_D)$ ,  $f_{XZ}(X, Z) = f_{XZ}(x_1, z_2, \dots, z_D)$ , and

$$f_{XZ_{-d}}(XZ_{-d}) = \int f_{XZ}(x_1, z_1, \dots, z_D) dz_d.$$

Define  $w_X(X, Z) \equiv \frac{f_X(X)f_Z(Z)}{f_{XZ}(X, Z)}$  and  $w_d(X, Z) \equiv \frac{f_d(Z_d)f_{XZ_{-d}}(XZ_{-d})}{f_{XZ}(X, Z)}$ . Using Rosenblatt density estimators, I can estimate  $w_X(X, Z)$  by  $\hat{w}_X(X, Z) = \frac{\hat{f}_X(X)\hat{f}_Z(Z)}{\hat{f}_{XZ}(X, Z)}$  and  $w_d(X, Z)$  by  $\hat{w}_d(X, Z) = \frac{\hat{f}_d(Z_d)\hat{f}_{XZ_{-d}}(XZ_{-d})}{\hat{f}_{XZ}(X, Z)}$ .

Multiply both sides of (3.6) by  $w_X$ :

$$w_X Y = w_X m_1(X_1) + \sum_{d=2}^D w_X m_d(Z_d) + u.$$

and take conditional expectation on both sides,

$$E(w_X Y | X = x) = E(w_X m_1(X) | X = x) + \sum_{d=2}^D E(w_X m_d(Z_d) | X = x).$$

The first term,

$$\begin{aligned} E(w_X m_1(X) | X = x) &= E\left(\frac{f_X(X) f_Z(Z)}{f_{XZ}(X, Z)} m_1(X) | X = x\right) \\ &= f_X(x) m_1(x) E\left(\frac{f_Z(Z)}{f_{XZ}(x, Z)} | X = x\right) \\ &= f_X(x) m_1(x) \int \left(\frac{f_Z(Z)}{f_{XZ}(x, Z)} \frac{f_{XZ}(x, Z)}{f_X(x)} dZ\right) \\ &= m_1(x), \end{aligned}$$

and the second term, for all  $d = 2, \dots, D$ ,

$$\begin{aligned} E(w_X m_d(Z_d) | X = x) &= E\left(\frac{f_X(X) f_Z(Z)}{f_{XZ}(X, Z)} m_d(Z_d) | X = x\right) \\ &= f_X(x) E\left(\frac{f_Z(Z) m_d(Z_d)}{f_{XZ}(x, Z)} | X = x\right) \\ &= f_X(x) \int \left(\frac{f_Z(Z) m_d(Z_d)}{f_{XZ}(x, Z)} \frac{f_{XZ}(x, Z)}{f_X(x)} dZ\right) \\ &= E(m_d(Z_d)) = 0. \end{aligned}$$

Thus,  $E(w_X Y | X = x) = m_1(x)$ . Since  $m_1(x)$  is discontinuous at  $x = 0$ , I use the Hestenes estimator to estimate  $E(w_X Y | X = x)$ . That is  $\hat{m}_1^p(x) = \hat{E}(w_X Y | X = x)$ .

Now, multiply both sides of (3.6) by  $w_d$ :

$$w_d Y = w_d m_1(X_1) + \sum_{d=2}^D w_d m_d(Z_d) + u,$$

and take conditional expectation,

$$E(w_d Y | Z_d = z_d) = E(w_d m_1(X_1) | Z_d = z_d) + \sum_{d=2}^D E(w_d m_d(Z_d) | Z_d = z_d).$$

The first term,

$$\begin{aligned}
E(w_d m_1(X)|Z_d = z_d) &= E\left(\frac{f_d(Z_d)f_{XZ-d}(XZ-d)}{f_{XZ}(X,Z)}m_1(X)|Z = z_d\right) \\
&= f_d(z_d)E\left(\frac{f_{XZ-d}(XZ-d)}{f_{XZ}(X,Z)}m_1(X)|Z = z_d\right) \\
&= f_d(z_d)\int\left(\frac{f_{XZ-d}(XZ-d)}{f_{XZ}(X,Z)}m_1(X)\frac{f_{XZ}(X,Z)}{f_d(z_d)}dXdZ-d\right) \\
&= E(m_1(X)),
\end{aligned}$$

and the second term,

$$\begin{aligned}
\sum_{i=2}^D E(w_d m_i(Z_i)|Z_d = z_d) &= E(w_d m_d(Z_d)|Z_d = z_d) \\
&= E\left(\frac{f_d(Z_d)f_{XZ-d}(XZ-d)}{f_{XZ}(X,Z)}m_d(Z_d)|Z_d = z_d\right) \\
&= f_d(z_d)m_d(z_d)E\left(\frac{f_{XZ-d}(XZ-d)}{f_{XZ}(X,Z)}|Z_d = z_d\right) \\
&= f_d(z_d)m_d(z_d)\int\left(\frac{f_{XZ-d}(XZ-d)}{f_{XZ}(X,Z)}\frac{f_{XZ}(X,Z)}{f_d(z_d)}dXdZ-d\right) \\
&= m_d(z_d).
\end{aligned}$$

Thus,  $E(w_d Y|Z_d = z_d) = E(m_1(X)) + m_d(z_d)$ . I estimate  $E(m_1(X))$  by  $\hat{E}(m_1(X)) = \hat{E}(Y) = \frac{1}{n}\sum_{i=1}^n Y_i$ , and I estimate  $E(w_d Y|Z_d = z_d)$  using the NW estimator. I obtain  $\hat{m}_d^p(z_d) = \hat{E}(w_d Y|Z_d = z_d) - \hat{E}(m_1(X))$ .

Following the explanation by [Kim et al. \(1999\)](#), I call  $\hat{m}_1^p(x)$  and  $\hat{m}_d^p(z_d)$  pilot estimators because they are not oracle efficient. To get oracle efficient estimators, I perform a one-step backfitting estimation. I define regression as

$$Y - \sum_{d=2}^D \hat{m}_d^p(Z_d) = m_1(X_1) + \gamma,$$

where  $E(\gamma|X_1, Z_2, \dots, Z_D) = 0$ . Then  $E(Y - \sum_{d=2}^D \hat{m}_d^p(Z_d)|X_1 = x) = m_1(x)$ . Using the Hestenes estimator to estimate  $m_1(x)$ , I obtain  $\hat{m}_1(x) = \hat{E}(Y - \sum_{d=2}^D \hat{m}_d^p(Z_d)|X_1 = x)$ .

Then I define regression as

$$Y - \hat{m}_1(X_1) - \sum_{i \neq d}^D \hat{m}_i^p(Z_i) = m_d(Z_d) + \lambda,$$

where  $E(\lambda|X_1, Z_2, \dots, Z_D) = 0$ .  $E(Y - \hat{m}_1(X)|Z_d = z_d) = m_d(z_d)$ . Using the NW estimator to estimate  $m_d(z_d)$ , I obtain  $\hat{m}_d(z_d) = \hat{E}(Y - \hat{m}_1(X)|Z_d = z_d)$ .

In summary, using instrument variables, I obtain the pilot estimators. Then, with the one-step backfitting estimation, I obtain the oracle efficient estimators for the continuous components of additional covariates. Generally speaking, IV estimation is computationally faster than MI estimation. For a sample of size  $n$  with  $D$  regressors, it performs  $2D * n$  univariate estimations rather than  $D * n^2$  multivariate estimations as in MI .

### 3.3.3 Backfitting Estimation

A widely used backfitting procedure for additive models was initially proposed by [Buja et al. \(1989\)](#). Through iterations and updating, backfitting estimation is another method to use univariate regressions instead of multivariate regressions to estimate an additive model. However, as shown by [Opsomer et al. \(1997\)](#) and [Opsomer \(2000\)](#) on their asymptotic properties, the backfitting estimators are not oracle efficient.

I use the following iterated procedure to estimate  $m_d(Z_d)$  for  $d = 2, \dots, D$ . I start by defining

$$Y_{i1} = Y_i - \sum_{d=2}^D m_d(Z_{id}) = Y_i = m_1(X_{i1}) + u_{i1}.$$

Since  $m_1$  is discontinuous, I split the sample by  $X_1 \geq 0$  or  $X_1 < 0$ , and use the Hestenes estimator to obtain  $\hat{m}_1^{[1]}(X_{i1})$ , for  $i = 1, \dots, n$ . I then define

$$Y_{i2} = Y_i - \hat{m}_1^{[1]}(X_{i1}) - \sum_{d=3}^D m_d(Z_d) = m_2(Z_{i2}) + u_{i2}.$$

Since  $m_2$  is continuous, I use the NW estimator to obtain  $\hat{m}_2^{[1]}(Z_{i2})$ , for  $i = 1, \dots, n$ . Next I define

$$Y_{i3} = Y_i - \hat{m}_1^{[1]}(X_{i1}) - \hat{m}_2^{[1]}(Z_{i2}) - \sum_{d=4}^D m_d(Z_d) = m_3(Z_{i3}) + u_{i3},$$

and use the NW estimator to obtain  $\hat{m}_3^{[1]}(Z_{i3})$ , for  $i = 1, \dots, n$ . I repeat these steps until

$$Y_{iD} = Y_i - \hat{m}_1^{[1]}(X_{i1}) - \sum_{d=2}^{D-1} \hat{m}_d^{[1]}(Z_{id}) = m_D(Z_{iD}) + u_{iD}.$$

Then, I use the NW estimator to obtain  $\hat{m}_D^{[1]}(Z_{iD})$ , for  $i = 1, \dots, n$ . I call  $\hat{m}_1^{[1]}(X_{i1}), \hat{m}_2^{[1]}(Z_{i2}), \dots, \hat{m}_D^{[1]}(Z_{iD})$  the one step backfitting estimators for  $m_1, \dots, m_D$ . Now, I choose  $\epsilon > 0$  and evaluate whether or not

$$\left| \sum_{i=1}^n Y_i^2 - \sum_{i=1}^n \left( Y_i - \hat{m}_1^{[1]}(X_{i1}) - \sum_{d=2}^D \hat{m}_d^{[1]}(Z_{id}) \right)^2 \right| < \epsilon.$$

If yes, I stop. Otherwise, I define,

$$Y_{i1}^{[1]} = Y_i - \sum_{d=2}^D \hat{m}_d^{[1]}(Z_{id}) = m_1(X_{i1}) + u_{i1}^{[1]},$$

and use the Hestenes estimator to obtain  $\hat{m}_1^{[2]}(X_{i1})$ , for  $i = 1, \dots, n$ . Next, I define

$$Y_{i2}^{[1]} = Y_i - \hat{m}_1^{[2]}(X_{i1}) - \sum_{d=3}^D \hat{m}_d^{[1]}(Z_{id}) = m_2(Z_{i2}) + u_{i2}^{[1]},$$

and use the NW estimator to obtain  $\hat{m}_2^{[2]}(Z_{i2})$ , for  $i = 1, \dots, n$ . I repeat these steps until I obtain  $\hat{m}_D^{[2]}(Z_{iD})$ . I call  $\hat{m}_1^{[2]}(X_{i1}), \hat{m}_2^{[2]}(Z_{i2}), \dots, \hat{m}_D^{[2]}(Z_{iD})$  the second step backfitting estimators for  $m_1, \dots, m_D$ . I evaluate whether or not

$$\left| \sum_{i=1}^n \left( Y_i - \hat{m}_1^{[1]}(X_{i1}) - \sum_{d=2}^D \hat{m}_d^{[1]}(Z_{id}) \right)^2 - \sum_{i=1}^n \left( Y_i - \hat{m}_1^{[2]}(X_{i1}) - \sum_{d=2}^D \hat{m}_d^{[2]}(Z_{id}) \right)^2 \right| < \epsilon.$$

If yes, I stop. Otherwise I repeat the backfitting steps until the step  $J$  such that

$$\left| \sum_{i=1}^n \left( Y_i - \hat{m}_1^{[J-1]}(X_{i1}) - \sum_{d=2}^D \hat{m}_d^{[J-1]}(Z_{id}) \right)^2 - \sum_{i=1}^n \left( Y_i - \hat{m}_1^{[J]}(X_{i1}) - \sum_{d=2}^D \hat{m}_d^{[J]}(Z_{id}) \right)^2 \right| < \epsilon.$$

I call  $\hat{m}_1^{[J]}(X_{i1}), \hat{m}_2^{[J]}(Z_{i2}), \dots, \hat{m}_D^{[J]}(Z_{iD})$  the  $J$  step backfitting estimators for  $m_1, \dots, m_D$ . A process with  $J$  iterations performs  $J * D * n$  univariate regression estimations.

### 3.3.4 B-Splines Estimation

As nonparametric estimation, spline estimators compute significantly faster than kernel estimators (Chen, 2007; Wang et al., 2007). Similar to global polynomial estimation, in spline estimation, coefficients of the basis functions come into the model in linear form. It only takes one ordinary least squared (OLS) estimation to estimate all coefficients. In contrast, kernel estimators perform  $n$  local estimations at  $n$  evaluation points. The advantage becomes even greater when a large number of regressors are present in the model – still, spline estimators only need one OLS estimation to obtain all coefficients. Of course, as the dimension of regressors and the sample size increases, the number of coefficients increases. The computational burden is to compute the inverse of a matrix of the size of the number of coefficients.

Consider an interval  $[a, b] \in \mathbb{R}$ , on which I define a knot set  $\vec{t} = \{t_1, \dots, t_N, t_{N+1}, \dots, t_{N+k}\}$  with  $t_1 = t_2 = \dots = t_k = a$ ,  $t_{N+1} = t_{N+2} = \dots = t_{N+k} = b$ , and  $t_i < t_{i+1}$  for all  $i = k, \dots, N$ . Thus I have  $N + k$  knots, which consists of  $N - k$  interior knots and  $2k$  boundary knots. I choose  $N$  according to the sample size  $n$  :  $N = \sqrt{n}$ .  $k$  is the order of the polynomial of basis functions. For example, if  $k = 1$ , a the basis function is a constant function,  $k = 2$ , a linear function,  $k = 3$ , a quadratic function, and so on.

Basis functions,  $B_{j,l}(x)$ , are defined recursively. At the first level,  $l = 1$ ,

$$B_{k+j-1,1}(x) = \begin{cases} 1, & \text{if } t_{k+j-1} \leq x \leq t_{k+j}, j = 1, \dots, N - k + 1 \\ 0, & \text{otherwise} \end{cases},$$

if  $t_j = t_{j+1}$ ,  $B_{j,l} = 0$ . That is,  $B_{1,1} = \dots = B_{k-1,1} = 0$ , and  $B_{N+1,1} = \dots = B_{N+k-1,1} = 0$ .

At the levels up to the  $k$ th level,  $1 < l \leq k$ ,

$$B_{j,l} = w_{j,l} B_{j,l-1} + (1 - w_{j+1,l}) B_{j+1,l-1},$$



where  $w_{j,l} = \frac{x-t_j}{t_{j+l-1}-t_j}$ , for  $j = 1, \dots, N + k - l$ .

Now, I define a B-Splines of order  $k$  as

$$S^k(x) = \sum_{j=1}^N \lambda_j B_{j,k}(x),$$

where  $\lambda_j \in \mathbb{R}$ , and  $B_{j,k}(x)$  are basis functions. The space of the spline function is of  $N$  dimensions,  $N = k + N - k$ , where  $k$  is the order of the polynomial of basis functions and  $N - k$  is the number of interior knots.

Using the B-Splines function to approximate the components of regression function in equation (3.5), I have

$$\begin{aligned} E(Y|X_1, Z_2, \dots, Z_D) &= S_1^k(X_1) + \sum_{d=2}^D S_d^k(Z_d) + N^{-1}O(1) \\ &= \sum_{j=1}^N \lambda_{j1} B_{j1,k}(X_1) + \sum_{d=2}^D \sum_{j=1}^N \lambda_{jd} B_{jd,k}(Z_d) + N^{-1}O(1). \end{aligned}$$

Let  $B$  be the  $(N * D) \times n$  matrix of the basis functions,

$$B = \begin{bmatrix} B_{11,k}(X_{11}) & \dots & B_{1N,k}(X_{11}) & B_{12,k}(Z_{21}) & \dots & B_{ND,k}(Z_{D1}) \\ B_{11,k}(X_{12}) & \dots & B_{1N,k}(X_{12}) & B_{12,k}(Z_{22}) & \dots & B_{ND,k}(Z_{D2}) \\ \vdots & & & & & \vdots \\ B_{11,k}(X_{1n}) & \dots & B_{1N,k}(X_{1n}) & B_{12,k}(Z_{2n}) & \dots & B_{ND,k}(Z_{Dn}) \end{bmatrix}$$

and  $\lambda$  be the  $(N * D) \times 1$  matrix of the coefficients,

$$\lambda = \begin{bmatrix} \lambda_{11} & \dots & \lambda_{N1} & \lambda_{12} & \dots & \lambda_{N2} & \dots & \lambda_{1D} & \dots & \lambda_{ND} \end{bmatrix}'$$

I can write  $E(Y|X_1, Z_2, \dots, Z_D) = B\lambda + N^{-1}O(1)$ . Then, using OLS estimation, I get the estimates of the coefficients,  $\hat{\lambda} = (B' B)^{-1} B' Y$ . Lastly, I calculate

$$\hat{m}_d(Z_{jd}) = \sum_{j=1}^G \hat{\lambda}_{jd} B_{jd,k}(Z_{id}) - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^N \hat{\lambda}_{jd} B_{jd,k}(Z_{id})$$

for  $d = 2, \dots, D$  to ensure  $\hat{m}_d(Z_{jd}) = 0$ .

An issue with spline estimators is the singularity problem. Basis functions,  $B_{j,k}(x)$ , are piecewise functions whose values are non-zeros only in  $k$  intervals of the knot set, so most elements in the matrix  $B$  are zeros. This causes the matrix to be close to singularity and generates inaccurate estimates of the coefficients  $\hat{\lambda}_{jd}$ , of the spline function. The problem becomes more severe when the sample size  $n$  increases since the number of knots  $N$  increases with the sample size  $n$  and the size of interval,  $\frac{b-a}{N-1}$ , becomes smaller.

### 3.4 Simulation

To investigate the finite sample performance of covariates-adjusted RD estimators, I conduct simulations to compare the four two-stage RD estimators in which I use MI, IV, BF, or SP estimators respectively as the first-stage estimators, and the Hestenes estimators as the second-stage estimators. I compare the performance of these estimators and with the four benchmark estimators that represent four different cases. In the first and second cases, I use the univariate Hestenes estimator and local linear estimators to estimate the jump without accounting for the additional covariates in the regression. The third case is the estimator proposed by [Calonico et al., 2018](#), where they use a local linear estimator to estimate an additive linear model. I describe the implementation in the next subsection. The last case considers oracle estimators in which the true regression components of the additional covariates are known. The descriptions of these estimator are as follows.

HT: the univariate Hestenes estimator that does not account for the additional covariates in the regression.

LL: the univariate local linear estimator that does not account for the additional covariates in the regression.

Cal: the estimator described in [Calonico et al., 2018](#) that uses additive linear model to account for the additional covariates in the regression.

Oracle: the oracle estimator in which the true regression components of the additional covariates are known,

MI-HT: the two-stage estimator that uses the marginal integration estimator at the first stage and the Hestenes estimator at the second stage.

IV-HT: the two-stage estimator that uses the instrument variable estimator at the first stage and the Hestenes estimator at the second stage.

BF-HT: the two-stage estimator that uses the backfitting estimator at the first stage and the Hestenes estimator at the second stage.

SP-HT: the two-stage estimator that uses the B-Splines estimator at the first stage and the Hestenes estimator at the second stage.

### 3.4.1 RD Estimators Using Additive Linear Model

I implement the jump estimator  $\hat{\alpha}$  described in [Calonico et al., 2018](#), which is a local linear estimator with additional covariates in linear form. For ease of notation, consider a bivariate regression model:

$$Y = m(X) + Z\gamma + \epsilon.$$

The estimator is obtained from

$$\hat{\alpha} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix} \hat{\beta}' - \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \hat{\beta}'.$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_-^1 \\ \hat{\beta}_-^2 \\ \hat{\beta}_+^1 \\ \hat{\beta}_+^2 \\ \hat{\gamma} \end{bmatrix} = \arg \min_{\beta_-^1, \beta_-^2, \beta_+^1, \beta_+^2, \gamma} \sum_{i=1}^n \left[ Y_i - I(X_i - x < 0) [1 \ X_i - x] \begin{bmatrix} \beta_-^1 \\ \beta_-^2 \end{bmatrix} - I(X_i - x \geq 0) [1 \ X_i - x] \begin{bmatrix} \beta_+^1 \\ \beta_+^2 \end{bmatrix} - Z_i \gamma \right]^2 K\left(\frac{X_i - x}{h}\right),$$

where  $I(A)$  is the indicator function of event  $A$ ,  $x$  is the thresh hold, and  $K(\cdot)$  is a kernel function.

Write  $R(x) = \begin{bmatrix} I(X_i - x < 0) [1 \ X_i - x] & I(X_i - x \geq 0) [1 \ X_i - x] & Z_i \end{bmatrix}_{i=1}^n$ , and  $P(x) = \text{diag}\{K \left(\frac{X_i - x}{h}\right)\}_{i=1}^n$ , then  $\hat{\beta} = (R(x)'P(x)R(x))^{-1}R(x)Y$ .

### 3.4.2 The Data Generating Process (DGP)

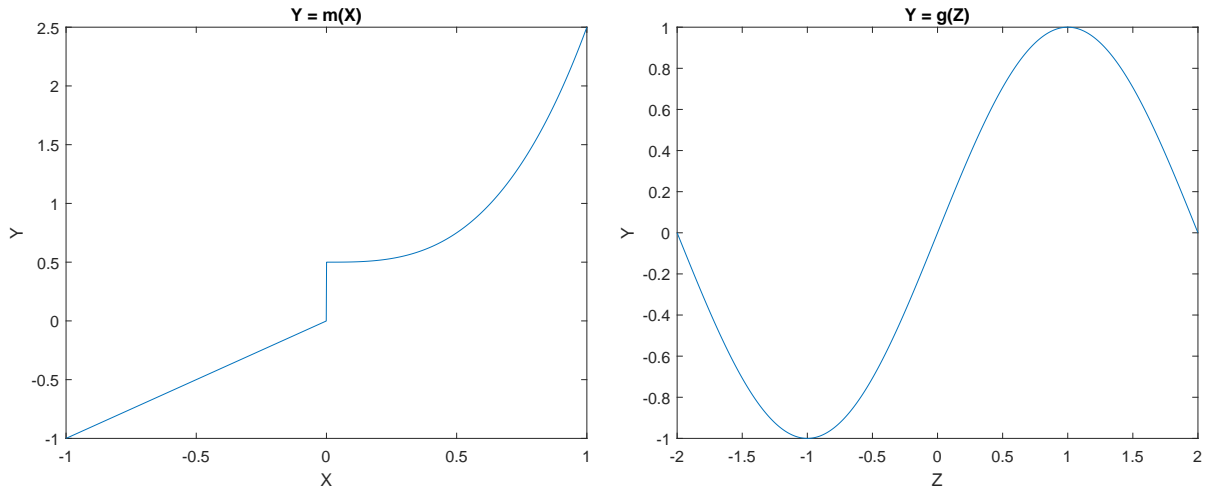
I specify the following bivariate additive model,

$$Y = m(X) + g(Z) + \epsilon,$$

where  $m(X) = \begin{cases} X & \text{if } -1 \leq X < 0 \\ J + kX^p & \text{if } 0 \leq X \leq 1 \end{cases}$ ,  $g(Z) = \sin(\pi Z/2)$ , and  $\epsilon \sim N(0, \sigma^2)$ .

The spline estimator is defined on samples with compact support, so I generate independent variables  $X$  and  $Z$  with uniform distributions:  $X \sim U[-1, 1]$  and  $Z \sim U[-2, 2]$ . Figure 3.1 provides the graph of the model over the relevant range of  $X$  and  $Z$ .

**Figure 3.1:** Regression Components of the True Model



$E(g(x)) = \int_{-2}^2 g(Z) f_Z(Z) dZ = \int_{-2}^2 \sin(\pi Z/2) \frac{1}{4} dZ = 0$ , and  $E(m(X)) = \int_{-1}^1 m(X) f_X(X) dX = -\frac{1}{4} + \frac{J}{2} + \frac{1}{2} \frac{k}{p+1}$ . If  $k = 2$ ,  $p = 3$ , and  $J = 0$ , then  $E(m(X)) = \frac{J}{2}$ .

To examine other sample distributions, I also generate  $X$  and  $Z$  with truncated normal distributions:  $X \sim N(0, 1)$  between  $(-1, 1)$  and  $Z \sim N(0, 2)$  between  $(-2, 2)$ . In both cases, I generate samples of size  $n = 200, 500, 1000$  respectively and conduct the simulation with 2000 repetitions.

### 3.4.3 Estimation Results

The findings of the finite performance of these estimators are in Table 3.1. To compare performance among nonparametric estimators, it is essential to choose a bandwidth that is fair to all estimators rather than favor some estimators. The common bandwidth selection methods include the plug-in, cross-validation, and the rule of thumb method. Optimally choosing the bandwidth is complicated for two-stage estimators of the additive model with discontinuity because it means to choose the bandwidths for each regressor at the first stage and for the regressor with the discontinuous component at the second stage to minimize the MSE at the discontinuous point. Since I do not have asymptotic properties of these estimators, I am not able to use plug-in methods. For simplicity and being fair to all estimators, I use the rule of thumb proposed by [Kim et al. \(1999\)](#) to choose bandwidth for all estimators.  $h_1 = n^{-1/5} \text{std}(X_1)$  and  $h_d = n^{-1/5} \text{std}(Z_d)$ , for  $d = 2, \dots, D$ . The bandwidths depend only on the sample size and variance, so all estimators have the same bandwidth.

For all sample sizes and distributions in my experiments, I observe these results: the Oracle, IV-HT, and BF-HT estimators have the smallest root mean squared error (RMSE), followed by the HT, MI-HT, and SP-HT estimators. The ones that have the largest RMSE are LL and Cal estimators. These results agree with my predictions. The Oracle estimators have the small RMSE because the model is correctly specified and the additional components are known. The four two-stage estimators use nonparametric functions to account for the additional components that are nonlinear, so they have smaller MSE than the estimators by [Calonico et al., 2018](#) who use linear functions to account for the additional components. The univariate Hestenes estimators have smaller RMSE than the univariate LL estimators, which is in accordance with my findings in Chapter 2.

It might be a little unexpected that the best performer is the IV-HT estimator rather than the Oracle estimator. This could be because the IV-HT estimator is an oracle efficient estimator. As  $n \rightarrow \infty$ , it has the same asymptotic distribution as the Oracle estimator whereas in finite samples, it could outperform the Oracle estimator. It is also evident in the experiment that as  $n$  increases, the RMSE of the two estimators become ever closer to each other. Another surprise is

**Table 3.1:** Simulation Results

| Sample Bias and Mean Squared Error of the Jump Estimators                                      |          |       |        |        |       |       |       |       |       |
|--|----------|-------|--------|--------|-------|-------|-------|-------|-------|
| Estimators   | HT       | LL    | Cal    | Oracle | MI-HT | IV-HT | BF-HT | SP-HT |       |
| Uniform distributions: $X \sim U[-1, 1]$ and $Z \sim U[-2, 2]$                                 |          |       |        |        |       |       |       |       |       |
| n=200  | Bias     | 0.678 | -0.366 | -0.366 | 0.679 | 0.681 | 0.671 | 0.667 | 0.685 |
|  | Variance | 0.381 | 1.375  | 1.361  | 0.374 | 0.408 | 0.372 | 0.378 | 0.427 |
|  | RMSE     | 0.917 | 1.228  | 1.222  | 0.914 | 0.933 | 0.906 | 0.907 | 0.947 |
|  | Rel_R    | 1.011 | 1.355  | 1.348  | 1.008 | 1.030 | 1.000 | 1.001 | 1.045 |
| n=500  | Bias     | 0.580 | -0.280 | -0.282 | 0.580 | 0.577 | 0.577 | 0.575 | 0.577 |
|  | Variance | 0.162 | 0.543  | 0.540  | 0.158 | 0.169 | 0.158 | 0.161 | 0.174 |
|  | RMSE     | 0.706 | 0.789  | 0.787  | 0.703 | 0.708 | 0.700 | 0.701 | 0.712 |
|  | Rel_R    | 1.008 | 1.126  | 1.124  | 1.004 | 1.011 | 1.000 | 1.001 | 1.016 |
| n=1000   | Bias     | 0.466 | -0.256 | -0.257 | 0.466 | 0.467 | 0.464 | 0.463 | 0.468 |
|  | Variance | 0.092 | 0.283  | 0.276  | 0.089 | 0.099 | 0.089 | 0.091 | 0.097 |
|  | RMSE     | 0.555 | 0.590  | 0.585  | 0.553 | 0.563 | 0.551 | 0.553 | 0.562 |
|  | Rel_R    | 1.007 | 1.071  | 1.061  | 1.004 | 1.021 | 1.000 | 1.003 | 1.020 |
| Truncated normal : $X \sim N(0, 1)$ between $(-1, 1)$ and $Z \sim N(0, 2)$ between $(-2, 2)$ . |          |       |        |        |       |       |       |       |       |
| n=200  | Bias     | 0.624 | -0.283 | -0.282 | 0.620 | 0.624 | 0.617 | 0.617 | 0.622 |
|  | Variance | 0.373 | 1.277  | 1.228  | 0.354 | 0.390 | 0.354 | 0.370 | 0.418 |
|  | RMSE     | 0.873 | 1.164  | 1.143  | 0.860 | 0.883 | 0.857 | 0.866 | 0.897 |
|  | Rel_R    | 1.019 | 1.359  | 1.334  | 1.003 | 1.030 | 1.000 | 1.011 | 1.047 |
| n=500  | Bias     | 0.515 | -0.262 | -0.261 | 0.513 | 0.512 | 0.510 | 0.511 | 0.514 |
|  | Variance | 0.161 | 0.533  | 0.524  | 0.158 | 0.168 | 0.157 | 0.160 | 0.172 |
|  | RMSE     | 0.653 | 0.775  | 0.769  | 0.649 | 0.656 | 0.646 | 0.649 | 0.660 |
|  | Rel_R    | 1.010 | 1.199  | 1.190  | 1.004 | 1.015 | 1.000 | 1.004 | 1.022 |
| n=1000   | Bias     | 0.428 | -0.222 | -0.222 | 0.428 | 0.426 | 0.426 | 0.426 | 0.425 |
|  | Variance | 0.091 | 0.269  | 0.262  | 0.088 | 0.094 | 0.088 | 0.091 | 0.097 |
|  | RMSE     | 0.523 | 0.564  | 0.557  | 0.521 | 0.526 | 0.519 | 0.521 | 0.527 |
|  | Rel_R    | 1.007 | 1.086  | 1.073  | 1.003 | 1.012 | 1.000 | 1.004 | 1.014 |

that the univariate Hestenes estimators that do not account for the additional components have smaller RMSE than the estimator by [Calonico et al., 2018](#) that account for the additional covariates. The explanation could be that the presence of additional variables make the regression component function with discontinuity less linear.

The reduction of RMSE by estimators that accounting for additional covariates is achieved through the reduction of variance. If I compare the LL estimator with the Cal estimator, and the HT estimator with the four two-stage estimators: MI-HT, IV-HT, BF-HT, and SP-HT, I can see that adding covariates has little or no effects on bias but significantly reduce the variance so the RMSE decreases.

The differences in performance among the four two-stage estimators are not dramatic. The estimators using instrument variable and backfitting as the first-stage estimators have the smallest RMSE, which is in line with the findings from [Martins-Filho and Yang \(2007\)](#), who conclude the backfitting estimators have the best finite performance among other estimators, such as MI and IV estimators, for additive models with continuous regression components. As the sample size increases, the RMSE of every estimator has reduced significantly. Noticeably, at small sample size  $n = 200$ , the mean squared error of each of the four two-stage estimators is about half of the mean squared error of the estimator by [Calonico et al., 2018](#).

The results suggest that in estimating a RDD model when additional covariates are present, researchers should first consider the IV-HT estimator because they have the best performance among the four two-stage estimators and outperform the commonly used local linear estimation with additional covariates in linear form. The second best choice is my Hestenes univariate regression estimators. They have similar RMSE to my two-stage estimators and smaller RMSE than both univariate and multivariate local linear regression estimators.

### 3.5 Conclusions

In regression discontinuity designs, it is very common for empirical researchers to include additional covariates when estimating the jump discontinuity. In this chapter, using an additive model and allowing flexible functional form for all components of covariates, I explore issues regarding the

estimation of the jump discontinuity in the presence of additional covariates. I find that accounting for additional covariates reduces the variance of estimates, and therefore, improves the precision of the estimation. Choosing from available nonparametric estimation alternatives for additive models, I find that instrument variable and backfitting estimators have the smallest RMSE in estimating the jump discontinuity. If the additional components of the regression function are nonlinear, ignoring the non-linearity results in larger RMSE. My findings recommend that when additional covariates are available, it is advantageous for empirical researchers to use my two-stage estimators to estimate the jump discontinuity. The second best choice is my Hestenes univariate regression estimator.

For future work, I plan to research asymptotic properties of the two-stage estimators that use B-splines as first-stage estimators for continuous components of the regression and use Hestenes estimators as the second-stage estimators for the discontinuous component. Spline estimators are computationally expedient, especially when a large number of regressors are present. However, in general, it is complicated to derive asymptotic properties of a second-stage estimator ([Wang et al., 2007](#)). Specifically, no one has derived properties for estimator in this context— a second-stage kernel estimator that is subsequently dependent on the first-stage series estimators with a known discontinuity in one of the regression component function. A computational challenge for spline estimator is to solve the singularity problem regarding its matrix.

In a broader sense, spline estimators have great potential but are less explored as nonparametric estimators. They possess both the local characteristics of nonparametric estimation and the global characteristics of parametric estimation. Kernel estimators are extensively explored partly because they have tractable asymptotic properties, but they are myopic: the estimation ignores the global trend in data and only uses local observations. Furthermore, they can be computationally intense and wasteful. Evaluation and estimation are the same: to evaluate a regression at  $m$  points means to estimate  $m$  times. Besides, only part of the estimation outcome is useful. For example, the estimates of local linear estimators naturally provide the intercept and slope of the local regression while only the intercept is the object of interest. Spline estimators do not have these problems.



## References

- Abadie, A., J. Angrist, and G. Imbens (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70(1), 91–117.
- Angrist, J. D. and M. Rokkanen (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association* 110(512), 1331–1344.
- Apostol, T. (1974). *Mathematical Analysis*. New York: Addison Wesley.
- Arai, Y. and H. Ichimura (2018). Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator. *Quantitative Economics* 9(1), 441–482.
- Armstrong, T. B. and M. Kolesár (2017). A simple adjustment for bandwidth snooping. *The Review of Economic Studies* 85(2), 732–765.
- Armstrong, T. B. and M. Kolesár (2018). Optimal inference in a class of regression models. *Econometrica* 86(2), 655–683.
- Barreca, A. I., J. M. Lindo, and G. R. Waddell (2016). Heaping-induced bias in regression-discontinuity designs. *Economic Inquiry* 54(1), 268–293.
- Bartalotti, O. and Q. Brummet (2017). Regression discontinuity designs with clustered data. In *Regression Discontinuity Designs: Theory and Applications*, pp. 383–420. Emerald Publishing Limited.
- Bertanha, M. (2017). Regression discontinuity design with many thresholds. *Available at SSRN 2712957*.
- Bertanha, M. and G. W. Imbens (2014). External validity in fuzzy regression discontinuity designs. Technical report, National Bureau of Economic Research.
- Bugni, F. A. and I. A. Canay (2018). Testing continuity of a density via g-order statistics in the regression discontinuity design. *arXiv preprint arXiv:1803.07951*.
- Buja, A., T. Hastie, R. Tibshirani, et al. (1989). Linear smoothers and additive models. *The Annals of Statistics* 17(2), 453–510.
- Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2018). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 1–10.
- Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82(6), 2295–2326.
- Calonico, S., M. D. Cattaneo, and R. Titiunik (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association* 110(512), 1753–1769.
- Card, D., D. S. Lee, Z. Pei, and A. Weber (2015). Inference on causal effects in a generalized regression kink design. *Econometrica* 83(6), 2453–2483.

- Cattaneo, M. and J. C. Escanciano (Eds.) (2017). *Regression Discontinuity Designs: Theory and Applications*, Volume 38 of *Advances in Econometrics*. Emerald Group Publishing.
- Cattaneo, M. D., B. R. Frandsen, and R. Titiunik (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference* 3(1), 1–24.
- Cattaneo, M. D., R. Titiunik, G. Vazquez-Bare, and L. Keele (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics* 78(4), 1229–1248.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics* 6, 5549–5632.
- Cheng, M.-Y. (1994). On boundary effects of smooth curve estimators. *Unpublished dissertation, Department of Statistics, University of North Carolina at Chapel Hill*.
- Cheng, M.-Y., J. Fan, J. S. Marron, et al. (1997). On automatic boundary corrections. *The Annals of Statistics* 25(4), 1691–1708.
- Chernozhukov, V. and C. Hansen (2005). An iv model of quantile treatment effects. *Econometrica* 73(1), 245–261.
- Chiang, H. D. and Y. Sasaki (2019). Causal inference by quantile regression kink designs. *Journal of Econometrics*.
- Cook, T. D. (2008). "waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics* 142(2), 636–654.
- Davezies, L. and T. Le Barbanchon (2017). Regression discontinuity design with continuous measurement error in the running variable. *Journal of Econometrics* 200(2), 260–281.
- Dong, Y. (2010). Jumpy or kinky? regression discontinuity without the discontinuity.
- Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics* 30(3), 422–446.
- Dong, Y. and A. Lewbel (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics* 97(5), 1081–1092.
- Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14(1), 153–158.
- Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87(420), 998–1004.
- Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 196–216.
- Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* 20, 2008–2036.
- Feir, D., T. Lemieux, and V. Marmer (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics* 34(2), 185–196.

- Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics* 31(7), 799–815.
- Frandsen, B. R., M. Frölich, and B. Melly (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics* 168(2), 382–395.
- Frölich, M. (2007). Regression discontinuity design with covariates.
- Gao, W. Y. (2018). Minimax linear estimation at a boundary point. *Journal of Multivariate Analysis* 165, 262–269.
- Gasser, T. and H.-G. Müller (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pp. 23–68. Springer.
- Gasser, T. and H.-G. Müller (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 171–185.
- Gelman, A. and G. Imbens (2014). Why high-order polynomials should not be used in regression discontinuity designs. Technical report, National Bureau of Economic Research.
- Hahn, J., P. Todd, and W. van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1), 201–209.
- Härdle, W. and O. Linton (1994). Applied nonparametric methods. *Handbook of econometrics* 4, 2295–2339.
- Hestenes, M. R. (1941). Extension of the range of a differentiable function. *Duke Mathematical Journal* 8(1), 183–192.
- Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies* 79(3), 933–959.
- Imbens, G. and S. Wager (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics* 101(2).
- Imbens, G. and T. Zajonc (2011). Regression discontinuity design with multiple forcing variables. *Report, Harvard University.[972]*.
- Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142(2), 615–635.
- Kamat, V. (2018). On nonparametric inference in the regression discontinuity design. *Econometric Theory* 34(3), 694–703.
- Keele, L. J. and R. Titiunik (2015). Geographic boundaries as regression discontinuities. *Political Analysis* 23(1), 127–155.
- Kim, W., O. B. Linton, and N. W. Hengartner (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics* 8(2), 278–297.
- Kolesár, M. and C. Rothe (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review* 108(8), 2277–2304.

- Kuersteiner, G. M., D. C. Phillips, and M. Villamizar-Villegas (2018). Effective sterilized foreign exchange intervention? Evidence from a rule-based policy. *Journal of International Economics* 113, 118–138.
- Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2), 281–355.
- Li, F., A. Mattei, and F. Mealli (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 1906–1931.
- Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.
- Linton, O. and J. P. Nielsen (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 93–100.
- Litschig, S. and K. M. Morrison (2013). The impact of intergovernmental transfers on education outcomes and poverty reduction. *American Economic Journal: Applied Economics* 5(4), 206–240.
- Martins-Filho, C. and K. Yang (2007). Finite sample performance of kernel-based regression methods for non-parametric additive models under common bandwidth selection criterion. *Journal of Nonparametric Statistics* 19(1), 23–62.
- Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics* 142(2), 829–850.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2), 698–714.
- Mynbaev, K. and C. Martins-Filho (2018). Unified estimation of densities on bounded and unbounded domains. *Annals of the Institute of Statistical Mathematics*.
- Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10(2), 1–21.
- Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* 73(2), 166–179.
- Opsomer, J. D., D. Ruppert, et al. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics* 25(1), 186–211.
- Otsu, T., K.-L. Xu, and Y. Matsushita (2013). Estimation and inference of discontinuity in density. *Journal of Business & Economic Statistics* 31(4), 507–524.
- Otsu, T., K.-L. Xu, and Y. Matsushita (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics* 186(1), 94–112.
- Papay, J. P., J. B. Willett, and R. J. Murnane (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics* 161(2), 203–207.

- Porter, J. (2003). Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*, 5–19.
- Porter, J. and P. Yu (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics* 189(1), 132–147.
- Reardon, S. F. and J. P. Robinson (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness* 5(1), 83–104.
- Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and methods* 14(5), 1123–1136.
- Shen, S. and X. Zhang (2016). Distributional tests for regression discontinuity: Theory and empirical examples. *Review of Economics and Statistics* 98(4), 685–700.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 689–705.
- Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology* 51(6), 309–317.
- Tjøstheim, D. and B. H. Auestad (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association* 89(428), 1398–1409.
- Trochim, W. M. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, Volume 6. Sage Publications, Inc.
- van der Klaauw, W. (2008). Regression–discontinuity analysis: A survey of recent developments in economics. *Labour* 22(2), 219–245.
- Wang, L., L. Yang, et al. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics* 35(6), 2474–2503.
- Wong, V. C., P. M. Steiner, and T. D. Cook (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics* 38(2), 107–141.
- Xu, K.-L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics* 201(1), 1–18.
- Xu, K.-L. (2018). A semi-nonparametric estimator of regression discontinuity design with discrete duration outcomes. *Journal of Econometrics* 206(1), 258–278.
- Ziegler, K. (2001). On approximations to the bias of the nadaraya-watson regression estimator. *Journal of Nonparametric Statistics* 13(4), 583–589.

## Appendix A

### Supporting Lemmas

**Lemma 1.** *Suppose the assumptions of Theorem 3 are holding and  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$ .*

*The covariance of  $\hat{g}(x)$  and  $\hat{f}(x)$  for  $x \geq 0$  has the following representation*

$$Cov(\hat{g}(x), \hat{f}(x)) = \begin{cases} \frac{1}{nh} \left\{ m^+(0)f(0) \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right\}, & x = 0 \\ \frac{1}{nh} \left\{ m^+(x)f(x) \int_{\mathbb{R}} K^2(u) du + o(1) \right\}, & x > 0 \end{cases}.$$

*Proof.* Let  $\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n \left[ K\left(\frac{X_i - x}{h}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{X_i + x}{h}\right) \right] Y_i d_i$ ,  $w_0 = -1$ ,  $k_0 = -1$ ,  $u_i = \sum_{j=0}^{s+1} \frac{k_j}{w_j} K\left(\frac{X_i + x}{h}\right)$  and  $K\left(\frac{X_i - x}{h}\right) = K_i$ . Then, since  $\{X_i\}_{i=1,2,\dots}$  forms an i.i.d. sequence and  $E(Y|X = x) = m^+(x)$  for  $x \geq 0$ ,

$$\begin{aligned} Cov(\hat{g}(x), \hat{f}(x)) &= \frac{n}{(nh)^2} E \left[ m^+(X_1) u_1 d_1 K_1 \right] + \frac{n(n-1)}{(nh)^2} E \left[ m^+(X_1) u_1 d_1 \right] E \left[ K_1 \right] \\ &\quad - \frac{1}{h^2} E \left[ m^+(X_1) u_1 d_1 \right] E \left[ K_1 \right] \\ &= \frac{1}{nh^2} E \left[ m^+(X_1) u_1 d_1 K_1 \right] - \frac{1}{nh^2} E \left[ m^+(X_1) u_1 d_1 \right] E \left[ K_1 \right]. \end{aligned}$$

Put  $\frac{1}{h} E \left[ m^+(X_1) u_1 d_1 K_1 \right] = T_1$  and  $\frac{1}{h^2} E \left[ m^+(X_1) u_1 d_1 \right] E \left[ K_1 \right] = T_2$ , then  $Cov(\hat{g}(x), \hat{f}(x)) = \frac{1}{nh} T_1 - \frac{1}{n} T_2$ . As with the variance, the covariance will be different for  $x = 0$  and  $x > 0$ .

Case ( $x > 0$ ):

$$\begin{aligned}
T_1 &= \frac{1}{h} \int_0^\infty K^2 \left( \frac{X_1 - x}{h} \right) m^+(X_1) f(X_1) dX_1 \\
&\quad + \frac{1}{h} \int_0^\infty K \left( \frac{X_1 - x}{h} \right) \sum_{j=1}^{s+1} \frac{k_j}{w_j} K \left( \frac{\frac{X_1}{w_j} + x}{h} \right) m^+(X_1) f(X_1) dX_1 \\
&\text{letting } \frac{x - X_1}{h} = u \text{ in the first term and } \frac{\frac{X_1}{w_j} + x}{h} = u \text{ in the second} \\
&= \int_{-\infty}^{\frac{x}{h}} K^2(u) m^+(x - hu) f(x - hu) du \\
&\quad + \sum_{j=1}^{s+1} k_j \int_{\frac{x}{h}}^\infty K \left( w_j u - (w_j + 1) \frac{x}{h} \right) K(u) m^+(w_j(hu - x)) f(w_j(hu - x)) du = I_{00} + \sum_{j=1}^{s+1} k_j I_{0j},
\end{aligned}$$

where

$$I_{00} = \int_{-\infty}^{\frac{x}{h}} K^2(u) m^+(x - hu) f(x - hu) du,$$

and

$$I_{0j} = \int_{\frac{x}{h}}^\infty K \left( w_j u - (w_j + 1) \frac{x}{h} \right) K(u) m^+(w_j(hu - x)) f(w_j(hu - x)) du.$$

Now,

$$\begin{aligned}
\left| I_{00} - m^+(x) f(x) \int_{\mathbb{R}} K^2(u) du \right| &= \left| \int_{\mathbb{R}} K^2(u) \left[ m^+(x - hu) f(x - hu) - m^+(x) f(x) \right] du \right. \\
&\quad \left. - \int_{\frac{x}{h}}^\infty K^2(u) m^+(x - hu) f(x - hu) du \right| \\
&\leq \left| \int_{|u| \leq C} K^2(u) \left[ m^+(x - hu) f(x - hu) - m^+(x) f(x) \right] du \right| \\
&\quad + \left| \int_{|u| > C} K^2(u) \left[ m^+(x - hu) f(x - hu) - m^+(x) f(x) \right] du \right| \\
&\quad + \left| \int_{\frac{x}{h}}^\infty K^2(u) m^+(x - hu) f(x - hu) du \right|, \text{ for } C > 0 \\
&\leq C \bar{p}(Ch, x) \int_{|u| \leq C} K^2(u) du + C \int_{|u| > C} K^2(u) du + C \int_{\frac{x}{h}}^\infty K^2(u) du,
\end{aligned}$$

where the last equality follows from the uniform boundedness of  $K$ ,  $f$  and  $m^+$  and where  $\bar{p}(Ch, x)$  is as defined in the proof of Theorem 2. By continuity of  $f$  and the fact that  $\int K^2(u) du < C$ , for

all  $\epsilon > 0$ ,

$$\left| I_{00} - m^+(x)f(x) \int_{\mathbb{R}} K^2(u)du \right| \leq \epsilon. \quad (1.1)$$

Similarly,

$$|I_{0j}| \leq C \int_{\frac{x}{h}}^{\infty} K(u) du < \epsilon \text{ for all } n \text{ sufficiently large.} \quad (1.2)$$

Consequently,  $|T_1 - m^+(x)f(x) \int_{\mathbb{R}} K^2(u)du| \leq \epsilon$ .

Turning to  $T_2$ , I first observe that from standard properties of  $\hat{f}$  I have

$$\frac{1}{h} E \left( K \left( \frac{X_1 - x}{h} \right) \right) \rightarrow f(x). \quad (1.3)$$

Now, letting  $\frac{x-X_1}{h} = u$  and  $\frac{X_1+x}{w_j} = u$  I have

$$\begin{aligned} h^{-1} E \left[ m^+(X_1) u_1 d_1 \right] &= h^{-1} E \left( m^+(X_1) u_1 \right) \\ &= \frac{1}{h} \int_0^{\infty} K \left( \frac{X_1 - x}{h} \right) m^+(X_1) f(X_1) dX_1 \\ &\quad + \frac{1}{h} \sum_{j=1}^{s+1} \frac{k_j}{w_j} \int_0^{\infty} K \left( \frac{X_1 + x}{w_j} \right) m^+(X_1) f(X_1) dX_1 \\ &= \int_{-\infty}^{\frac{x}{h}} K(u) m^+(x - hu) f(x - hu) du \\ &\quad + \sum_{j=1}^{s+1} k_j \int_{\frac{x}{h}}^{\infty} K(u) m^+(w_j(hu - x)) f(w_j(hu - x)) du \\ &= I_1 + \sum_{j=1}^{s+1} I_{2j}. \end{aligned}$$

Using arguments similar to those used in the study of  $T_1$ , I have

$$\begin{aligned} \left| I_1 - m^+(x)f(x) \int_{\mathbb{R}} K(u)du \right| &\leq C\bar{p}(Ch, x) \int_{|u| \leq C} K(u)du + 2C \int_{|u| > C} K(u)du \\ &\quad + C \int_{\frac{x}{h}}^{\infty} K(u)du. \end{aligned}$$



By continuity of  $f(x)$  I have, for all  $\epsilon > 0$ ,  $|I_1 - m^+(x)f(x) \int_{\mathbb{R}} K^2(u)du| \leq \epsilon$ . Similarly,

$$|I_{2j}| \leq C \sum_{j=0}^{s+1} k_j \int_{\frac{x}{h}}^{\infty} |K(u)| du \leq \epsilon,$$

for sufficiently large  $n$ . Therefore,

$$\left| \frac{1}{h} E \left( m^+(X_1)u_1 \right) - m^+(x)f(x) \int_{\mathbb{R}} K^2(u)du \right| \leq \epsilon. \quad (1.4)$$

Thus for  $x > 0$ ,  $Cov(\hat{g}(x), \hat{f}(x)) = \frac{1}{nh}T_1 - \frac{1}{n}T_2 = \frac{1}{nh} (m^+(x)f(x) \int_{\mathbb{R}} K^2(u)du + o(1))$ .

Case ( $x = 0$ ): Repeating the change in variables used above, I have

$$T_1 = I_{00} + \sum_{j=1}^{s+1} \frac{k_j}{w_j} I_{0j},$$

where

$$I_{00} = \int_{-\infty}^0 K^2(u)m^+(-hu)f(-hu)du$$

and

$$I_{0j} = \int_0^{\infty} K \left( \frac{u}{w_0} \right) K \left( \frac{u}{w_j} \right) m^+(hu)f(hu)du.$$

Now,

$$\begin{aligned} \left| I_{00} - m^+(0)f(0) \int_{-\infty}^0 K^2(u)du \right| &= \left| \int_{-\infty}^{-C} K^2(u) \left[ m^+(-hu)f(-hu) - m^+(0)f(0) \right] du \right. \\ &\quad \left. + \int_{-C}^0 K^2(u) \left[ m^+(-hu)f(-hu) - m^+(0)f(0) \right] du \right| \\ &\leq C \int_{-\infty}^{-C} K^2(u)du + \bar{p}(Ch, 0) \int_0^C K^2(u)du. \end{aligned}$$

For  $n$  sufficiently large and all  $\epsilon > 0$ ,

$$\left| I_{00} - m^+(0)f(0) \int_{-\infty}^0 K^2(u)du \right| \leq \epsilon. \quad (1.5)$$

Similar arguments give,

$$\left| I_{0j} - m^+(0)f(0) \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du \right| < \epsilon. \quad (1.6)$$

Consequently, (1.5) and (1.6) give

$$\left| T_1 - m^+(0)f(0) \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du \right| < \epsilon.$$

Turning to  $T_2 = \frac{1}{h^2} E [m^+(X_1)u_1d_1] E \left[ K\left(\frac{X_1}{h}\right) \right]$  I have from the properties of  $\hat{f}$  that  $\frac{1}{h} E \left( K\left(\frac{X_1}{h}\right) \right) \rightarrow f(0)$ . Now, again changing variables,

$$\begin{aligned} \frac{1}{h} E (m^+(X_1)u_1d_1) &= \frac{1}{h} \int_0^\infty K\left(\frac{X_1}{h}\right) m^+(X_1)f(X_1)dX_1 \\ &\quad + \frac{1}{h} \sum_{j=1}^{s+1} \frac{k_j}{w_j} \int_0^\infty K\left(\frac{X_1}{w_jh}\right) m^+(X_1)f(X_1)dX_1 \\ &= I_1 + \sum_{j=1}^{s+1} \frac{k_j}{w_j} I_{2j}, \end{aligned}$$

where  $I_1 = \int_{-\infty}^0 K(u)m^+(-hu)f(-hu)du$  and  $I_{2j} = \int_0^\infty K(u)m^+(w_jhu)f(w_jhu)du$ . Using the same arguments as in the first case ( $x > 0$ ),

$$\left| I_1 - m^+(0)f(0) \int_{-\infty}^0 K(u)du \right| \leq C \int_{-\infty}^{-C} K(u)du + \bar{p}(Ch, 0) \int_0^C K(u)du.$$

For  $h$  be sufficiently small and continuity of  $f$  I have, for all  $\epsilon > 0$ ,  $\left| I_1 - m^+(0)f(0) \int_{-\infty}^0 K(u)du \right| \leq$

$\epsilon$ .

$$\begin{aligned}
\left| I_2 - m^+(0)f(0) \int_0^\infty K\left(\frac{u}{w_j}\right) du \right| &= \left| \int_0^\infty K\left(\frac{u}{w_j}\right) [m^+(hu)f(hu) - m^+(0)f(0)] du \right| \\
&= \left| \int_0^C K\left(\frac{u}{w_j}\right) [m^+(hu)f(hu) - m^+(0)f(0)] du \right. \\
&\quad \left. + \int_C^\infty K\left(\frac{u}{w_j}\right) [m^+(hu)f(hu) - m^+(0)f(0)] du \right| \\
&\leq \bar{p}(hC, 0) \int_0^C \left| K\left(\frac{u}{w_j}\right) \right| du + C \int_C^\infty K\left(\frac{u}{w_0}\right) K\left(\frac{u}{w_j}\right) du,
\end{aligned}$$

where for all  $\epsilon > 0$ ,  $\left| \int_C^\infty K\left(\frac{u}{w_j}\right) du \right| < \epsilon$  and for sufficiently small  $h$ ,  $\bar{p}(hC, 0) < \epsilon$ . Thus,

$$\left| I_2 - m^+(0)f(0) \int_0^\infty K\left(\frac{u}{w_j}\right) du \right| < \epsilon.$$

Consequently,

$$\left| \frac{1}{h} E\left(m^+(X_1)u_1\right) - m^+(0)f(0) \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K\left(\frac{u}{w_j}\right) du \right| < \epsilon. \quad (1.7)$$

Thus, for  $x = 0$ ,  $Cov(\hat{g}(x), \hat{f}(x)) = \frac{1}{nh}T_1 - \frac{1}{n}T_2 = \frac{1}{nh} \left[ m^+(0)f(0) \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right]$ .

In summary, I have

$$Cov(\hat{g}(x), \hat{f}(x)) = \begin{cases} \frac{1}{nh} \left\{ m^+(0)f(0) \sum_{j=0}^{s+1} \frac{k_j}{w_j} \int_0^\infty K(u) K\left(\frac{u}{w_j}\right) du + o(1) \right\}, & x = 0 \\ \frac{1}{nh} \left\{ m^+(x)f(x) \int_{\mathbb{R}} K^2(u) du + o(1) \right\}, & x > 0. \end{cases}$$

□

**Lemma 2.** Under the assumptions of Theorem 3,  $E(|Z_n(x)|) = O\left(\left(\frac{1}{nh}\right)^{\frac{3}{2}}\right)$ .

*Proof.*

$$\begin{aligned}
Z_n(x) &= 3(\hat{g}(x) - E\hat{g}(x))(\hat{f}(x) - E\hat{f}(x))^2 \int_0^1 \frac{(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^3} dt \\
&\quad - 3(\hat{f}(x) - E\hat{f}(x))^3 \int_0^1 (1-t)^2 \frac{[E\hat{g}(x) + t(\hat{g}(x) - E\hat{g}(x))]}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt = 3(J_1 - J_2).
\end{aligned}$$

Letting,

$$s_n = E(\hat{g}(x) - E\hat{g}(x)|X_1 \cdots, X_n) = \frac{1}{nh} \sum_{i=1}^n (z_{i,n} - E(z_{i,n})),$$

where  $z_{i,n} = K \left( \frac{X_i - x}{h} \right) I_{\{X_i \geq 0\}} m^+(x_i)$ , I have that

$$E(J_1) = \int s_n(\hat{f}(x) - E\hat{f}(x))^2 \int_0^1 \frac{(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^3} dt f(X) dX,$$

and

$$\begin{aligned} E(J_2) &= (\hat{f}(x) - E\hat{f}(x))^3 E(\hat{g}(x)) \int_0^1 \frac{(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt f(X) dX \\ &+ s_n(\hat{f}(x) - E\hat{f}(x))^3 \int_0^1 \frac{t(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt f(x) dx. \end{aligned}$$

By the Cauchy-Schwartz inequality

$$|EJ_1| \leq \left( \int s_n^2 (\hat{f}(x) - E\hat{f}(x))^4 f(X) dX \right)^{\frac{1}{2}} \left( \int \left( \int_0^1 \frac{(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^3} dt \right)^2 f(X) dX \right)^{\frac{1}{2}}.$$

Now, since  $(1-t)^2 \leq 1$  for  $0 \leq t \leq 1$ , letting  $d = (1-t)E\hat{f}(x) + t\hat{f}(x)$  I see that  $d > 0$  since  $E(\hat{f}(x)) \geq 0$ . Consequently,  $\int_0^1 \frac{1}{d^3} (1-t)^2 dt \leq \int_0^1 \frac{1}{d^3} dt$ . But,

$$\int_0^1 \frac{1}{d^3} dt = \frac{\hat{f}(x) + E\hat{f}(x)}{2\hat{f}(x)^2 (E\hat{f}(x))^2} = \frac{1}{2\hat{f}(x)E\hat{f}(x)^2} + \frac{1}{2\hat{f}(x)^2 E\hat{f}(x)}.$$

Now, since  $0 < B \leq E\hat{f}(x)$  I have  $\int_0^1 \frac{1}{d^3} dt \leq \frac{1}{B} \max \left\{ \frac{1}{\hat{f}(x)^2 B}, \frac{1}{\hat{f}(x)^2} \right\}$ . Taking  $\max \left\{ \frac{1}{\hat{f}(x)B}, \frac{1}{\hat{f}(x)} \right\} = \frac{1}{\hat{f}(x)B}$  I have

$$|EJ_1| \leq \left( \int s_n^2 (\hat{f}(x) - E\hat{f}(x))^4 f(X) dX \right)^{\frac{1}{2}} \left( \frac{1}{B^2} \int \left( \frac{1}{\hat{f}(x)} \right)^2 f(X) dX \right)^{\frac{1}{2}}.$$

Now,

$$\begin{aligned}
& |E(J_2)| \\
& \leq |J_{2,1}| + |J_{2,2}| \\
& \leq \left( \int \left( (\hat{f}(x) - E\hat{f}(x))^3 E\hat{g}(x) \right)^2 f(X) dX \right)^{\frac{1}{2}} \left( \int \left( \int_0^1 \frac{(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt \right)^2 f(X) dX \right)^{\frac{1}{2}} \\
& + \left( \int \left( s_n(\hat{f}(x) - E\hat{f}(x))^3 \right)^2 f(X) dX \right)^{\frac{1}{2}} \left( \int \left( \int_0^1 \frac{t(1-t)^2}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt \right)^2 f(X) dX \right)^{\frac{1}{2}}.
\end{aligned}$$

Now,  $t(1-t)^2 \leq t \leq 1$ , hence

$$\begin{aligned}
& \int_0^1 \frac{1}{[E\hat{f}(x) + t(\hat{f}(x) - E\hat{f}(x))]^4} dt \\
& = \frac{(E\hat{f}(x))^2 + E\hat{f}(x)\hat{f}(x) + \hat{f}(x)^2}{3E\hat{f}(x)^3\hat{f}(x)^3} \\
& = \frac{1}{3E\hat{f}(x)\hat{f}(x)^3} + \frac{1}{3E\hat{f}(x)^2\hat{f}(x)^2} + \frac{1}{3E\hat{f}(x)^3\hat{f}(x)} \\
& \leq \frac{1}{B} \max \left\{ \frac{1}{\hat{f}(x)^3}, \frac{1}{B\hat{f}(x)^2}, \frac{1}{B^2\hat{f}(x)} \right\}.
\end{aligned}$$

Suppose  $\max \left\{ \frac{1}{\hat{f}(x)^3}, \frac{1}{B\hat{f}(x)^2}, \frac{1}{B^2\hat{f}(x)} \right\} = \frac{1}{\hat{f}(x)B^2}$ . Then,

$$\begin{aligned}
|EJ_2| & \leq \left\{ \left( \int \left( (\hat{f}(x) - E\hat{f}(x))^3 E\hat{g}(x) \right)^2 f(x) dx \right)^{\frac{1}{2}} + \left( \int \left( s_n(\hat{f}(x) - E\hat{f}(x))^3 \right)^2 f(X) dX \right)^{\frac{1}{2}} \right\} \\
& \times \left( \frac{1}{B^4} \int \left( \frac{1}{\hat{f}(x)} \right)^2 f(X) dX \right)^{\frac{1}{2}}.
\end{aligned}$$

Now,

$$\int \left( \frac{1}{\hat{f}(x)} \right)^2 f(X) dX = \int_0^\infty \int e^{-\lambda\hat{f}(x)^2} f(X) dX d\lambda = \int_0^\infty E \left( e^{-\lambda\hat{f}(x)^2} \right) d\lambda.$$

Under the conditions of Theorem 3, and by Slutsky Theorem,  $\hat{f}(x)^2 \xrightarrow{P} f(x)^2$ . Thus, by Lebesgue's

dominated convergence Theorem

$$h_n(x, \lambda) = E \left( e^{-\lambda \hat{f}(x)^2} \right) \rightarrow E \left( e^{-\lambda f(x)^2} \right) = \left( e^{-\lambda f(x)^2} \right) = h(x, \lambda).$$

since  $|h_n(x, \lambda)| \leq \int |e^{-\lambda \hat{f}(x)^2}| f(x) dx \leq 1$  for all  $n, \lambda$ . Thus,  $h_n(x, \lambda)$  is bounded and convergent on  $[0, \infty)$ . Then, by Arzèla's Theorem [Apostol, 1974](#), p. 228

$$\lim_{n \rightarrow \infty} \int_0^\infty h_n(x, \lambda) d\lambda \rightarrow \int_0^\infty E \left( e^{-\lambda \hat{f}(x)^2} \right) d\lambda = \frac{1}{f(x)^2} \leq C$$

since  $0 < B < f(x)$ . Thus,

$$\begin{aligned} |E(J_1) + E(J_2)| &\leq C \left( \left( \int s_n^2 (\hat{f}(x) - E\hat{f}(x))^4 f(X) dX \right)^{\frac{1}{2}} + \left( \int ((\hat{f}(x) - E\hat{f}(x))^3 E\hat{g}(x))^2 f(X) dX \right)^{\frac{1}{2}} \right. \\ &\quad \left. + \left( \int (s_n (\hat{f}(x) - E\hat{f}(x))^3)^2 f(x) dx \right)^{\frac{1}{2}} \right) \end{aligned}$$

Now,

$$\begin{aligned} E|\hat{f} - E\hat{f}|^3 &\leq \left( E|\hat{f} - E\hat{f}|^2 \right)^{\frac{1}{2}} \left( E|\hat{f} - E\hat{f}|^4 \right)^{\frac{1}{2}} \\ &= (V\hat{f})^{\frac{1}{2}} \left( E \left( \hat{f} - E\hat{f} \right)^4 \right)^{\frac{1}{2}} \\ &= O((nh)^{-1})^{\frac{1}{2}} (O((nh)^{-2}))^{\frac{1}{2}} \end{aligned}$$

by Hölder's Inequality and the fact that  $E \left( \hat{f} - E\hat{f} \right)^4 = O((nh)^{-2})$  from [Ziegler, 2001](#). Then,

$$E|\hat{f} - E\hat{f}|^3 \leq (nh)^{-\frac{1}{2}} (nh)^{-1} O(1) = (nh)^{-\frac{3}{2}} O(1),$$

$E\hat{g}(x) = O(1)$ , and

$$s_n = (\hat{g}(x) - E(\hat{g}(x)|X_1, \dots, X_n)) = O_p((nh)^{-\frac{1}{2}}).$$

Thus,  $E(Z_n(x)) = O \left( \left( \frac{1}{nh} \right)^{\frac{3}{2}} \right)$ . □

## Appendix B

### Matlab Codes for Simulations

In this appendix, I provide Matlab codes that construct the regression estimators under the study. They are not the complete list of codes but building blocks to implement the simulation, empirical application, and estimation strategy described in Chapter 2 and 3.

#### B.1 Regression Estimators

**Listing B.1:** kays.m Generate Hestenes Parameters

```
% This function calculates Hestenes coefficients.
%INPUTS:
% s: a natural number denotes the smoothness of the function.
% b: If b=0, the sequence  $w_i=1/i$ ; if b=1,  $w_i=i$ .
% OUTPUTS:
% Hestenes coefficients, an  $(s+1) \times 2$  matrix of  $\{k_j, w_j\}$  for  $j = 1, \dots, s+1$ 

function res_out = kays(s,b)
    if b==0
        w = 1./(1:1:s+1); %elementary wise division
        M = ones(s+1,s+1);
        for r=2:s+1
            M(r,:) = (-w).^(r-1);
        end
        k=M\ones(s+1,1);
        res_out = [w' k];
    elseif b==1
        w = (1:1:s+1); %start from 1, increase by 1, until s+1
        M = ones(s+1,s+1);
        for r=2:s+1
            M(r,:) = (-w).^(r-1);
        end
        k=M\ones(s+1,1); %M^(-1)*ones
        res_out = [w' k];
    end
end
```

**Listing B.2:** unwm Univariate Nadaraya-Watson Estimator

```
% This function performs univariate Nadaraya-Watson estimation
% INPUTS:
```

```

% y: an (nX1) column vector containing the observations on the regressand
% x: an (nX1) matrix containing the observations on the regressors
% evl: a (pX1) matrix containing the points where it is desired to
%     evaluate the estimator.  p can be any integer {1,2,...}
% Kn: the kernel. if Kn=1 the gaussian kernel is used; if K=2 the
%     Epanechnikov kernel is used; if K=3 the triangular kernel is used;
%     if K=4 the rectangular kernel is used;
% h: a scaler of bandwidths.
% s: a natural number. the smoothness of the function
% b: If b=0, the sequence w_i=1/i; if b=1, w_i=i.
% OUTPUTS:
% mhat: a vector of estimates of dimension p X 1
function mhat = unw(y,x,evl,Kn,h)

d=(1/h)*gsubtract(x,evl');

    if Kn==1 % Gaussian Kernel

        k = (1/sqrt(2*pi))*exp(-0.5*d.^2);

    elseif Kn==2 % Epanechnikov

        k = (0.75*(1-d.^2)).*(abs(d)<= 1);

    elseif Kn==3 % Triangular

        k = (1-abs(d)).*(abs(d)<= 1);

    elseif Kn==4 % Rectangular kernel

        k = double(abs(d)<= 1);

    end

mhat = ((y'*k)./sum(k))';
end

```

### Listing B.3: mvnw.m Multivariate Nadaraya-Watson Estimator

```

% This function performs multivariate Nadaraya-Watson estimation
% based the Gaussian kernel.
% INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% x: an (nXnreg) matrix containing the observations on the regressors
% evl: a (pXnreg) matrix containing the points where it is desired to
%     evaluate the estimator.  p can be any integer {1,2,...}
% h: is a (nregX1) vector of bandwidths.
% OUTPUTS:
% mhat: a matrix of estimates of dimension p X nreg

function mvnw = mvnw(y,x,evl,h)
n=size(y,1);

```



```

p=size(evl,1);
K=ones(n,p);
nreg=size(x,2);
for j=1:nreg
    a=(1/h(j))*gsubtract(x(:,j),evl(:,j)');
    k=(1/sqrt(2*pi))*exp(-0.5*a.^2); % Gaussian Kernel
    K=K.*k;
end

mvnw = ((y'*K)./sum(K))';
end

```

#### Listing B.4: ulle.m Univariate Local Linear Estimator

```

% This code does the Fan modification for a univariate local linear (help
% avoiding) singularity. Univariate local linear estimation with different
% kernels
% INPUTS
% y: a vector of regressands of dimension n X 1
% x: a vector of regressors of dimension n X 1
% e: a vector with points of evaluation of dimension p X 1
% h: bandwidth
% Kn: the kernel. if Kn=1 the gaussian kernel is used; if K=2 the
%     Epanechnikov kernel is used; if K=3 the triangular kernel is used;
%     if K=4 the rectangular kernel is used;
% OUTPUTS
% mhat: a vector of estimates of dimension p X 1
function [ mhat ] = ulle(y,x,e,Kn,h)
    n=size(y,1);
    p = size(e,1);
    mhat=zeros(p,1);
    diff=gsubtract(x,e'); % estimate point
    d=(1/h)*diff;

    if Kn==1 % Gaussian Kernel

        K = (1/sqrt(2*pi))*exp(-0.5*d.^2);

    elseif Kn==2 % Epanechnikov

        K = (0.75*(1-d.^2)).*(abs(d)<= 1);

    elseif Kn==3 % Triangular

        K = (1-abs(d)).*(abs(d)<= 1);

    elseif Kn==4 % Rectangular kernel

        K = double(abs(d)<= 1);

    end

```

```

    for j=1:p
        R = [ones(n,1) diff(:,j)];
        P = diag(K(:,j));
        m = R'*P*R;
        s(j,:) = (1/(det(m)+n^(-2)))*[m(2,2) -m(1,2)]*R'*P;
    end
    mhat = s*y;

end

```

### Listing B.5: mvllin.m Multivariate Local Linear Estimator

```

% This function performs multivariate local linear estimation
% based the Gaussian kernel.
%INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% x: an (nXnreg) matrix containing the observations on the regressors
% evl: a (pXnreg) matrix containing the points where it is desired to
% evaluate the estimator. p can be any integer {1,2,...}
% h: is a (nregX1) vector of bandwidths.

function mvllin = mvllin(y,x,evl,h)
n=size(y,1);
p=size(evl,1);
nreg=size(x,2);
K=ones(n,p);
s=zeros(p,n);
e = [1 zeros(1,nreg)];
for j=1:nreg
    a=(1/h(j))*gsubtract(x(:,j),evl(:,j)');
    k=(1/sqrt(2*pi))*exp(-0.5*a.^2); % Gaussian Kernel
    K=K.*k;
end

for i=1:p
    b=gsubtract(x,evl(i,:)); %b: nXnreg
    R=[ones(n,1) b];
    P=diag(K(:,i));
    s(i,:)=e*((R'*P*R)\(R'*P)); % R: nX(1+nreg) p: nXn s: pXn
end
mvllin = s*y;

end

```

### Listing B.6: uhest.m Univariate Hestenes Estimator

```

% This function performs univariate Hestenes estimation
% INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% X: an (nX1) matrix containing the observations on the regressors
% x: a (pX1) matrix containing the points where it is desired to
% evaluate the estimator. p can be any integer {1,2,...}
% Kn: the kernel: if Kn=1 the gaussian kernel is used; if K=2 the

```

```

%      Epanechnikov kernel is used; if K=3 the triangular kernel is used;
%      if K=4 the rectangular kernel is used;
% h: is a scaler of bandwidths.
% s: a natural number denotes the smoothness of the function.
% b: If b=0, the sequence  $w_i=1/i$ ; if b=1,  $w_i=i$ .
% OUTPUTS:
% mhat: vector of estimates of dimension p X 1
function m = uhest(Y,X,x,K,h,s,b)

d=(1/h)*gsubtract(x,X')';
result=kays(s,b);
w=result(:,1);
k=result(:,2);
ratio = k./w;
nk = zeros(size(X,1),size(x,1));

    if K ==1 % Gaussian Kernel

        k0 = (1/sqrt(2*pi))*exp(-0.5*d.^2);

    elseif K==2 % Epanechnikov

        k0 = (0.75*(1-d.^2)).*(abs(d)<= 1);

    elseif K==3 % Triangular

        k0 = (1-abs(d)).*(abs(d)<= 1);

    elseif K==4 % Rectangular kernel

        k0 = double(abs(d)<= 1);

    end

for j=1:s+1
    Z=X./w(j);
    a=(1/h)*gsubtract(x,-Z')';
    nk = nk+ratio(j)*(1/sqrt(2*pi))*exp(-0.5*a.^2);
end
denom =sum(k0+nk);
m = ((Y'*(k0+nk))./denom)';
end

```

## B.2 Density Estimators

### Listing B.7: mvfrp.m Multivariate NW Density Estimator

```

% This function calculates the Rosenblatt-Parzen kernel density estimator.
% The output is a column vector with dimension equal to that of the vector
% x. X is a column vector of dimension n (number of observations), K=1

```

```

% means a standard gaussian kernel is used, K=2 means an Epanechnikov
% kernel is used. x is a m X 1 vector, where m is the number of points
% where the estimator is evaluated.
% based the Gaussian kernel .
%INPUTS:
% x: an (nXnreg) matrix containing the observations on the regressors
% evl: a (pXnreg) matrix containing the points where it is desired to
% evaluate the estimator. p can be any integer {1,2,...}
% h: is a (nregX1) vector of bandwidths.

function fhat = mvfrp(x,evl,h)

    n=size(x,1);
    p=size(ev1,1);
    nreg=size(x,2);
    K=ones(n,p);
    H=1;
    s=zeros(p,n);
        for j=1:nreg
            a=(1/h(j))*gsubtract(x(:,j),evl(:,j)');
            k=(1/sqrt(2*pi))*exp(-0.5*a.^2); % Gaussian Kernel
            K=K.*k;
            H=H*h(j);
        end
    fhat =(1/(n*H))*sum(K)';

end

```

**Listing B.8:** Univariate LL Density Estimator with a Boundary on the Right

```

% This function calculates the local linear binning density estimates on
% the left side of the discontinuity point. The method is proposed by
% Cheng(1994).
% INPUTS:
% X: an n X 1 vector containing all observations (left and right of the
% discontinuity)
% x: an p X 1 vector containing all points of desired evaluation
% (left and right of the discontinuity)
% Kn: the kernel. if Kn=1 the gaussian kernel is used; if K=2 the
% Epanechnikov kernel is used; if K=3 the triangular kernel is used;
% if K=4 the rectangular kernel is used;
% h: a scaler of bandwidths.
% OUTPUTS:
% mhat: a vector of estimates of dimension p X 1
function fhat_ng = fll_jn(X,x,K,h)

n=size(X,1);

%Rule of Thumb Bin Size (b) for the entire line

b=2*std(X,1)*(n^(-1/2));
J=floor((max(X)-min(X))/b)+1;
c=zeros(2*J,1);

```

```

G=zeros(2*J,1);
Y=zeros(2*J,1);

for i=1:2*J
    G(i)=((i-J)-0.5)*b;
    c(i)=sum( G(i)-0.5*b<=X & X<G(i)+0.5*b);
    Y(i)=(1/(n*b))*c(i);
end

% Grid (G) and normalized counts (Y) over the entire real line

D=[G Y];

% Choose only the negative side for points of evaluation, grid and counts
xn = x(x<=0);%
Dn = D(G<0,:);
fhat_ng = ulle(Dn(:,2),Dn(:,1),xn,K,h);

end

```

**Listing B.9:** Univariate LL Density Estimator with a Boundary on the Left

```

% This function calculates the local linear binning density estimates on
% the right side of the discontinuity point. The method is proposed by
% Cheng(1994).
% INPUTS:
% X: an n X 1 vector containing all observations (left and right of the
%     discontinuity)
% x: an p X 1 vector containing all points of desired evaluation
%     (left and right of the discontinuity)
% Kn: the kernel. if Kn=1 the gaussian kernel is used; if K=2 the
%     Epanechnikov kernel is used; if K=3 the triangular kernel is used;
%     if K=4 the rectangular kernel is used;
% h: a scaler of bandwidths.
% OUTPUTS:
% mhat: a vector of estimates of dimension p X 1

function fhat_pg = fll_jp(X,x,K,h)

% X: an n X 1 vector containing all observation (left and right of the
% discontinuity)
% x: an n X 1 vector containing all points of desired evaluation
%(left and right of the discontinuity)
% K is the kernel: if K=1 the gaussian kernel is used, if K=2 the
% Epanechnikov kernel is used and if K=3 the triangular kernel is used.

n=size(X,1);

%Rule of Thunb Bin Size (b) for the entire line

b=2*std(X,1)*(n^(-1/2));
J=floor((max(X)-min(X))/b)+1;
c=zeros(2*J,1);

```

```

G=zeros(2*J,1);
Y=zeros(2*J,1);

for i=1:2*J
    G(i)=((i-J)-0.5)*b;
    c(i)=sum( G(i)-0.5*b<=X & X<G(i)+0.5*b);
    Y(i)=(1/(n*b))*c(i);
end

% Grid (G) and normalized counts (Y) over the entire real line
D=[G Y];

% Choose only the positive side for points of evaluation, grid and counts
xp = x(x>=0);
Dp = D(G>=0,:);
fhat_pg = ulle(Dp(:,2),Dp(:,1),xp,K,h);
end

```

**Listing B.10:** Univariate Hestnes Density Hestenes Estimator with a Boundary on the Right

```

% This function calculates the Hestenes density estimates on
% the left side of the discontinuity point. The method is proposed by
% Mynbaev and Martins-Filho(2018)
% INPUTS:
% X: an n X 1 vector containing all observations (left and right of the
%     discontinuity)
% x: an p X 1 vector containing all points of desired evaluation
%     (left and right of the discontinuity)
% Kn: the kernel. if Kn=1 the gaussian kernel is used; if K=2 the
%     Epanechnikov kernel is used; if K=3 the triangular kernel is used;
%     if K=4 the rectangular kernel is used;
% h: a scaler of bandwidths.
% s: a natural number denotes the smoothness of the function.
% b: If b=0, the sequence w_i=1/i; if b=1, w_i=i.
% OUTPUTS:
% mhat: a vector of estimates of dimension p X 1

function fhat_ = fhest_jn(X,x,K,h,s,b)

n = size(X,1);

X_ = X(X<0);
X_=(-1)*X_;

x_ = x(x<=0);
x_ = (-1)*x_;

a0_ = (1/h)*gsubtract(x_,X_')';

result=kays(s,b);
w=result(:,1);
k=result(:,2);
ratio = k./w;

```

```

nk_ = zeros(size(X_,1),size(x_,1));

if K==1

    k0_ = (1/sqrt(2*pi))*exp(-0.5*a0_.^2);

    for j=1:s+1
        Z_ = X_/w(j);
        a_=(1/h)*gsubtract(x_,-Z_')';
        nk_ = nk_+ratio(j)*(1/sqrt(2*pi))*exp(-0.5*a_.^2);
    end

    fhat_ =(1/(n*h))*sum(k0_+nk_')';

elseif K==2
    k0_ = 0.75*(1-a0_.^2).*(abs(a0_)<=1);

    for j=1:s+1
        Z_ = X_/w(j);
        a_=(1/h)*gsubtract(x_,-Z_')';
        nk_ = nk_+ ratio(j)*0.75*(1-a_.^2).*(abs(a_)<=1);
    end

    fhat_ =(1/(n*h))*sum(k0_+nk_')';

elseif K==3

    k0_=(1-abs(a0_)).*(abs(a0_)<=1);

    for j=1:s+1
        Z_ = X_/w(j);
        a_=(1/h)*gsubtract(x_,-Z_')';
        nk_ = nk_+ ratio(j)*(1-abs(a_)).*(abs(a_)<=1);
    end

    fhat_ =(1/(n*h))*sum(k0_+nk_')';

end

end

end

```

**Listing B.11:** Univariate Hestnes Density Hestenes Estimator with a Boundary on the Left

```

% This function calculates the Hestenes density estimates on
% the left side of the discontinuity point. The method is proposed by
% Mynbaev and Martins-Filho(2018)
% INPUTS:
% X: an n X 1 vector containing all observations (left and right of the
%     discontinuity)
% x: an p X 1 vector containing all points of desired evaluation
%     (left and right of the discontinuity)
% Kn: the kernel. if Kn=1 the gaussian kernel is used; if K=2 the
%     Epanechnikov kernel is used; if K=3 the triangular kernel is used;

```

```

%      if K=4 the rectangular kernel is used;
% h: a scaler of bandwidths.
% s: a natural number denotes the smoothness of the function.
% b: If b=0, the sequence  $w_i=1/i$ ; if b=1,  $w_i=i$ .
% OUTPUTS:
% mhat: a vector of estimates of dimension  $p \times 1$ 

```

```

function fhat = fhest_jp(X,x,K,h,s,b)

n = size(X,1);
X_p = X(X>=0);
x_p = x(x>=0);

a0 = (1/h)*gsubtract(x_p,X_p')';

result=kays(s,b);
w=result(:,1);
k=result(:,2);
ratio = k./w;

nk = zeros(size(X_p,1),size(x_p,1));

if K==1

    k0 = (1/sqrt(2*pi))*exp(-0.5*a0.^2);

    for j=1:s+1
        Z = X_p./w(j);
        a=(1/h)*gsubtract(x_p,-Z')';
        nk = nk +ratio(j)*(1/sqrt(2*pi))*exp(-0.5*a.^2);
    end

    fhat = (1/(n*h))*sum(k0+nk)';

elseif K==2

    k0 = (0.75*(1-a0.^2)).*(abs(a0)<=1);

    for j=1:s+1
        Z = X_p./w(j);
        a=(1/h)*gsubtract(x_p,-Z')';
        nk = nk + ratio(j)*0.75*(1-a.^2).*(abs(a)<=1);
    end

    fhat = (1/(n*h))*sum(k0+nk)';

elseif K==3

    k0 = (1-abs(a0)).*(abs(a0)<=1);

    for j=1:s+1

```



```

Z = X_p./w(j);
a=(1/h)*gsubtract(x_p,-Z')';
nk = nk + ratio(j)*(1-abs(a)).*(abs(a)<=1);
end

fhat =(1/(n*h))*sum(k0+nk)';

```

```
end
```

```
end
```

### B.3 Estimators for the Additive Model

#### Listing B.12: Multivariate Marginal Integration Estimator

```

% This function performs multivariate marginal integration estimation
% based
% on the Gaussian kernel
% INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% x: an (nXD) matrix containing the observations on the regressors
% evl: a (pXD) matrix containing the points where it is desired to
% evaluate the estimator. p can be any integer {1,2,...}
% h: a (DX1) matrix of bandwidths.
% OUTPUTS:
% mhat: a (PXD) vector of estimates

function mhat = mi_estimate(y,x,evl,h)

P = size(ev1,1); % the number of evaluation points
D = size(x,2); % the number of regressors
n = size(x,1); % sample size

for d = 1:D
    for p = 1:P
        % form the n evaluation points
        xg = x;
        x1g = repmat(ev1(p,d),n,1);
        xg(:,d)=x1g;
        mdp = mvnw(y,x,xg,h);
        mhat(d,p) = mean(mdp);
    end
end

mhat = mhat';

end

```

#### Listing B.13: Multivariate Instrument Variable Estimator

```

% This function performs multivariate instrument variable estimation
% based on the Gaussian kernel.
% INPUTS:

```

```

% y: an (nX1) column vector containing the observations on the regressand
% x: an (nXD) matrix containing the observations on the regressors
% evl: a (pXD) matrix containing the points where it is desired to
%     evaluate the estimator. p can be any integer {1,2,...}
% h: a (DX1) matrix of bandwidths.
% OUTPUTS:
% mhat: a (PXD) vector of estimates

function mhat = iv_estimate(y,x,evl,h)

    P = size(ev1,1); % the number of evaluation points
    D = size(x,2); % the number of regressors
    n = size(x,1); % sample size

    % First stage: get T and pilot estimators and evaluate on
    % all sample points
    for d = 1:D
        fd = mvfrp(x(:,d),x(:,d),h); % marginal density
            fd: nX1
        xd = x ;
        xd(:,d) = []; % delete the d column
        f_d = mvfrp(xd,xd,h); % marginal density f(-d):
            nX1

        fj = mvfrp(x,x,h); % joint density
        T(:,d) = ((fd.*f_d)./fj).*y;
        mhat(d,:) = mvnw(T(:,d),x(:,d),evl(:,d),h)'; %
            univariate estimation nX1
    end

    mhat = mhat';

end

```

#### Listing B.14: Multivariate Backfitting Estimator

```

% This function performs multivariate Backfitting estimation
% based the Gaussian kernel.
% INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% x: an (nXD) matrix containing the observations on the regressors
% evl: a (pXD) matrix containing the points where it is desired to
%     evaluate the estimator. p can be any integer {1,2,...}
% h: a (DX1) matrix of bandwidths.
% OUTPUTS:
% mhat: a (PXD) vector of estimates

function mhat = backfitting_estimate(y,x,evl,h)

    % split the sample

```

```

YX = sortrows([y,x,evl],2);
r = x(:,1) ;
Yn = YX(r<0,1);
Yp = YX(r>=0,1);

Xn = x(r<0,:);
Xp = x(r>=0,:);
x = [Xn ; Xp];

En = evl(r<0,:);
Ep = evl(r>=0,:);
evl = [En ; Ep];

P = size(evl,1); % the number of evaluation points
D = size(x,2); % the number of regressors

res2(1) = 0;
res2(2) = sum(y.^2);
mhat(1:P,1:D) = 0;

j = 2; s=1; b = 1 ;
while abs(res2(j)-res2(j-1))> 0.01
    for d = 1:D
        lag = mhat;
        lag(:,d)=[]; % delete the current column
        yd = y - sum(lag,2);
        if d == 1
            m_n = uhest(Yn,Xn(:,d),En(:,d),h,s,b);
            m_p = uhest(Yp,Xp(:,d),Ep(:,d),h,s,b);
            mhat(:,d) = [m_n ; m_p];
        else
            mhat(:,d) = unw(yd,x(:,d),evl(:,d),h); %
                univariate estimation nX1
        end
    end

    end

    j = j+1;
    res2(j) = sum((y - sum(mhat,2)).^2);

    if j >100
        break
    end

end

yhat = sum(mhat,2); % sum of each row

end

```

### Listing B.15: Multivariate BSpine Estimator

```

% This function performs multivariate B_Spine estimation
% INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% x: an (nXP) matrix containing the observations on the regressors
% K: the order of polynomial
% OUTPUTS:
% mhat: a (nXP) vector of estimates

function mhat = bspline_estimate(y,x,k,h)
n = size(y,1);
nk = sqrt(n); % it can be changed to other formulas

P = size(x,2) ; % the dimensions of X
X = []; % basis functions evaluated at the evaluation points

% split the sample by X1
yx = sortrows([y,x],2);
yxn = yx(x(:,1)<0,:);
yxp = yx(x(:,1)>=0,:);

% Get basis functions for X1
t = augknt(linspace(min(yxn(:,2)), max(yxn(:,2)), sqrt(size(yxn,1)
)), k);
vn = size(t,2)- k; % the # of control points v = the # of knots m
- the order k
% Produce B-Spline basis matrix evaluated at sample points
Bn = bspline_basismatrix(k,t,yxn(:,2)');
A = zeros(size(yxp,1),size(Bn,2));
Bn = [Bn;A];

t = augknt(linspace(min(yxp(:,2)), max(yxp(:,2)), sqrt(size(yxp,1)
)), k);
vp = size(t,2)- k; % the # of control points v = the # of knots m
- the order k
Bp = bspline_basismatrix(k,t,yxp(:,2)');
A = zeros(size(yxn,1),size(Bp,2));
Bp = [A;Bp];

X = [X Bn Bp];

% Get basis functions for X2, X3 ...
for i = 3:P+1

% Construct all the knots
t = augknt(linspace(min(yx(:,i)), max(yx(:,i)), nk), k);
v = size(t,2)- k; % the dimension v = the # of knots m - the

```

```

        order k
    % Produce B-Spline basis matrix evaluated at sample points
    B = bspline_basismatrix(k,t,yx(:,i)');
    X = [X B] ;

end

% estimate beta-hat

    X = [ones(n,1) X]; % Add a column of ones to the matrix
    beta = (X'*X)\(X'*yx(:,1));

%     yhat = X*beta;

% estimate m-hat
    mhat = zeros(n,1); % mhat1 is a place holder
    sp = 0;
    ep = 1 + vn + vp; % skip the terms of m1

    for i = 2:P
        sp = ep + 1 ; % starting index
        ep = ep + v ; % end index
        C = beta(sp:ep);
        m = X(:,sp:ep)*C ;
        mhat = [mhat m];
    end
end

```

### Listing B.16: LL OLS Estimator

```

% This function calculates the local linear estimates for an additive
% linear model. The method is proposed by Calonico et al.(2016)
% INPUTS:
% y: an (nX1) column vector containing the observations on the regressand
% x: an (nX1) matrix containing the observations on the regressors
% z: an (nX(nreg-1)) matrix containing additional covariates
% t: a (pX1) matrix containing the points where it is desired to
%     evaluate the estimator. p can be any integer {1,2,...}
% h: a (pX1) vector of bandwidths.
% OUTPUTS:
% mhat: a vector of estimates of dimension p X 1
function theta = localOLS(y,x,z,t,h)
n=size(y,1);
vect = t*ones(n,1);
a=(1/h)*(x-vect);
K=(1/sqrt(2*pi))*exp(-0.5*a.^2);
R = [ (x-vect<0) x.*(x-vect<0) (x-vect>=0) x.*(x-vect >=0) z];
P=diag(K);
theta=((R'*P*R)\(R'*P))*y;

```