Essays in Nonparametric Econometrics

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Tomasz Olma

im Frühjahrs-/Sommersemester 2021

Abteilungssprecher:	Prof. Volker Nocke, Ph.D.
Referent:	Prof. Christoph Rothe, Ph.D.
Koreferent:	Prof. Yoshiyasu Rai, Ph.D.

Tag der Verteidigung:	20.07.2021

Acknowledgments

My doctoral studies have been both an exceptionally stimulating and challenging experience. During this time, I have received a lot of support from the academic community in Mannheim, for which I am sincerely thankful.

First of all, I would like to express my gratitude to my advisor Christoph Rothe for his invaluable support and guidance. Throughout the last four years, I have learned from him a lot about econometrics and academic work in general. Christoph was always available for discussions about research, and I benefited from his expertise and ability to convey the intuition behind complex problems in a simple manner. His insightful feedback, encouragement, and reliable strategic advice at crossroads provided a constant stimulus to advance my research.

I would also like to thank Yoshiyasu Rai for his constructive feedback, which helped me improve both my research and the way I communicate the results. I am grateful to Carsten Trenkler for valuable comments and discussions following all my seminar presentations, and to Markus Frölich for his support during the academic job market. I am also thankful to Antonio Ciccone for broadening my perspective on economic research.

I thank Claudia, my fellow student in econometrics, for being a great companion throughout the Ph.D. program. We learned together a lot about nonparametric econometrics and causal inference, and we shared the struggles of novice researchers. I am grateful for all that and for the fruitful collaboration on the third chapter of this thesis.

During my time in Mannheim, I experienced great kindness of the whole econometrics group. I thank in particular Alex, Cathrine, Giovanni, Jonas, and Majed for valuable feedback after my presentations and the very friendly atmosphere in the group.

I gratefully acknowledge all the administrative support and funding received during my graduate studies, which enabled me to focus primarily on research. This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences. I received further funding from the German Academic Exchange Service, from the German Research Foundation through the Collaborative Research Center Transregio 224 (Project A04; PI: Antonio Ciccone), and from the European Research Council through grant SH1-77202 (PI: Christoph Rothe).

I thank all my fellow Ph.D. students. It was wonderful to go through the first-year courses together. I also always enjoyed eating out together, watching and doing sports, and all other social activities. This friendly environment greatly contributed to making my stay in Mannheim a very enjoyable experience.

My gratitude goes also to all those who influenced me prior to my graduate studies. Starting from my school education in Bielsko-Biała, to the studies at the Warsaw School of Economics and the University of Warsaw, I was always fortunate to encounter dedicated teachers and mentors as well as fantastic fellow students, who stimulated my scientific curiosity and serve as role models for me up to this very day.

Finally, I would like to thank my parents for their constant support, sage advice, and encouragement to pursue my goals. *Dziękuję!*

Contents

List of Figures

List of Tables

Preface

Econometric methods serve to estimate functional relationships between economic variables. Classic, parametric techniques rely on modeling the observed data using a specification that is known up to a finite number of estimable parameters. Such procedures perform very well if the selected model is correctly specified, but they can lead to false conclusions under misspecification. Nonparametric methods, in contrast, allow researchers to model the relationships of interest in a flexible way, without imposing any functional form assumptions, subject only to smoothness conditions. For example, the local linear estimator of the conditional expectation of an outcome variable given covariates is based on the linear regression using observations in a small neighborhood of the covariate value of interest. If the neighborhood shrinks at an appropriate rate, under regularity conditions, the local linear estimator can estimate well any curve with bounded second-order derivative.

Owing to their flexibility, nonparametric methods have become popular in various areas of economics. Prominent examples include randomized experiments and regression discontinuity designs. In a randomized experiment, a treatment, e.g. social assistance or job training, is assigned to randomly selected units. To estimate the treatment effect conditional on covariates, one can employ nonparametric methods to flexibly estimate the conditional expectations in the treatment and control groups. In a regression discontinuity design, in turn, units are assigned a treatment if and only if their running variable, e.g. a poverty score, exceeds a fixed cutoff value. Under standard assumptions, a local average treatment effect can be identified by comparing units just to the left and just to the right of the cutoff. The corresponding conditional expectations are typically estimated using local linear methods.

There exists an extensive literature on estimation in the canonical settings for randomized experiments and regression discontinuity designs, and various refinements and robustifications have been developed. This dissertation provides two, practically relevant contributions to this literature. First, it revisits settings where the comparability of units in the treatment and control group breaks down due to some sample selection mechanism, in which case treatment effects are only partially identified. A novel method for estimation of bounds on conditional treatment effects is proposed for such settings. Second, it develops a procedure that efficiently incorporates covariates into an analysis of standard regression discontinuity designs.

On the technical side, the unifying theme of the proposed methods is two-stage estimation that is robust with respect to the first-stage estimation error. In the considered settings, the object to be estimated is a scalar parameter, but it involves an unknown, nuisance function that has to be estimated in the first stage. In general, the first-stage estimation may affect the properties of the final estimator in a potentially complicated manner. In order to avoid this issue, the proposed estimators are constructed in a way that renders them very insensitive to estimation of the nuisance function. This property has attractive theoretical and practical implications. The asymptotic results are valid under weak requirements on the first-stage estimator, and standard methods for conducting statistical inference can be easily adapted to the considered settings.

This dissertation consists of three chapters. Chapters 1 and 3 are self-contained. Chapter 2 builds on the estimation method developed in Chapter 1, but it can also be read independently of Chapter 1.

Chapter 1 concerns nonparametric estimation of truncated conditional expectation functions. Such functions are objects of interest in a wide range of economic applications, including income inequality measurement, financial risk management, and impact evaluation. They typically involve truncating the outcome variable above or below certain quantiles of its conditional distribution. In this estimation problem, the conditional quantile function is a nuisance parameter, which has to be estimated in the first stage. I propose an estimator that is immunized against the first-stage estimation error owing to the use of a Neyman-orthogonal moment in the second stage. This construction ensures that the proposed estimator has favorable bias properties and that inference methods developed for the standard nonparametric regression can be readily adapted to conduct inference on truncated conditional expectation functions. As an extension, I consider estimation with an estimated truncation quantile level. The proposed estimator is applied in two empirical settings: sharp regression discontinuity designs with a manipulated running variable and program evaluation with sample selection.

Chapter 2 is joint work with Christoph Rothe. We consider estimation and inference in fuzzy regression discontinuity designs with a manipulated running variable. In the manipulation framework proposed by Gerard, Rokkanen, and Rothe (2020), we derive alternative bounds on the treatment effect of interest. The proposed bounds are not sharp, but they take a simple form, and they can be estimated using the estimator of truncated conditional expectation functions developed in Chapter 1. We propose a method for constructing confidence sets for the partially identified treatment effect using test inversion in the spirit of Anderson-Rubin confidence sets.

Chapter 3 is joint work with Claudia Noack and Christoph Rothe. We propose a novel class of covariate-adjusted regression discontinuity estimators that can have a smaller variance than the estimators used in the literature. Our procedure accommodates a wide range of covariate adjustments under mild conditions. We consider classic parametric and nonparametric, as well as machine learning methods, so that suitable estimators can be chosen for any given type of covariates. We allow for discrete and continuous covariates in low- and high-dimensional settings. The proposed estimators are easily applicable because the tuning parameters, both in the first and second stage, can be selected and confidence intervals can be constructed following standard methods used in the literature. We characterize the covariate adjustments that lead to the smallest variance in this class of regression discontinuity estimators.

CHAPTER 1

Nonparametric Estimation of Truncated Conditional Expectation Functions

1.1. INTRODUCTION

A truncated sample mean is the mean calculated after discarding some of the highest and/or lowest values in a sample. Such quantities, which estimate the corresponding truncated expectations, are used in a wide range of economic applications. In studies of inequality, income dispersion can be summarized by reporting the mean income in different quintiles of its distribution, i.e., the mean income of the 20% of households with the lowest income, followed by the mean income of households between the 20th and 40th percentile of the income distribution, etc. (e.g. Semega et al., 2020). In finance, the expected shortfall denotes the expected value of a certain proportion, e.g. 5%, of top losses. It is a widely-used risk measure informing about the performance of a portfolio of assets in the worst-case scenarios (e.g. Chen, 2008). Truncated means are also used in settings with contaminated data, where the sharp bounds on the true expected outcome take the form of truncated expectations (Horowitz and Manski, 1995). The partial identification approach underlying this result has been adapted to several impact evaluation settings to address sample selection problems; see, e.g., Zhang and Rubin (2003); Lee (2009); Chen and Flores (2015).

In all the above examples, the analysis can be enriched by incorporating covariates. First, the anatomy of income inequality can be better understood when analyzed conditionally on characteristics such as age or work experience. Second, an estimator of the expected shortfall can be more informative if it takes into account covariates. Third, in impact evaluation, the heterogeneity of treatment effects can be explored based on individuals' characteristics. Furthermore, Gerard et al. (2020) apply the trimming approach of Horowitz and Manski (1995) to regression discontinuity designs with a manipulated running variable, which necessarily involve conditioning on a covariate.

In this paper, I propose a novel, nonparametric estimator of truncated expectations defined conditionally on covariates. As in the above examples, I consider setups where the outcome variable needs to be truncated above or below certain quantiles of its conditional distribution. For ease of exposition, I focus on one-sided truncation. I consider a nonparametric setting with a continuous outcome variable, denoted by Y, and a vector of continuous covariates, denoted by X.[1] For a quantile level $\eta \in (0,1)$ and x in the support of X, let $Q(\eta, x)$ be the conditional η-quantile of Y given $X = x$. My aim is to nonparametrically estimate the following function:

$$m(\eta, x) = \mathbb{E}[Y | Y \leq Q(\eta, X), X = x]. \tag{1.1.1}$$

I refer to η in the above definition as the truncation quantile level. It might be chosen by the analyst, in which case it is a fixed, known number, but in some applications it has to be estimated from the data. My setting is nonparametric, meaning that I do not impose any parametric restrictions on the joint distribution of (X, Y). In particular, the functions $m(\eta, x)$ and $Q(\eta, x)$ can be of any form, subject only to smoothness restrictions.

In this estimation problem, the function $Q(\eta, \cdot\,)$ is a nuisance parameter. If it was known, then based on a sample $\{(X_i, Y_i)\}_{i=1}^n$ from the distribution of (X, Y), one could estimate $m(\eta, x)$ using standard nonparametric regression techniques, e.g., kernel estimators, applied to the sample restricted to observations with $Y_i \leq Q(\eta, X_i)$. Alternatively, motivated by the equivalent representation of the estimand as:

$$m(\eta, x) = \frac{1}{\eta}\mathbb{E}[Y\mathbb{1}(Y \leq Q(\eta, X))|X = x], \tag{1.1.2}$$

one could run a nonparametric regression with $\frac{1}{\eta}Y_i\mathbb{1}(Y_i \leq Q(\eta, X_i))$ as the outcome variable. Feasible versions of these two estimators, however, require estimating the function $Q(\eta, \cdot\,)$ in the first stage. This additional estimation step affects the properties of the resulting estimators in a potentially complicated manner.

In order to avoid the transmission of the first-stage estimation error to the final estimator, I propose a modification of the latter approach, which utilizes a conditional moment that is Neyman-orthogonal to the nuisance function $Q(\eta, \cdot\,)$. Specifically, my estimation approach is based on the following representation of the estimand:

$$m(\eta, x) = \frac{1}{\eta}\mathbb{E}[Y\mathbb{1}(Y \leq Q(\eta, X)) - Q(\eta, X)(\mathbb{1}(Y \leq Q(\eta, X)) - \eta)|X = x]. \tag{1.1.3}$$

[1] If the covariates take on only a small number of distinct values, then the truncated conditional expectation function can be estimated using sample truncated means binned by covariate values.

Compared to (1.1.2), the conditional moment in (1.1.3) contains an additional term, which, however, is mean-zero conditional on X.[2] Its inclusion renders the conditional moment in (1.1.3) insensitive to small perturbations of $Q(\eta, \cdot)$ in the following sense. For the quantile level η and $q \in \mathbb{R}$, let

$$\psi(\eta, q) = \frac{1}{\eta}\left[Y\mathbb{1}(Y \leq q) - q(\mathbb{1}(Y \leq q) - \eta)\right]. \tag{1.1.4}$$

Equation (1.1.3) can be expressed as $m(\eta, x) = \mathbb{E}[\psi(\eta, Q(\eta, X))|X = x]$. This expression is insensitive to small perturbations of the conditional quantile function because the derivative of $\mathbb{E}[\psi(\eta, q)|X = x]$ with respect to q evaluated at the true conditional quantile $Q(\eta, x)$ is zero,

$$\frac{\partial}{\partial q}\mathbb{E}[\psi(\eta, q)|X]|_{q=Q(\eta,X)} = 0, \text{ a.s.} \tag{1.1.5}$$

Such orthogonal, or immunized, conditional moments feature prominently in the modern literature in setups where a nuisance parameter has to be estimated in the first stage (e.g. Belloni et al., 2017; Chernozhukov et al., 2018). In this literature, it is well understood that the orthogonality property immunizes the estimator against the first-stage estimation error.

Based on the orthogonal conditional moment in equation (1.1.3), my proposed estimator is constructed in two steps using local linear methods (Fan and Gijbels, 1996). In the first stage, I estimate the local linear approximation of the function $Q(\eta, \cdot)$. In the second stage, I run a local linear regression with a generated outcome variable corresponding to the expression under the conditional expectation in (1.1.3). The estimator is computationally easy to implement, and I show that the tuning parameters (bandwidths in the two local linear regressions) can be selected as in the standard nonparametric regression.

This paper contains two main theoretical results. First, I show that my estimator is asymptotically equivalent to the corresponding oracle estimator using the true function $Q(\eta, \cdot)$. Given this result, the asymptotic distribution follows from the standard theory of local linear estimation. The proposed estimator has good bias properties, and it is straightforward to adapt existing inference methods to do inference on truncated conditional expectation functions. Second, I study the asymptotic properties of my estimator when the truncation quantile level is estimated from the data. Under a high-level assumption on $\widehat{\eta}$, I derive an expansion of the estimator of the truncated conditional expectation function evaluated at $\widehat{\eta}$ about the estimator evaluated at the true value η. This expansion can be used on a case-by-case basis to derive the asymptotic distribution of the estimator evaluated at $\widehat{\eta}$ for specific estimators of the truncation quantile level.

[2] The conditional moment in (1.1.3) is the quantity of interest when the outcome variable has mass points, but, I show in this paper that there are reasons to consider this formula even with a continuous outcome variable.

I apply the proposed estimator in two empirical settings. First, I estimate bounds on the local average treatment effect in regression discontinuity designs with a manipulated running variable (Gerard et al., 2020). Second, I estimate bounds on the conditional wage effect of a job training program (Lee, 2009). These bounds involve truncated conditional expectation functions with truncation quantile levels that need to be estimated from the data.

Related Literature. The proposed two-stage procedure is similar to that of Linton and Xiao (2013). In the first stage, they estimate $Q(\eta, X_i)$ in a local polynomial quantile regression at X_i. In the second stage, they apply the Nadaraya-Watson estimator to the data with a generated outcome variable corresponding to the conditional moment in (1.1.3). My analysis, however, is different in three aspects. First, I employ a local linear estimator in the second stage, which is well-known to have favorable bias properties compared to the Nadaraya-Watson estimator.[3] Second, I estimate the function $Q(\eta, \cdot)$ based on a single local linear quantile regression. If one is interested in $m(\eta, x)$ for a specific covariate value, my approach is much simpler to implement than using a separate local polynomial quantile regression for each data point included in the second-stage regression. Third, and most importantly, the analysis of Linton and Xiao (2013) applies specifically to setups where the conditional variance of the outcome variable is infinite. While the presence of an infinite variance generally complicates the derivation of the asymptotic distribution, which is a non-normal, stable law, it makes some aspects of the analysis easier. Specifically, Linton and Xiao (2013) exploit the fact that under their assumptions the first-stage local polynomial quantile estimator converges faster than the respective oracle estimator. Their proof does not directly apply to models with finite variance of the outcome variable considered in this paper, where the first-stage and the oracle estimators have the same rates of convergence.

Other nonparametric estimators of truncated conditional expectation functions have been developed by Scaillet (2005), Cai and Wang (2008), and Kato (2012), who construct their estimators based on first-stage estimators of the conditional cumulative distribution function (c.d.f.) of the outcome variable. This estimation strategy, however, is not well-suited for estimation at boundary points of the support of the conditioning variables. The Nadaraya-Watson estimator of the conditional c.d.f.,[4] employed by Scaillet (2005), exhibits the so-called boundary effects in that its bias is of larger order at the boundary than in the interior. The bias properties of the Nadaraya-Watson can be improved upon using the local linear estimator, but it is not guaranteed to produce a proper c.d.f., as

[3] Linton and Xiao (2013) mention the possibility of using a local polynomial regression with a generated outcome variable $\frac{1}{\eta} Y_i \mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i))$, but they did not investigate it further.

[4] Estimation of a conditional c.d.f. can be cast as a regression problem with $\mathbb{1}(Y_i \leq y)$ as the outcome variable.

the resulting function can be nonmonotone and is not restricted to lie between zero and one. For that reason, Cai and Wang (2008) and Kato (2012) use the weighted Nadaraya-Watson estimator, which, for interior points, is asymptotically equivalent to the local linear estimator, but it yields a proper c.d.f. The weighted Nadaraya-Watson estimator, however, is not defined for boundary points.

Various ways of estimating truncated conditional expectation functions have also been proposed in parametric settings. In early work, Koenker and Bassett Jr (1978), Ruppert and Carroll (1980), and Jurečková (1984) consider generalizations of truncated means to linear models. In the first stage, they estimate quantile regressions, and in the second stage they run a regression on a sample truncated according to the first-stage estimates. Conceptually related to my paper is the work of Barendse (2020), who also runs a regression with a generated outcome variable based on the orthogonal moment. He additionally considers efficient weighting, analogous to, possibly nonlinear, weighted least squares. Dimitriadis et al. (2019) develop a joint quantile and expected shortfall estimation framework and find estimators that can be more efficient than the simple two-stage procedure described above. The efficiency gains of Dimitriadis et al. (2019) and Barendse (2020), however, are specific to parametric models, and they do not carry over to the nonparametric setting.

The cited papers—developed for the conditional expected shortfall estimation or robust estimation—assume that the truncation quantile level is chosen by the analyst. A setting with estimated conditional truncation quantile levels and possibly continuous covariates is studied by Semenova (2020). She exploits a moment that is similar to (1.1.3), but it includes additional terms, which render the expression orthogonal also to the truncation quantile level.[5] Her focus, however, is on integrated truncated conditional expectations, and she does not provide conditional estimates.[6] Estimated trimming proportions have also been studied in the unconditional setting, e.g., by Shorack et al. (1974) and Lee (2009).

Outline of the Paper. The remainder of this paper is structured as follows. In Section 1.2, I formally introduce the proposed estimator. I study its asymptotic properties in Section 1.3. In Section 1.4, I discuss inference, estimation with an estimated truncation quantile level, and related approaches. I present a Monte Carlo study in Section 1.5. In Section 1.6, I consider two empirical applications: (i) sharp regression discontinuity designs with a manipulated running variable and (ii) estimation of the conditional wage effect of a job training program. Section 1.7 concludes.

[5] This property is achieved using a specific conditional moment defining the truncation quantile level.

[6] Semenova (2020) considers a setting with many covariates, which requires regularization in the first step. I do not consider such aspects.

1.2. ESTIMATOR

In this section, I formally introduce my proposed estimator. To simplify the exposition, in the main text, I consider a univariate X. A natural extension for the multivariate case is presented in Appendix 1.A.1. Throughout the paper, I consider estimation of the truncated conditional expectation function at a selected covariate value x_0.

In the first stage, I estimate the conditional η-quantile function $Q(\eta, \cdot)$. For the second-stage estimator it suffices if $Q(\eta, \cdot)$ is estimated well for covariate values close to x_0. The level and slope of the function $Q(\eta, \cdot)$ at x_0 are estimated in a local linear quantile regression as

$$(\widehat{q}_0(\eta, x_0; a), \widehat{q}_1(\eta, x_0; a))^\top = \operatorname*{arg\,min}_{(\beta_0,\beta_1)} \sum_{i=1}^{n} k_a(X_i - x_0)\rho_\eta(Y_i - \beta_0 - \beta_1(X_i - x_0)),$$

where $\rho_\eta(v) = v(\eta - \mathbb{1}(v \leq 0))$ is the 'check' function, $k(\cdot)$ is a kernel function, a is a bandwidth, and $k_a(v) = k(v/a)/a$. Using these estimates, I estimate $Q(\eta, x)$ as

$$\widehat{Q}^{ll}(\eta, x; x_0, a) = \widehat{q}_0(\eta, x_0; a) + \widehat{q}_1(\eta, x_0; a)(x - x_0).$$

For a given η, $\widehat{Q}^{ll}(\eta, x; x_0, a)$ is a linear (random) function in x indexed by x_0 and a.

In the second stage, I run a local linear regression with $\psi_i(\eta, \widehat{Q}^{ll}(\eta, X_i; x_0, a))$ as the outcome variable, where

$$\psi_i(\eta, q) = \frac{1}{\eta}\left[Y_i \mathbb{1}(Y_i \leq q) - q(\mathbb{1}(Y_i \leq q) - \eta)\right].$$

The final estimator is given by

$$\widehat{m}(\eta, x_0; a, h) = e_1^\top \operatorname*{arg\,min}_{(\beta_0,\beta_1)} \sum_{i=1}^{n} k_h(X_i - x_0)\Big(\psi_i(\eta, \widehat{Q}^{ll}(\eta, X_i; x_0, a)) - \beta_0 - \beta_1(X_i - x_0)\Big)^2,$$

where h is another bandwidth, which can be different from the first-stage bandwidth.

1.3. ASYMPTOTIC PROPERTIES

In this section, I introduce the assumptions and study the asymptotic properties of the proposed estimator. I use the following notation. I put $\partial_x^k m(\eta, x_0) = \frac{\partial^k}{\partial x^k} m(\eta, x)|_{x=x_0}$ and $\partial_x^k Q(\eta, x_0) = \frac{\partial^k}{\partial x^k} Q(\eta, x)|_{x=x_0}$. For positive sequences b_n and c_n, I write $b_n \prec c_n$ if $b_n/c_n \to 0$, and $b_n \asymp c_n$ if $C_1 b_n \leq c_n \leq C_2 b_n$ for some positive constants C_1 and C_2.

1.3.1. **Assumptions.** As the canonical case, I consider estimation based on independent and identically distributed (i.i.d.) data. This modeling assumption is appropriate for microeconometric applications. The asymptotic analysis could be extended to allow for dependent data satisfying an α-mixing condition under restrictions on the rates of the mixing coefficients similarly Masry and Fan (1997).

Assumption 1.1. *(a)* $\{(X_i, Y_i)\}_{i=1}^{n}$ *are continuous i.i.d. random variables; (b)* $\eta \in (0, 1)$.

I follow the classic literature on local polynomial modeling methods and assume that the covariate is continuous. The density of X is denoted by $f_X(x)$, and its support is denoted by $\mathcal{X}$. The conditional distribution function of Y given X is denoted by $F_{Y|X}(y|x)$, and the corresponding conditional density by $f_{Y|X}(y|x)$.

Subsequent assumptions involve smoothness requirements for the functions $Q(\eta, \cdot\,)$ and $m(\eta, \cdot\,)$. I adopt the following convention. For a point on the left (right) boundary of $\mathcal{X}$, I define the derivative with respect to the covariate value as the right (left) derivative at that point.

Assumption 1.2. *(a)* $\partial_x^2 Q(\eta, x)$ *is continuous in* x *on* $\mathcal{X}$*; (b)* $\mathcal{X}$ *is a bounded interval and* $f_X(x)$ *is continuous and positive on* $\mathcal{X}$*; (c)* $f_{Y|X}(y|x)$ *is continuous in* x *and* y *on* $\{(x, y) : x \in \mathcal{X},\, y \in [Q(\eta, x) \pm \epsilon]\}$ *for some* $\epsilon > 0$*. Moreover,* $f_{Y|X}(Q(\eta, x)|x) > 0$.

Assumption 1.2 comprises standard conditions for the asymptotic analysis of the local linear quantile estimator. A continuous second-order derivative of $Q(\eta, x)$ with respect to x is required to control the bias introduced by approximating the possibly nonlinear function $Q(\eta, \cdot\,)$ with its first-order Taylor expansion. The restrictions on the density $f_X(x)$ ensure that there are observations around the estimation point. The restrictions on the conditional density $f_{Y|X}(y|x)$ ensure that the conditional η-quantile function can be precisely estimated.

Assumption 1.3. *(a)* $\partial_x^2 m(\eta, x)$ *is continuous in* x *on* $\mathcal{X}$*; (b)* $Var(Y|X = x, Y \leq Q(\eta, x))$ *is finite, positive, and continuous in* x *on* $\mathcal{X}$*; (c)* $\mathbb{E}[|Y|^{2+\xi}\mathbb{1}(Y \leq Q(\eta, X))|X = x]$ *is bounded uniformly over* x *in* $\mathcal{X}$ *for some* $\xi > 0$.

Assumption 1.3 is a natural adaptation of the standard conditions for the local linear estimator in the nonparametric mean regression to the problem of estimating truncated conditional expectation functions. Even if the function $Q(\eta, \cdot\,)$ was known, a continuous second-order derivative of $m(\eta, x)$ with respect to x would be required to control the bias introduced by approximating the function $m(\eta, \cdot\,)$ with its first-order Taylor expansion. Parts (b) and (c) are needed to obtain asymptotic normality.

Assumption 1.4. *(a) The kernel* k *is a bounded and symmetric density function with compact support, say* $[-1, 1]$*; (b) As* $n \to \infty$, $h \to 0$, $a \to 0$, $nh \to \infty$, *and* $na \to \infty$.

The restrictions on the kernel are standard. The requirements on the bandwidths are necessary for ensuring consistency.

1.3.2. **Asymptotic Distribution.** In this section, I analyze the asymptotic properties of my estimator. The key result is that the feasible estimator $\widehat{m}$ is asymptotically equivalent to the oracle estimator employing the true function $Q(\eta, \cdot)$, which is given by

$$\widetilde{m}(\eta, x_0; h) = e_1^\top \operatorname*{arg\,min}_{(\beta_0,\beta_1)} \sum_{i=1}^n k_h(X_i - x_0)(\psi_i(\eta, Q(\eta, X_i)) - \beta_0 - \beta_1(X_i - x_0))^2.$$

This asymptotic equivalence result is stated in Theorem 1.1.

Theorem 1.1. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then*

$$R(\eta, x_0; a, h) \equiv \widehat{m}(\eta, x_0; a, h) - \widetilde{m}(\eta, x_0; h) = O_p(w_n(nh)^{-1/2} + w_n^2),$$

where $w_n = a^2 + h^2 + (a+h)(a^3 n)^{-1/2}$. *In particular, if* $a \asymp h$, *then* $R(\eta, x_0; a, h) = O_p(h^4 + (nh)^{-1})$.

The remainder $R(\eta, x_0; a, h)$ is driven by the estimation error from the first stage on the interval $\mathcal{X}(x_0, h) \equiv [x_0 - h, x_0 + h] \cap \mathcal{X}$, which is relevant for the second-stage estimator. There are two sources of this estimation error. First, the function $Q(\eta, \cdot)$ is replaced with its local linear approximation, which results in an error of order $O(h^2)$. Second, the intercept and slope of this approximation are estimated at rates $O_p(a^2 + (an)^{-1/2})$ and $O_p(a + (a^3 n)^{-1/2})$, respectively.[7] As a result, the estimated conditional quantile function satisfies

$$\sup_{x \in \mathcal{X}(x_0,h)} |\widehat{Q}^{ll}(\eta, x; x_0, a) - Q(\eta, x)| = O_p(w_n). \tag{1.3.1}$$

If $h(nh)^{-1/3} \prec a$, then $w_n \to 0$, and $R(\eta, x_0; a, h)$ is of order smaller than $O_p(w_n)$. This low sensitivity to the first-stage estimation error is obtained by construction, owing to the use of an orthogonal moment.

Theorem 1.1 holds regardless of whether the variance of the outcome variable is finite or infinite. If Assumption 1.3 holds in addition to the assumptions of Theorem 1.1, then the asymptotic normal distribution follows from the standard theory of local linear estimation (e.g. Masry and Fan, 1997). If the variance of the outcome variable is infinite, then the asymptotic distribution can be obtained under alternative assumptions following the steps of Linton and Xiao (2013). I focus on the former case.

The asymptotic distribution is presented in Corollary 1.1. It involves typical kernel constants, which differ depending on whether x_0 lies in the interior or on the boundary of the support of X, but I leave this dependence implicit in the notation. If x_0 lies in the interior of $\mathcal{X}$, I define $\mu = \int v^2 k(v) dv$ and $\kappa = \int k(v)^2 dv$. If x_0 lies on the boundary of $\mathcal{X}$, I define $\mu = (\bar{\mu}_2^2 - \bar{\mu}_1\bar{\mu}_3)/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2)$ and $\kappa = \int_0^\infty (k(v)(\bar{\mu}_1 v - \bar{\mu}_2))^2 dv/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2)^2$, where $\bar{\mu}_j = \int_0^\infty v^j k(v) dv$.

[7] In fact, these are the only properties of the first-stage estimator required in the proof of Theorem 1.1.

Corollary 1.1. *Suppose that Assumptions 1.1–1.4 hold, and $h(nh)^{-1/6} \prec a \prec \sqrt{h}$, e.g., $a = h$. Then*

$$\sqrt{nh}\Big(\widehat{m}(\eta, x_0; a, h) - m(\eta, x_0) - \mathcal{B}(\eta, x_0, h)\Big) \xrightarrow{d} \mathcal{N}(0, V(\eta, x_0)),$$

where

$$\begin{aligned} \mathcal{B}(\eta, x_0, h) &= \frac{1}{2}\mu\, \partial_x^2 m(\eta, x_0) h^2 + o_p(h^2), \\ V(\eta, x_0) &= \frac{\kappa}{\eta f_X(x_0)}\Big(Var(Y|Y \leq Q(\eta, x_0), X = x_0) + (1-\eta)\left(Q(\eta, x_0) - m(\eta, x_0)\right)^2\Big). \end{aligned}$$

The additional conditions imposed on the bandwidths ensure that the remainder $R(\eta, x_0; a, h)$ is of order $o_p(h^2 + (nh)^{-1/2})$, and as such, it does not affect the first-order asymptotic distribution of $\widehat{m}$. These conditions admit certain degrees of both under- and oversmoothing in the first stage relative to the second stage. For example, if $h \asymp n^{-1/5}$, then I require that $n^{-1/3} \prec a \prec n^{-1/10}$. Subject to these restrictions, the choice of the first-stage bandwidth does not affect the first-order asymptotic distribution. In practice, the two bandwidths can be set equal.

As in the standard nonparametric regression, the leading bias is proportional to the second derivative of the function that is being estimated. The variance is fully analogous to the variance of the unconditional truncated mean.

1.4. DISCUSSION

In this section, I discuss statistical inference based on the asymptotic result in Corollary 1.1, estimation with an estimated quantile level, and related approaches.

1.4.1. **Inference.** The asymptotic distribution obtained in Corollary 1.1 forms the basis for conducting statistical inference. As in the standard nonparametric regression, constructing a confidence interval (CI) requires estimating the variance and accounting for the bias. The asymptotic variance $V(\eta, x_0)$ can be consistently estimated using the Eicker-Huber-White (EHW) estimator based on the residuals from the second stage. Let $\widehat{se}(h)$ denote the resulting estimate of the standard error. The asymptotic bias can be handled in any of the three following ways adapted from the nonparametric regression literature.

The first, classic approach is called undersmoothing (US). It relies on choosing a 'small' bandwidth, which ensures that the bias is negligible. If $h \prec n^{-1/5}$, or equivalently $nh^5 \to 0$, then the bias is of smaller order than the standard error. As a result, an asymptotically valid $1-\alpha$ CI can be formed as

$$CI_\alpha^{US} = [\widehat{m}(\eta, x_0; h, h) \pm z_{1-\alpha/2} \cdot \widehat{se}(h)], \tag{1.4.1}$$

where z_u is the u-quantile of the standard normal distribution. The two further approaches allow for bandwidths of order $n^{-1/5}$. This case is relevant as it covers, i.a., the bandwidth optimal in terms of the asymptotic mean squared error.

The second approach is analogous to the robust bias corrections proposed by Calonico et al. (2014). It involves subtracting an estimate of the leading bias term and accounting for the additional variation in the bias-corrected estimator when forming a CI. The bias correction term can be constructed using the estimator of $\partial_x^2 m(\eta, x_0)$ proposed in Section 1.A.2. The CI takes the form as in (1.4.1), except that a bias-corrected estimator and an adjusted standard error are used.

The third approach follows Armstrong and Kolesár (2020), who propose 'honest' CIs that account for the largest possible bias under restrictions on the smoothness of the function that is being estimated. Suppose that $|\partial_x^2 m(\eta, x_0)|$ is bounded by some known constant M. Then the leading bias term is bounded in absolute value by $\frac{1}{2}|\mu| M h^2$. It follows from Armstrong and Kolesár (2020) that an asymptotically valid $1-\alpha$ confidence interval can be formed as

$$CI_\alpha = [\widehat{m}(\eta, x_0; h, h) \pm \mathrm{cv}_{1-\alpha}(\widehat{r}(h)) \cdot \widehat{se}(h)], \tag{1.4.2}$$

where $\widehat{r}(h) = \frac{1}{2}|\mu| M h^2 / \widehat{se}(h)$ and $\mathrm{cv}_{1-\alpha}(t)$ is the $1-\alpha$ quantile of the folded normal distribution $|\mathcal{N}(t, 1)|$.[8] One can also account for the maximal bias of the oracle estimator conditional on the realizations of the covariate. The bandwidth can be chosen so as to minimize the worst-case mean squared error or the length of the CI. Implementation of bandwidth selectors and of the CIs requires imposing a bound on $\partial_x^2 m(\eta, x_0)$. See Armstrong and Kolesár (2020) and Noack and Rothe (2021) for discussions of the choice of the smoothness constant in the standard nonparametric regression.

1.4.2. **Estimated Truncation Quantile Level.** In some applications, the truncation quantile level of interest has to be estimated from the data. In this section, I study the properties of my estimator evaluated at an estimated truncation quantile level. Specifically, under a high-level assumption on the estimator $\widehat{\eta}$ of η, I provide an expansion of the estimator $\widehat{m}(\widehat{\eta}, x_0)$ about the estimator $\widehat{m}(\eta, x_0)$. This result can be used on a case-by-case basis to derive the asymptotic distribution of $\widehat{m}(\widehat{\eta}, x_0)$ for specific estimators $\widehat{\eta}$. I analyze two such examples in Section 1.6.

To keep the exposition transparent, I restrict the analysis to bandwidths such that $a \asymp h$. In comparison to Theorem 1.1, I impose two further assumptions. First, I require that the estimator $\widehat{\eta}$ converges at a rate not slower than the estimator $\widehat{m}(\eta, x_0; a, h)$ does.

[8] I do not discuss coverage properties uniform in the data generating processes, which would require ensuring that the remainder in Theorem 1.1 is uniformly small.

Assumption 1.5. *There exists a deterministic sequence η_n such that $\eta_n - \eta = O(h^2)$ and $\widehat{\eta} - \eta_n = O_p\Big((nh)^{-1/2}\Big)$.*

Second, I slightly strengthen Assumption 1.2(a), which is needed to control the bias of the first-stage local linear quantile estimator for quantile levels close to η.

Assumption 1.6. *$\partial_x^2 Q(u,x)$ is continuous in u and x on $[\eta-\epsilon, \eta+\epsilon] \times \mathcal{X}$ for some $\epsilon > 0$.*

Theorem 1.2 provides an expansion of the estimator with an estimated truncation quantile level about the estimator using the true quantile level.

Theorem 1.2. *Suppose that Assumptions 1.1–1.6 hold and $a \asymp h$. Then*

$$\widehat{m}(\widehat{\eta}, x_0; a, h) = \widetilde{m}(\eta, x_0; h) + C(\eta, x_0)(\widehat{\eta} - \eta) + O_p(h^4 + (nh)^{-1}),$$

where $C(\eta, x_0) = \partial_\eta m(\eta, x_0) = \frac{1}{\eta}(Q(\eta, x_0) - m(\eta, x_0))$.

The coefficient on $(\widehat{\eta} - \eta)$ in the above expansion is equal to the derivative of $m(\eta, x_0)$ with respect to the truncation quantile level, which is in line with Lemma 1 of Shorack et al. (1974) and Proposition 3 of Lee (2009), who study the unconditional truncated mean with random trimming proportions. In Theorem, 1.2 it is essential that $\eta < 1$, assumed in Assumption 1.1(b). Otherwise, if Y has unbounded support, the derivative $\partial_\eta m(\eta, x_0)$ is infinite, and the expansion in Theorem 1.2 is not valid.

1.4.3. **Related Approaches.** Local linear methods can be used to construct two further estimators, which have not been formally studied in the literature so far. I discuss them briefly in this section, and I provide a detailed asymptotic analysis in Appendix 1.B. I argue that the first one has an undesirable property in that it is not translation invariant. The second one has good asymptotic properties only in one special case, when the same bandwidth is used in both stages.

The non-orthogonal conditional moment (NM) in (1.1.2) motivates running a local linear regression without the second term included in the generated outcome variable based on the orthogonal moment. Let

$$\begin{aligned}&\widehat{m}^{NM}(\eta, x_0; a, h)\\ &\quad = e_1^\top \operatorname*{arg\,min}_{(\beta_0, \beta_1)} \sum_{i=1}^n k_h(X_i - x_0) \left(\frac{1}{\eta} Y_i \mathbb{1}(Y_i \leq \widehat{Q}^{ll}(\eta, X_i; x_0, a)) - \beta_0 - \beta_1(X_i - x_0)\right)^2.\end{aligned}$$

Under assumptions, this estimator is consistent and asymptotically normal. However, it has one unappealing property—it is not translation invariant. Adding a constant to all outcomes and subtracting it from the result can yield a different estimate than applying the estimator to the original data.[9] The estimator $\widehat{m}$ is free of this deficiency.

Another estimator, motivated by the definition of the estimand in (1.1.1), can be obtained by running a local linear regression on a truncated sample (TS) restricted to observations that fall below the estimated conditional η-quantile function.[10] Let

$$\begin{aligned}&\widehat{m}^{TS}(\eta, x_0; a, h)\\&\quad = e_1^\top \operatorname*{arg\,min}_{(\beta_0,\beta_1)} \sum_{i=1}^{n} k_h(X_i - x_0)\,(Y_i - \beta_0 - \beta_1(X_i - x_0))^2\,\mathbb{1}(Y_i \leq \widehat{Q}^{ll}(\eta, X_i; x_0, a)).\end{aligned}$$

This estimator is translation invariant. Unlike in the case of $\widehat{m}$, the asymptotic distribution of $\widehat{m}^{TS}$ explicitly depends on the first-stage bandwidth, and in general it involves more complicated bias and variance formulas than those in Corollary 1.1. Only in the special case when the bandwidths in both stages are equal, is $\widehat{m}^{TS}$ asymptotically equivalent to the oracle estimator $\widetilde{m}$, and hence it has the asymptotic distribution given in Corollary 1.1. However, for boundary points, the remainder in the Bahadur representation of $\widehat{m}^{TS}(\eta, x_0; h, h)$ is in general of larger order than $O_p(h^4 + (nh)^{-1})$ obtained in Theorem 1.1 for bandwidths converging at the same rates.

The estimator based on the truncated sample with equal bandwidths corresponds most closely to the unconditional truncated mean, where the same (full) sample is used to first estimate the quantile and then to calculate the truncated mean. However, I advocate using the estimator $\widehat{m}$, as it makes the parallel between estimation of conditional expectation functions and truncated conditional expectation functions explicit.[11] The very small remainder in Theorem 1.1 provides a strong theoretical justification for conducting inference as if the oracle estimator was available.

I remark that the two-stage procedure yielding $\widehat{m}^{TS}$ with equal bandwidths provides an intuitive decomposition of the asymptotic variance $V(\eta, x_0)$ defined in Corollary 1.1. The asymptotic variance of the infeasible local linear estimator using observations with $Y_i \leq Q(\eta, X_i)$ equals $\frac{\kappa}{\eta f_X(x_0)}\mathrm{Var}(Y|Y \leq Q(\eta, x_0), X = x_0)$, which is the first component of $V(\eta, x_0)$. The second, strictly positive, component of $V(\eta, x_0)$ is due to the first-step estimation.[12]

[9] This difference is asymptotically very small in the case when the same bandwidth is used in both stages, but even then, the estimator is not numerically translation invariant.

[10] This approach has been proposed in a working paper by Gerard et al. (2016), but they do not derive its asymptotic distribution.

[11] Standard inference methods cannot be simply applied to the truncated sample.

[12] An analogous decomposition holds for the unconditional truncated mean. A similar point is also made by Dimitriadis et al. (2019, Remark 2.9) in a parametric model.

1.5. MONTE CARLO STUDY

In this section, I present simulation evidence for two claims. First, I show that the feasible estimator $\widehat{m}$ is close to the oracle estimator $\widetilde{m}$ in terms of the mean squared difference. Second, I show that inference based on $\widehat{m}$ performs almost identically as inference based on the oracle estimator $\widetilde{m}$. In this simulation study, I use the third approach discussed in Section 1.4.1, which exploits a bound on $\partial_x^2 m(\eta, x)$.[13] The qualitative conclusions about the very similar performance of the feasible and oracle estimators are the same for undersmoothing and robust bias corrections.

I generate data from a location-scale model of the form

$$Y = m(X) + sd(X)\varepsilon, \tag{1.5.1}$$

where X is uniformly distributed on $[-1, 1]$ and $\varepsilon \sim \mathcal{N}(0, 1)$. I consider three specifications for the conditional expectation function, which were used by Armstrong and Kolesár (2020) in their Monte Carlo study comparing different inference methods. Let

$$\begin{aligned} m_1(x) &= x^2 - 2s(|x| - 0.25), \\ m_2(x) &= x^2 - 2s(|x| - 0.2) + 2s(|x| - 0.5) - 2s(|x| - 0.65), \\ m_3(x) &= (x + 1)^2 - 2s(x + 0.2) + 2s(x - 0.2) - 2s(x - 0.4) + 2s(x - 0.7) - 0.92, \end{aligned}$$

where $s(x) = \max\{x, 0\}^2$ is the square of the plus function. These functions are depicted in Figure 1.1. Their second derivatives are bounded in absolute value by $M = 2$. I consider homoskedastic and hetersokedastic residuals, induced by functions $sd_1(x) = 0.5$ and $sd_2(x) = 0.5 + x$, respectively.

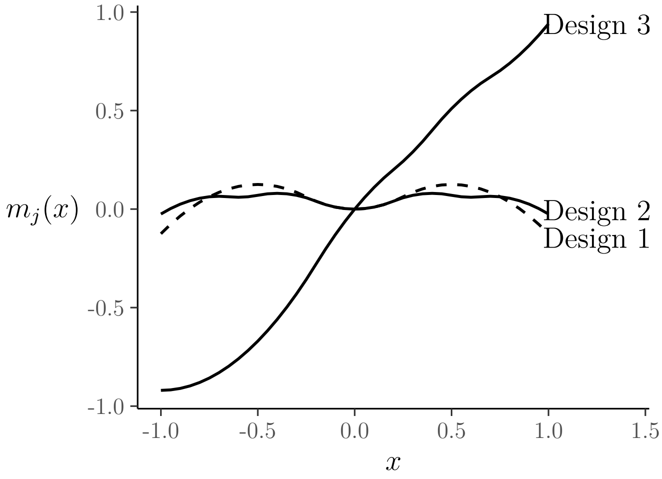


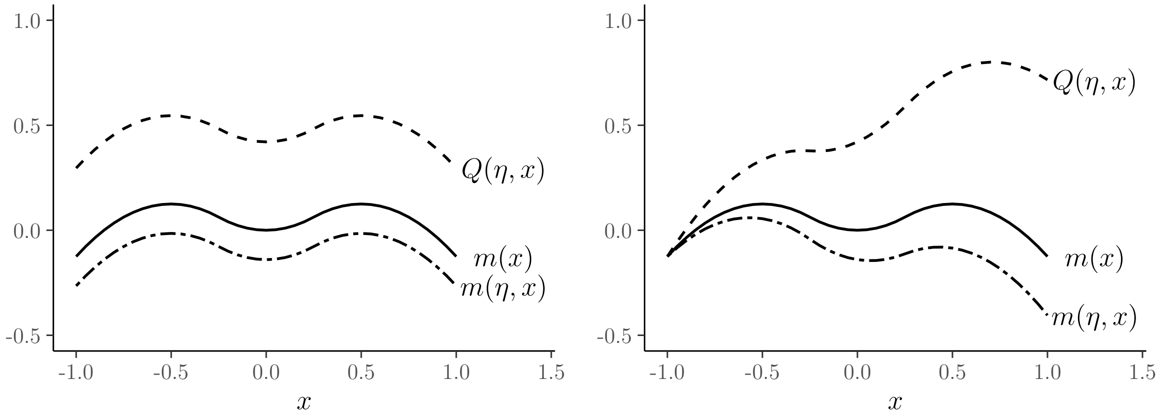
Figure 1.1: Conditional expectation functions $m_j(x)$.

[13] In simulations, I account for the exact worst-case bias of the oracle estimator conditional on the realizations of the covariate, rather than only for the leading bias term.

Due to normality of the residuals, the truncated conditional expectation functions have a simple, closed-form expression. It holds that

$$m(\eta, x) = m(x) - \frac{\phi(q_\eta)}{\eta} sd(x), \tag{1.5.2}$$

where $\phi(\cdot)$ is the density and q_η is the η-quantile of the standard normal distribution, respectively. With homoskedastic residuals, the truncated conditional expectation functions have the same shape as the respective conditional expectation functions, but they are shifted downwards. With heteroskedastic residuals, the slopes change as well, but this type of heteroskedasticity does not affect the curvature. Figure 1.2 illustrates that for $\eta = 0.8$ and $m(x) = m_1(x)$. Other cases are analogous.



(a) Homoskedastic case, $sd(x) = 0.5$. (b) Hetersokedastic case, $sd(x) = 0.5 \cdot (1 + x)$.

Figure 1.2: Truncated conditional expectation functions for $m(x) = m_1(x)$ and $\eta = 0.8$.

In all simulations, the sample size is $n = 1,000$, and the number of replications is $S = 10,000$. I estimate truncated conditional expectation functions for $x_0 = 0$ and three quantile levels, $\eta \in \{0.2, 0.5, 0.8\}$. I use the triangular kernel and the EHW variance estimator.

In Table 1.1, I report the root mean squared error (RMSE) of the oracle estimator $\widetilde{m}$ and the feasible estimator $\widehat{m}$, as well as the root mean squared error difference between the two. The estimators are evaluated with the RMSE-optimal bandwidth chosen for the oracle estimator using the bandwidth selector of Armstrong and Kolesár (2020) employing the true smoothness constant ($M = 2$). In all cases, the difference between the oracle and feasible estimators is small compared to their mean squared errors.[14] Moreover, the results are very similar in the homoskedastic and heteroskedastic settings, which shows that the estimator adapts to different slopes of the conditional quantile and truncated expectation functions very well.

[14] This qualitative conclusion remains the same when using the true bound multiplied or divided by two.

Table 1.1: RMSE and root mean squared distance to the oracle.

		RMSE			Dist. to the oracle		
	Design for m_j:	1	2	3	1	2	3
Homoskedastic errors							
$\eta = 0.2$	Oracle $\widetilde{m}$	5.044	5.002	5.146	-	-	-
	Feasible $\widehat{m}$	5.273	5.222	4.965	0.563	0.569	0.575
$\eta = 0.5$	Oracle $\widetilde{m}$	4.094	4.068	4.134	-	-	-
	Feasible $\widehat{m}$	4.202	4.174	4.041	0.277	0.280	0.282
$\eta = 0.8$	Oracle $\widetilde{m}$	3.742	3.721	3.759	-	-	-
	Feasible $\widehat{m}$	3.804	3.782	3.707	0.164	0.165	0.166
Heteroskedastic errors							
$\eta = 0.2$	Oracle $\widetilde{m}$	5.095	5.032	5.177	-	-	-
	Feasible $\widehat{m}$	5.306	5.236	5.006	0.548	0.551	0.556
$\eta = 0.5$	Oracle $\widetilde{m}$	4.126	4.091	4.157	-	-	-
	Feasible $\widehat{m}$	4.230	4.192	4.070	0.271	0.271	0.273
$\eta = 0.8$	Oracle $\widetilde{m}$	3.766	3.742	3.782	-	-	-
	Feasible $\widehat{m}$	3.825	3.800	3.731	0.161	0.160	0.161

Notes: All values are multiplied by 100. The estimators are evaluated with the RMSE-optimal bandwidth for the oracle estimator based on the true smoothness constant. The sample size is $n = 1,000$, and the number of simulations is $S = 10,000$.

In Table 1.2, I present results regarding the bandwidth choice as well as empirical coverage and length of 95% confidence intervals. Here, I also use the true smoothness constant ($M = 2$). The bandwidth selector for the feasible estimator chooses virtually the same bandwidth as would be chosen for the oracle estimator, and the coverage is nearly identical. I note that even for the oracle estimator, the CI based on the true smoothness constant can have coverage below the nominal confidence level despite correctly accounting for maximal bias. The reason for that is that although Y is conditionally normally distributed, the outcome variable $\psi(\eta, Q(\eta, X))$ is not. The non-normality is more severe for lower truncation quantile levels. In Appendix 1.D, I discuss a rule of thumb for choosing the smoothness constant, which performs well in this simulation setting.

1.6. APPLICATIONS

I discuss two empirical settings in which my estimator can be applied: (i) sharp regression discontinuity designs with a manipulated running variable and (ii) program evaluation under sample selection. They involve estimated truncation quantile levels.

1.6.1. **Sharp Regression Discontinuity Designs with Manipulation.**

Gerard et al. (2020) study regression discontinuity (RD) designs with a manipulated running variable. They develop a complex estimation approach applicable to fuzzy RD designs, which

Table 1.2: Coverage, average bandwidth, and average length of the 95% CI.

		Coverage			Bandwidth			CI length		
	Design for m_j:	1	2	3	1	2	3	1	2	3
Homoskedastic errors										
$\eta = 0.2$	Oracle $\widetilde{m}$	92.1	92.4	96.1	0.373	0.372	0.369	0.099	0.099	0.099
	Feasible $\widehat{m}$	92.1	92.3	96.1	0.366	0.368	0.374	0.100	0.100	0.098
$\eta = 0.5$	Oracle $\widetilde{m}$	93.5	93.7	96.0	0.334	0.334	0.333	0.080	0.080	0.080
	Feasible $\widehat{m}$	93.6	93.8	95.9	0.331	0.332	0.335	0.081	0.081	0.080
$\eta = 0.8$	Oracle $\widetilde{m}$	94.4	94.6	95.7	0.319	0.319	0.318	0.073	0.073	0.073
	Feasible $\widehat{m}$	94.4	94.5	95.9	0.318	0.318	0.320	0.074	0.074	0.073
Heteroskedastic errors										
$\eta = 0.2$	Oracle $\widetilde{m}$	92.1	92.7	96.3	0.382	0.384	0.379	0.100	0.100	0.100
	Feasible $\widehat{m}$	92.5	93.0	96.1	0.375	0.380	0.385	0.101	0.101	0.099
$\eta = 0.5$	Oracle $\widetilde{m}$	93.4	93.8	96.2	0.341	0.344	0.341	0.081	0.081	0.081
	Feasible $\widehat{m}$	93.6	94.0	96.0	0.337	0.342	0.344	0.081	0.081	0.080
$\eta = 0.8$	Oracle $\widetilde{m}$	94.4	94.6	95.8	0.325	0.328	0.326	0.074	0.074	0.074
	Feasible $\widehat{m}$	94.4	94.6	95.8	0.323	0.327	0.328	0.074	0.074	0.074

Notes: The estimators are evaluated with their respective RMSE-optimal bandwidths based on the true smoothness constant. The sample size is $n = 1,000$, and the number of simulations is $S = 10,000$.

encompass sharp RD designs as a special case. Their inference is based on a bootstrap procedure. I study a simpler approach tailored specifically to sharp RD designs, which allows me to derive the asymptotic distribution of the estimator of the bounds.

Partial Identification under Manipulation. In a sharp RD design, the treatment is assigned and taken up if and only if a special covariate, the running variable, exceeds a fixed cutoff value. If the distribution of units' potential outcomes varies smoothly with the running variable around the cutoff, then the (local to the cutoff) average treatment effect is identified by the difference in average outcomes of the treated and untreated units whose realization of the running variable is just to the right or just to the left of the cutoff, respectively. The key identifying assumption, however, is often questionable if the running variable is not exogenously determined.

To allow for violations of the smoothness assumption, Gerard et al. (2020) develop a framework where there are two unobservable types of units: *always-assigned* units, for which the realization of the running variable is always to the right of the cutoff, and hence they are assigned the treatment; and *potentially-assigned* units, whose density of the running variable is smooth around the cutoff, and hence they satisfy the standard assumptions of an RD design. Gerard et al. (2020) show that the average treatment effect for the subpopulation of potentially-assigned units at the cutoff, denoted by Γ, is partially identified. Under their behavioral model, the share of always-assigned units just to the right of the cutoff, denoted by τ, is identified by the discontinuity in the density of the running variable at the cutoff as

$$\tau = 1 - \frac{f(x_0^-)}{f(x_0^+)},$$

where x_0 is the cutoff value.[15] Given τ, the sharp bounds on Γ are obtained by considering the 'extreme' scenarios in which the always-assigned units constitute the proportion τ of the units with the lowest or the highest outcomes among the treated. The resulting lower and upper bound are given by:

$$\begin{aligned}\Gamma^L &= \mathbb{E}[Y|X = x_0^+, Y \leq Q(1-\tau, x_0^+)] - \mathbb{E}[Y|X = x_0^-],\\ \Gamma^U &= \mathbb{E}[Y|X = x_0^+, Y \geq Q(\tau, x_0^+)] - \mathbb{E}[Y|X = x_0^-].\end{aligned}$$

Estimation and Inference. I discuss the main ingredients of the bounds estimator and its asymptotic properties. The details are given in Appendix 1.C.1. The bounds Γ^L and Γ^U involve truncated conditional expectation functions, which I estimate using the estimator $\widehat{m}$ developed in this paper.[16] Since τ is the proportion of truncated data, the quantile level η in the previous sections corresponds to $1-\tau$, i.e. η is the proportion of potentially-assigned units just to the right of the cutoff. The first step is to estimate τ. The density limits can be estimated using estimators such as the linear smoother of the histogram (Cheng, 1997; McCrary, 2008), the linear smoother of the empirical density function (Jones, 1993; Lejeune and Sarda, 1992), or the local quadratic smoother of the empirical distribution function of (Cattaneo et al., 2020).

Under regularity conditions, the resulting estimator of the truncation quantile level, $\widehat{\eta} = 1 - \widehat{\tau}$, satisfies the high-level assumption of Theorem 1.2. Moreover, since $\widehat{\eta}$ depends only on the running variable, it is conditionally uncorrelated with the estimators of the truncated conditional expectations with known η, which simplifies the asymptotic variance formula. The conditional expectation just to the left of the cutoff, $\mathbb{E}[Y|X = x_0^-]$, can be estimated using a standard local linear estimator. The estimators of the bounds have an asymptotically normal distribution, which can be used to form confidence intervals.

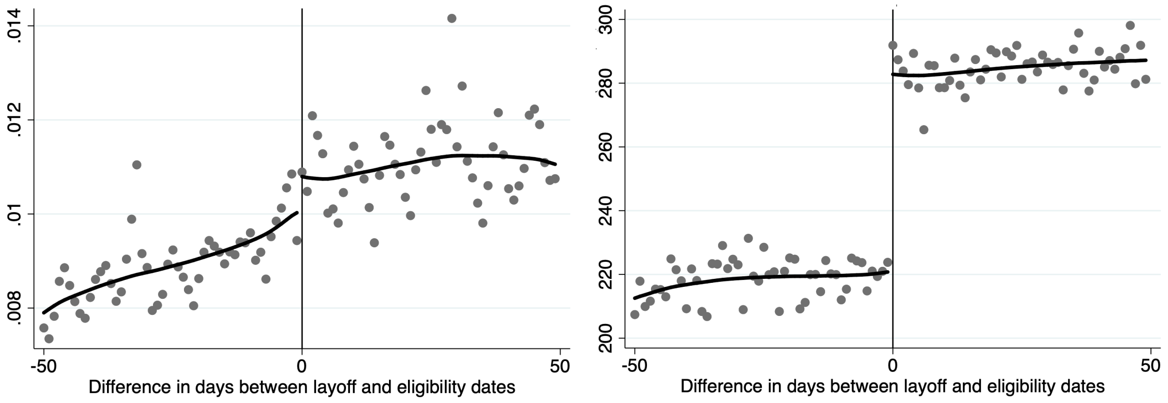
Empirical Application. I evaluate the procedure that I propose by implementing it for the empirical application of Gerard et al. (2020).[17] They investigate the effect of unemployment insurance (UI) benefits on the formal reemployment in Brazil. They exploit the rule that a worker involuntarily laid off from a private-sector firm is eligible for the UI benefit only if there was at least 16 months between the date of her layoff and the date of the last

[15] For a generic function $g(\cdot)$, I put $g(x_0^+) = \lim_{x \to x_0^+} g(x)$ and $g(x_0^-) = \lim_{x \to x_0^-} g(x)$.

[16] Estimation with truncation from below can be performed using the procedure developed for estimation with truncation from above by taking the negative of the estimator applied to the data $\{X_i, -Y_i\}_{i=1}^n$.

[17] The authors kindly implemented my procedure on their restricted-use data for comparison purposes.

layoff after which she applied for and drew UI benefits. This rule creates a discontinuity in the eligibility for UI benefits, which is reflected in a 70pp increase in the actual take-up of UI benefits. In the following, I focus on an intention-to-treat analysis, where the eligibility for UI benefits is the treatment, and the outcome of interest is the duration without a formal job after the layoff.



(a) Frequency.

(b) Duration without a formal job.

Figure 1.3: Graphical evidence for the intention-to-treat analysis.
Notes: The dots represent the frequency (left panel) and the average duration of unemployment censored at 24 months (right panel) by day. The figure is based on 169,575 observations. *Source: Gerard et al. (2020).*

Despite the 16-month rule being rather arbitrary, Gerard et al. (2020) point out the following ways in which violations of the standard RD assumptions may arise in this setup. Some workers may provoke their layoffs or ask their employers to report their quit as involuntary once they become eligible for a UI benefit. Other workers may have managed to delay their layoff to a date when they were eligible for the UI benefit. All theses workers are always-assigned units in the manipulation framework outlined in the previous subsection.

Figure 1.3 reproduces the graphical evidence for this RD design. The running variable is the difference in days between the layoff date and the eligibility date, so that the cutoff is at 0. In the left panel, I present the density of the running variable. The share of always-assigned units is estimated to be 6.4%, which is relatively well separated from zero. This is essential for the good quality of the normal approximation of the asymptotic distribution of $\hat{\tau}$. In the right panel, the dots represent the average outcome by day (of all observations). There is a marked jump in the mean duration without a formal job at the cutoff. I note that a substantial share, about 12–14%, of duration outcomes is censored at 24 months. This, however, does not require any adjustment in my estimation and inference procedure.

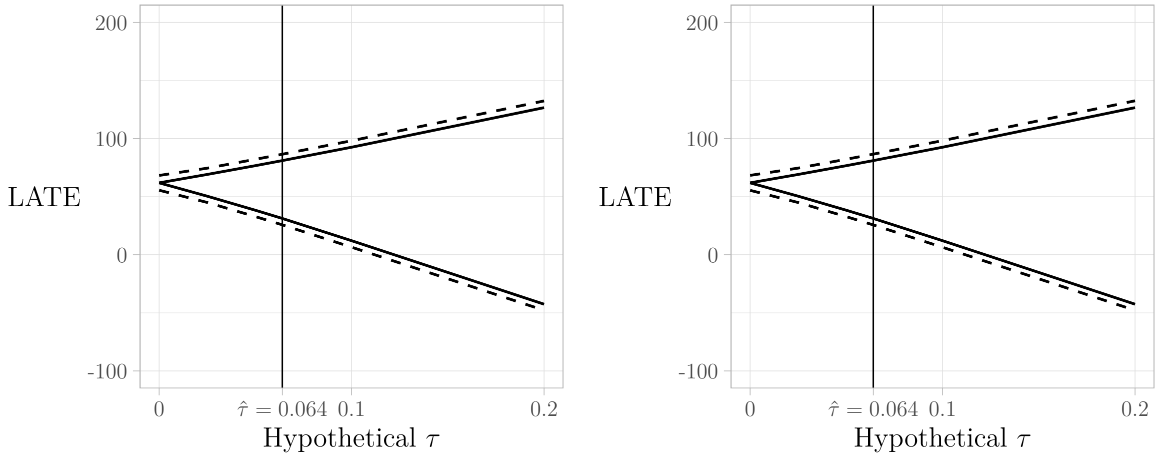
Following Gerard et al. (2020), I conduct two types of analysis. First, I estimate bounds on Γ using an estimated proportion of the always-assigned units to the right of the cutoff. Second, I conduct a sensitivity analysis, where I report bounds for different levels of potential manipulation. I report my results along with the original estimates of Gerard et al. (2020). Their estimator is based on a local linear estimator of the conditional c.d.f., and they conduct inference via bootstrap. All estimators use a 30-day bandwidth, and the confidence intervals are formally justified by undersmoothing.

In Table 1.3, I present estimates of the bounds and the 95% confidence intervals for Γ with estimated τ. As a reference point, the point estimate ignoring the possibility of manipulation indicates that the eligibility for UI benefits increases the duration of unemployment by about 62 days. When accounting for manipulation, however, the estimated identified set spans the range from 31 to 81 days. In the second part of the analysis, I do inference presuming a certain hypothetical, fixed degree of manipulation in the data. The results are presented in Figure 1.4. The vertical black line marks the estimated proportion of always-assigned units just to the right of the cutoff.

Table 1.3: Estimated effects of UI benefits on the duration without a formal job in days.

	Results of Gerard et al. (2020)		My results	
	Estimate	95% CI	Estimate	95% CI
Share of always-assigned units	0.064	[0.038; 0.089]		
LATE: Ignoring manipulation	61.9	[55.7; 68.1]	61.9	[55.5; 68.3]
LATE: Bounds for Γ	[31.4; 80.9]	[18.9; 89.6]	[31.4; 80.9]	[19.4; 89.5]

Note: There are 102,791 observations in the 30-day estimation window.



(a) Procedure of Gerard et al. (2020).

(b) Estimation with $\widehat{m}$.

Figure 1.4: Fixed-manipulation inference.

Notes: The horizontal axis displays the hypothetical proportion of potentially-assigned workers. The solid lines present the estimates of the bounds and the dashed lines mark 95% confidence intervals. The figures are based on 102,791 observations.

The results are nearly identical when using the procedure of Gerard et al. (2020) and mine. This similarity, however, is specific to this dataset, where the conditional quantile functions at the truncation quantile levels are flat. I show in Appendix 1.B.3 that compared to my estimator, approaches based on first-stage estimates of the conditional c.d.f. have an additional bias term when the conditional quantile function has a nonzero slope.

1.6.2. **Conditional Lee Bounds.** Lee (2009) studies the effect of a job training program on wage rates. In this analysis, he uses conditional estimates to narrow down the bounds on the unconditional effect (see also Semenova, 2020). The conditional treatment effects, however, may be of interest in their own right.

Partial Identification of the Wage Effect. Evaluation of the wage effect of a job training program is complicated by the fact that job training affects not only the wage rates but also the employment status. As a result, individuals in the treatment and control groups are not comparable conditional on being employed even if the treatment was random assigned. Lee (2009) derives bounds on the wage effect for the subpopulation of *always-observed* individuals, i.e. those who would work regardless of whether they obtained the treatment. In the first step, he identifies the proportion of individuals whose employment status is affected by the treatment status. By random assignment to the program, this proportion is given by the difference in the employment rates in the treatment and control group. If the training program weakly encourages to work, then the bounds on the wage rates of the always-observed in the treatment group are obtained by considering the extreme scenarios in which the always-observed individuals have the highest or the lowest wage rates among the employed.[18] This reasoning holds unconditionally as well as conditionally on covariates.

To state these bounds formally, let D be the treatment indicator and S the employment indicator. Further, let X be some additional covariate. The conditional proportion of individuals among the employed in the treatment group who are employed if and only if they are treated is identified as

$$p(x) = 1 - \frac{\mathbb{P}(S = 1|D = 0, X = x)}{\mathbb{P}(S = 1|D = 1, X = x)}.$$

The lower and upper bounds on the local average treatment effect on wage rates are given by (Lee, 2009, Proposition 1b)

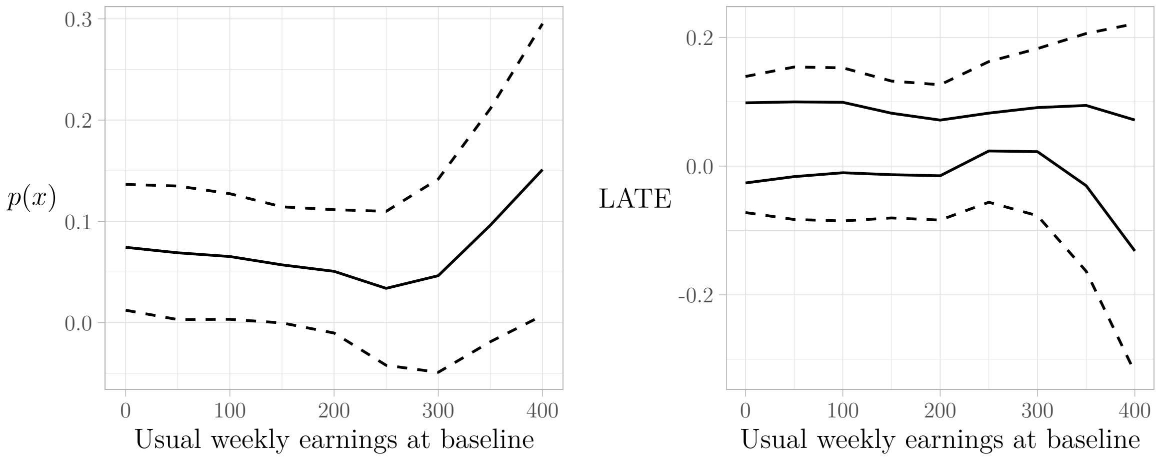
$$\Delta^L(x) = \mathbb{E}[Y|D = 1, S = 1, Y \leq Q_{DS}(1 - p(x), x), X = x] - \mathbb{E}[Y|D = 0, S = 1, X = x],$$
$$\Delta^U(x) = \mathbb{E}[Y|D = 1, S = 1, Y \geq Q_{DS}(p(x), x), X = x] - \mathbb{E}[Y|D = 0, S = 1, X = x],$$

[18] If the treatment discourages from working, then the control group would need to be truncated.

where $Q_{DS}(u,x)$ denotes the u-quantile of Y conditional on $D=1$, $S=1$, and $X=x$. Note that $p(x)$ is the proportion of data to be truncated conditional on $X=x$, so that $\eta = 1-p(x)$ in the notation from Section 1.2.

Lee (2009) conducts an intention-to-treat analysis, where the assignment to the training program is the treatment itself. Chen and Flores (2015) derive bounds on the treatment effect for the subpopulation of *always-employed compliers*, i.e. the individuals who comply with their treatment assignment and would be employed whether or not they obtained the treatment. Their bounds also involve truncated expectations. My estimator could be also applied to estimate the conditional versions of these bounds.

Estimation and Inference. I discuss the main ingredients of the bounds estimator. The details are given in Appendix 1.C.2. The conditional probabilities $\mathbb{P}(S=1|D=d, X=x)$ can be estimated using a local linear estimator with S_i as the outcome and X_i as a regressor, run on the sample restricted to observations with $D_i = d$ for $d \in \{0,1\}$,. Under regularity conditions, the resulting estimator $\widehat{\eta} = 1-\widehat{p}(x_0)$ satisfies the high-level assumption of Theorem 1.2. The truncated conditional expectations in the definition of $\Delta^L(x)$ and $\Delta^U(x)$ can be estimated using the estimator proposed in this paper and the conditional expectation function in the control group can be estimated using the standard local linear estimator. Restricting the samples based on the values of indicators S_i and D_i does not cause any complications in the asymptotic analysis. The estimators of the bounds have an asymptotically normal distribution, which can be used to form confidence intervals.



(a) The proportion of the employed induced to work by the treatment.

(b) Bounds on the LATE for the always observed (log wages).

Figure 1.5: Conditional Lee bounds for the Job Corps program.

Notes: The solid lines present the estimates of the bounds on the average treatment effect conditional on usual weekly earnings at baseline. The dashed lines mark pointwise 95% confidence intervals.

Empirical Application. I evaluate the effect of the job training offered under the Job Corps program in the United States. I use data from the National Job Corps Study conducted in mid 90s. I follow Lee (2009) closely in terms of the sample definition. The individuals who applied to the program were followed for four years after random assignment. There are 3599 individuals in the control group and 5546 in the treatment group, giving a total of 9145 observations. I investigate the effect on wage rates four years after the random assignment, conditioning on the usual weekly earnings at the most recent job reported at the baseline.

The results are presented in Figure 1.5. The bandwidth is selected based on smoothness constants calibrated through the procedure described in Appendix 1.D. The point estimates indicate that the treatment encourages taking up employment. The bounds on the treatment effect on wage rates are relatively flat for low weakly earnings at the baseline, where they are very similar to the unconditional estimates of Lee (2009). I note that there is a mass point in the distribution of the covariate at zero, but this does not invalidate the results.

1.7. CONCLUSIONS

I propose a nonparametric estimator of truncated conditional expectation functions based on an orthogonal conditional moment and local linear methods. When the truncation quantile level is known, I show that the feasible estimator is asymptotically equivalent to the oracle estimator, which uses the true conditional quantile function, and I find its asymptotic distribution. I also consider estimation with an estimated truncation quantile level. I considered estimation in two empirical settings: (i) sharp regression discontinuity designs with a manipulated running variable and (ii) program evaluation with sample selection.

Appendix

1.A. EXTENSIONS

In the main text, I consider local linear procedures with one covariate. It is straightforward to generalize the results to allow for a vector of covariates, and to use an arbitrary order of polynomials. I provide extensions in these two directions separately to avoid cumbersome notation, and to highlight different orders of the remainder term in the respective asymptotic equivalence results.

1.A.1. **Multivariate Case.** Let d be the dimension of X, and let $a = (a_1, ..., a_d)$ and $h = (h_1, ..., h_d)$ be vectors of bandwidths. Let $k(v) = \prod_{j=1}^d \mathcal{K}(v_j)$ be a d-dimensional product kernel built from the univariate kernel function $\mathcal{K}(\cdot)$. I put $|h| = \prod_{j=1}^d h_j$ and $k_h(v) = \prod_{j=1}^d \mathcal{K}(v_j/h_j)/h_j$, and similarly for a.

In the first step, I run a multivariate local linear quantile regression,

$$\begin{bmatrix} \widehat{q}_0(\eta, x_0; a) \\ \widehat{q}_1(\eta, x_0; a) \end{bmatrix} = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^n \rho_\eta(Y_i - \beta_0 - \beta_1^\top (X_i - x_0)) k_a(X_i - x_0).$$

Further,

$$\widehat{Q}^{ll}(\eta, x; x_0, a) = \widehat{q}_0(\eta, x_0; a) + \widehat{q}_1(\eta, x_0; a)^\top (x - x_0).$$

Finally,

$$\widehat{m}(\eta, x_0; a, h) = e_1^\top \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^n k_h(X_i - x_0)(\psi_i(\eta, \widehat{Q}^{ll}(\eta, X_i; x_0, a)) - \beta_0 - \beta_1^\top (X_i - x_0))^2,$$

where $e_1 = (1, 0, ..., 0)^\top$ is a $(d+1)$-dimensional vector. Likewise, the oracle estimator $\widetilde{m}(\eta, x_0; h)$ is defined as above but with $\psi_i(\eta, Q(\eta, X_i))$ as the outcome variable.

I maintain the smoothness assumptions on $Q(\eta, \cdot\,)$ with the understanding that for boundary points the derivatives exist in the directions in which x can be perturbed within $\mathcal{X}$. The assumptions on the kernel and the bandwidths are as follows.

Assumption 1.4*. *(a) Kernel: $\mathcal{K}$ is a bounded, symmetric density function with compact support, say $[-1, 1]$; (b) As $n \to \infty$, $\max_j h_j \to 0$, $\max_j a_j \to 0$, $n|h| \to \infty$, and $n|a| \to \infty$.*

Theorem 1.A.1 is the multivariate version of Theorem 1.1.

Theorem 1.A.1 (General d)**.** *Suppose that Assumptions 1.1, 1.2, and 1.4* hold, $h_j \asymp a_j$ for $j \in \{1, ..., d\}$, and that $\mathcal{X}$ is a convex set. Then*

$$\widehat{m}(\eta, x_0; a, h) = \widetilde{m}(\eta, x_0; h) + O_p\Big(\sum_j h_j^4 + (n|h|)^{-1}\Big).$$

For $d > 1$ the variance component of the remainder in Theorem 1.A.1 is of larger order than it is in Theorem 1.1. However, this result can still be used to obtain asymptotic normality because the oracle estimator has a bias of order $O_p(\sum h_j^2)$ and variance of order $O((n|h|)^{-1/2})$, which are smaller than the remainder in Theorem 1.A.1.

1.A.2. **Higher-Order Polynomials and Derivatives.** I introduce notation analogous to that in Section 1.2, making the dependence on p explicit. The local polynomial quantile estimates are given by

$$\widehat{q}^\top(\eta, x_0; a, p) = \underset{(\beta_0,...,\beta_p)^\top}{\arg\min} \sum_{i=1}^n k_h(X_i - x_0)\rho_\eta\Big(Y_i - \sum_{j=0}^p \frac{1}{j!}\beta_j(X_i - x_0)^j\Big).$$

I define the estimated p-th order approximation of $Q(\eta, \cdot\,)$ as

$$\widehat{Q}(\eta, x; x_0, a, p) = \sum_{j=0}^p \frac{1}{j!}\widehat{q}_j(\eta, x_0; a, p)(x - x_0)^j.$$

The estimator of the r-th derivative of $m(\eta, x)$ with respect to x at x_0, $\partial_x^r m(\eta, x_0)$, is defined as

$$\begin{aligned}&\widehat{m}_r(\eta, x_0; a, h, p)\\ &\quad = e_{r+1}^\top \underset{\beta}{\arg\min} \sum_{i=1}^n k_h(X_i - x_0)\Big(\psi_i(\eta, \widehat{Q}(\eta, X_i; x_0, a, p)) - \sum_{j=0}^p \frac{1}{j!}\beta_j(X_i - x_0)^j\Big)^2,\end{aligned}$$

where e_{r+1} is a $(p+1)$-dimensional vector with 1 at the $(r+1)$-th position and 0 otherwise. Likewise, the oracle estimator $\widetilde{m}_r(\eta, x_0; h, p)$ is defined as above but with $\psi_i(\eta, Q(\eta, X_i))$ as the outcome variable.

In order to prove an analog of Theorem 1.1, I require one natural modification of Assumption 1.2. I assume that the function $Q(\eta, x)$ is $p+1$ times continuously differentiable with respect to x (instead of twice).

Assumption 1.2*. *$\partial_x^{p+1}Q(\eta, x)$ is continuous in x, and Assumptions 2(b) and 2(c) hold.*

Theorem 1.A.2. *Suppose that Assumptions 1.1, 1.2*, and 1.4 hold, and that $h \asymp a$. Then*

$$\widehat{m}_r(\eta, x_0; a, h, p) = \widetilde{m}_r(\eta, x_0; h, p) + O_p(h^{-r}(h^{2(p+1)} + (nh)^{-1})).$$

With this result, under modified Assumption 1.3, asymptotic normality follows e.g. from the results of Hong (2003). The stochastic part of $h^r(\widetilde{m}_r(\eta, x_0; h, p) - \partial_x^r m(\eta, x_0))$ is of order $O_p((nh)^{-1/2})$, and its leading bias is of order $O_p(h^{p+1})$ or $O_p(h^{p+2})$. Theorem 1.A.2 allows to characterize the leading bias for all orders p and derivatives $r \leq p$, both for interior and boundary points, except for the local constant estimator for interior points. Its leading bias is of order $O_p(h^2)$, which is the same as the order of the remainder in the above theorem. This case is discussed by Kato (2012).

1.B. ALTERNATIVE APPROACHES

I discuss in detail the two alternative approaches introduced in Section 1.4.3. As reference points, I also present the asymptotic distributions of the corresponding oracle estimators employing the true conditional quantile function. Next, for interior points, I contrast my approach from Section 1.2 with the weighted Nadaraya-Watson estimator of Kato (2012).

1.B.1. **Local Linear Estimator Based on a Non-Orthogonal Moment.**

First, I show that in the special case when the same bandwidth is used in both stages, the estimator $\widehat{m}^{NM}(\eta, x_0; h, h)$ is asymptotically equivalent to the oracle estimator $\widetilde{m}(\eta, x_0; h)$, and I give the exact rate of the remainder. Second, I derive the asymptotic distribution in the general case allowing for different bandwidths.

Proposition 1.B.1. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then*

$$R^{NM}(\eta, x_0; h) \equiv \widehat{m}^{NM}(\eta, x_0; h, h) - \widetilde{m}(\eta, x_0; h) = O_p((h + (nh)^{-1/2})(h^2 + (nh)^{-1/2})).$$

Suppose additionally that $f(x)$ is continuously differentiable and x_0 is an interior point, or that $\partial_x^1 Q(\eta, x_0) = 0$. Then $R^{NM}(\eta, x_0; h) = O_p(h^4 + (nh)^{-1})$.

Let $\widetilde{m}^{NM}(x_0, \eta; h)$ be the oracle estimator corresponding to $\widehat{m}^{NM}(x_0, \eta; a, h)$, i.e. a local linear estimator with $\frac{1}{\eta} Y_i \mathbb{1}(Y_i \leq Q(\eta, X_i))$ as the outcome variable.

Proposition 1.B.2. *Suppose that Assumptions 1.1–1.4 hold, and $h/a \to \rho \in (0, \infty)$. Then*

(i) $\sqrt{nh}\Big(\widetilde{m}^{NM}(x_0, \eta; h) - m(\eta, x_0) - \widetilde{\mathcal{B}}^{NM}(\eta, x_0, h)\Big) \xrightarrow{d} \mathcal{N}(0, \widetilde{V}^{NM}(\eta, x_0)),$

where

$$\widetilde{\mathcal{B}}^{NM}(\eta, x_0, h) = \frac{1}{2}\mu \partial_x^2 m(\eta, x_0) h^2 + o_p(h^2),$$
$$\widetilde{V}^{NM}(\eta, x_0) = \frac{\kappa}{\eta f_X(x_0)} \Big(Var(Y | Y \leq Q(\eta, x_0), X = x_0) + (1 - \eta) m(\eta, x_0)^2 \Big).$$

(ii) $\sqrt{nh}\Big(\widehat{m}^{NM}(x_0, \eta; a, h) - m(\eta, x_0) - \mathcal{B}^{NM}(\eta, x_0, a, h)\Big) \xrightarrow{d} \mathcal{N}(0, V^{NM}(\eta, x_0, \rho)),$

where

$$\begin{aligned}\mathcal{B}^{NM}(\eta,x_0,a,h) &= \frac{1}{2}\mu\left(\partial_x^2 m(\eta,x_0)h^2 + C^{NM}(\eta,x_0)\partial_x^2 Q(\eta,x_0)(a^2-h^2)\right) + o_p(h^2),\\ V^{NM}(\eta,x_0,\rho) &= \frac{\kappa}{\eta f_X(x_0)} Var(Y|Y\le Q(\eta,x_0),X=x_0) + \frac{1-\eta}{\eta f(x_0)(\mu_0\mu_2-\mu_1^2)^2}\\ &\quad\times\int_{\mathcal{D}}\left(k(v)(\mu_2-\mu_1 v)\frac{1}{\eta}m(\eta,x_0) + \rho k(v\rho)(\mu_2-\mu_1 v\rho)\frac{1}{\eta}Q(\eta,x_0)\right)^2 dv\end{aligned}$$

with $C^{NM}(\eta,x_0)=\frac{1}{\eta}f_{Y|X}(Q(\eta,x_0)|x_0)Q(\eta,x_0)$*,* $\mathcal{D}=[-1,1]$ *if* x_0 *lies in the interior of* $\mathcal{X}$*,* $\mathcal{D}=[0,1]$ *if* x_0 *lies on the boundary of* $\mathcal{X}$*, and* $\mu_j=\int_{\mathcal{D}}k(v)v^j dv$*.*

Both bandwidths appear in the bias formula and the ratio ρ appears in the asymptotic variance. When ρ is small, i.e. a is large relative to h, then the variance of the feasible estimator is close to the variance of the oracle estimator because $V^{NM}(\eta,x_0,0)=\widetilde{V}^{NM}(\eta,x_0)$.

In the proof, I give an expansion of the feasible estimator $\widehat{m}^{NM}$ about the infeasible $\widetilde{m}^{NM}$. The bias $\mathcal{B}^{NM}(\eta,x_0,a,h)$ differs from the oracle bias due to the fact that, first, $Q(\eta,\cdot)$ is replaced by its local linear approximation, and, second, this approximation is estimated. The factor $C^{NM}(\eta,x_0)$ equals the derivative of $\frac{1}{\eta}\mathbb{E}[Y\mathbb{1}(Y\le y)|X=x_0]$ with respect to y evaluated at $Q(\eta,x_0)$,

$$C^{NM}(\eta,x_0)=\frac{d}{dy}\mathbb{E}\left[\frac{1}{\eta}Y\mathbb{1}(Y\le y)|X=x_0\right]\bigg|_{y=Q(\eta,x_0)}.$$

1.B.2. **Local Linear Estimator on a Truncated Sample.** First, I show that in the special case when the same bandwidth is used in both stages, the estimator $\widehat{m}^{TS}(\eta,x_0;h,h)$ is asymptotically equivalent to the oracle estimator $\widetilde{m}(\eta,x_0;h)$, and I give the exact rate of the remainder. Second, I derive the asymptotic distribution in the general case allowing for different bandwidths.

Proposition 1.B.3. *Suppose that Assumptions 1.1–1.4 hold. Then*

$$R^{TS}(\eta,x_0;h)\equiv\widehat{m}^{TS}(\eta,x_0;h,h)-\widetilde{m}(\eta,x_0;h)=O_p((h+(nh)^{-1/2})(h^2+(nh)^{-1/2})).$$

Suppose additionally that $f(x)$ *is continuously differentiable and* x_0 *is an interior point, or that* $\partial_x^1 Q(\eta,x_0)=\partial_x^1 m(\eta,x_0)$*. Then* $R^{TS}(\eta,x_0;h)=O_p(h^4+(nh)^{-1})$*.*

Let $\widetilde{m}^{TS}(x_0,\eta;h)$ be the oracle estimator corresponding to the estimator $\widehat{m}^{TS}(x_0,\eta;a,h)$, i.e. a local linear estimator using observations with $Y_i\le Q(\eta,X_i)$.

Proposition 1.B.4. *Suppose that Assumptions 1.1–1.4 hold, and* $h/a\to\rho\in(0,\infty)$*. Then* *(i)* $\sqrt{nh}\Big(\widetilde{m}^{TS}(\eta, x_0; h) - m(\eta, x_0) - \widetilde{\mathcal{B}}^{TS}(\eta, x_0, h)\Big) \xrightarrow{d} \mathcal{N}(0, \widetilde{V}^{TS}(\eta, x_0))$, *where*

$$\widetilde{\mathcal{B}}^{TS}(\eta, x_0, h) = \frac{1}{2}\mu \partial_x^2 m(\eta, x_0) h^2 + o_p(h^2),$$
$$\widetilde{V}^{TS}(\eta, x_0) = \frac{\kappa}{\eta f_X(x_0)} Var(Y|Y \le Q(\eta, x_0), X = x_0).$$

(ii) $\sqrt{nh}\Big(\widehat{m}^{TS}(\eta, x_0; a, h) - m(\eta, x_0) - \mathcal{B}^{TS}(\eta, x_0, a, h)\Big) \xrightarrow{d} \mathcal{N}(0, V^{TS}(\eta, x_0, \rho))$, *where*

$$\mathcal{B}^{TS}(\eta, x_0, a, h) = \frac{1}{2}\mu\left(\partial_x^2 m(\eta, x_0) h^2 - C^{TS}(\eta, x_0)\partial_x^2 Q(\eta, x_0)(h^2 - a^2)\right) + o_p(h^2),$$
$$V^{TS}(\eta, x_0, \rho) = \frac{\kappa}{\eta f_X(x_0)}\Big(Var(Y|Y \le Q(\eta, x_0), X = x_0) + \rho(1-\eta)\left(Q(\eta, x_0) - m(\eta, x_0)\right)^2 \Big)$$

with $C^{TS}(\eta, x_0) = \frac{1}{\eta} f_{Y|X}(Q(\eta, x_0)|x_0)(Q(\eta, x_0) - m(\eta, x_0))$.

As in the case of the estimator using a non-orthogonal moment, both bandwidths appear in the bias formula, and the ratio ρ appears in the asymptotic variance. When ρ is small, i.e. a is large relative to h, then the variance of the feasible estimator is close to the variance of the oracle estimator because $V^{TS}(\eta, x_0, 0) = \widetilde{V}^{TS}(\eta, x_0)$.

The factor $C^{TS}(\eta, x_0)$ equals the derivative of $\mathbb{E}[Y|X = x_0, Y \le y]$ with respect to y evaluated at $Q(\eta, x_0)$,

$$C^{TS}(\eta, x_0) = \frac{d}{dy}\mathbb{E}[Y|X = x_0, Y \le y]\Big|_{y=Q(\eta, x_0)}.$$

1.B.3. **Weighted Nadaraya-Watson Estimation for Interior Points.** I contrast my estimator $\widehat{m}$ with the estimator of Kato (2012) based on the weighted Nadaraya-Watson (WNW) estimator of the conditional c.d.f. For interior points, the WNW estimator is asymptotically equivalent to the local linear estimator. Additionally, the WNW estimator of $F_{Y|X}(y|x_0)$, i.e. applied to the data with $\mathbb{1}(Y_i \le y)$ as the outcome variable, is monotone in y, and it lies between 0 and 1. Both these properties are not shared by the local linear estimator.[19] I emphasize that the WNW estimator is not defined for boundary points, but for interior points the estimator of Kato (2012) bears some similarity with the approaches developed in this paper.

In the first step, Kato (2012) estimates the conditional c.d.f. as

$$\widehat{F}^{WNW}_{Y|X}(y|x_0; h) = \frac{\sum_{i=1}^n p_i(x_0) k_h(X_i - x_0)\mathbb{1}(Y_i \le y)}{\sum_{i=1}^n p_i(x_0) k_h(X_i - x_0)}, \tag{1.B.1}$$

where $p_i(x_0) \ge 0$ are the empirical likelihood weights, which maximize $\sum_{i=1}^n \log(p_i(x_0))$

[19] Nevertheless, the asymptotic properties remain the same when the weighted Nadaraya-Watson estimator is replaced with the local linear estimator.

subject to the constraints $\sum_{i=1}^n p_i(x_0) = 1$ and $\sum_{i=1}^n p_i(x_0)(X_i - x_0)k_h(X_i - x_0) = 0$.[20] He estimates $Q(\eta, x_0)$ as $\widehat{Q}^{WNW}(\eta, x_0; h) = \inf\{y : \eta \leq \widehat{F}_{Y|X}^{WNW}(y|x_0; h)\}$, and $m(\eta, x_0)$ as

$$\widehat{m}^{WNW}(\eta, x_0; h) = \frac{\sum_{i=1}^n p_i(x_0)k_h(X_i - x_0)Y_i \mathbb{1}(Y_i \leq \widehat{Q}^{WNW}(\eta, x_0; h))}{\sum_{i=1}^n p_i(x_0)k_h(X_i - x_0)\mathbb{1}(Y_i \leq \widehat{Q}^{WNW}(\eta, x_0; h))},$$

which is essentially the WNW estimator with $\frac{1}{\eta}Y_i\mathbb{1}(Y_i \leq \widehat{Q}^{WNW}(\eta, x_0; h))$ as the outcome variable. Kato (2012) shows that, under suitable assumptions, the estimator $\widehat{m}^{WNW}$ is asymptotically equivalent to the WNW estimator (and hence to the local linear estimator) with $\psi_i(\eta, Q(\eta, x_0))$ as the outcome variable. In consequence, it is asymptotically normal with asymptotic variance $V(\eta, x_0)$ defined in Corollary 1.1,[21] and its leading bias is given by

$$B^{WNW}(\eta, x_0, h) = \frac{1}{2}\mu \frac{d^2}{dx^2}\mathbb{E}[\psi(\eta, Q(\eta, x_0)|X = x)]|_{x=x_0}h^2.$$

The difference between the WNW approach and my approach, for interior points, results from the fact that they estimate different curves which coincide only at the evaluation point x_0. The two approaches have the same asymptotic variance but their biases are different, as shown in Proposition 1.B.5.

Proposition 1.B.5. *Suppose that $F_{Y|X}(y|x)$ is twice continuously differentiable. Then*

$$B^{WNW}(\eta, x_0, h) = B(\eta, x_0, h) - \frac{1}{2\eta}\mu f_{Y|X}(Q(\eta, x_0)|x_0)(\partial_x Q(\eta, x_0))^2 h^2.$$

The second term of the difference on the right-hand side is always non-negative, so that $B^{WNW}(\eta, x_0, h) \leq B(\eta, x_0, h)$. However, which of the two biases is larger in absolute value, depends on the specific data generating process. For example, it is possible that $B^{WNW}(\eta, x_0, h) = 0$ and $B(\eta, x_0, h) > 0$, or that $B^{WNW}(\eta, x_0, h) < 0$ and $B(\eta, x_0, h) = 0$.

However, I remark that in a simple location-scale model with a linear conditional expectation function and homoskedastic residuals, my estimator has no bias, whereas $|B^{WNW}(\eta, x_0, h)|$ can be arbitrarily large.

1.C. APPLICATIONS: ESTIMATION AND INFERENCE DETAILS

I formally introduce the estimators of the bounds in RD designs with a manipulated running variable discussed in Section 1.6.1 and of the conditional Lee bounds discussed in Section 1.6.2. Their asymptotic distributions follow easily from Theorems 1.1 and 1.2, and hence are stated without proofs.

[20] When x_0 lies on the boundary, so that all $X_i - x_0$ have the same sign, it is not possible to find non-negative weights satisfying the last constraint.

[21] Kato (2012) considers time series data, but the asymptotic variance of his estimator is the same as for i.i.d. data because of the localization effect (see his discussion following Theorem 1).

1.C.1. **Estimation in RD Designs with Manipulation.** I normalize the cutoff to zero. Let $\bar{\mu} = (\bar{\mu}_2^2 - \bar{\mu}_1\bar{\mu}_3)/(\bar{\mu}_2\bar{\mu}_0 - \mu_1^2)$ and $\bar{\kappa} = \int_0^\infty (k(v)(\bar{\mu}_1 v - \bar{\mu}_2))^2 dv/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2)^2$, where $\bar{\mu}_j = \int_0^\infty v^j k(v)dv$. Further, I define $k_h^-(v) = \mathbb{1}(v < 0)k_h(v)$ and $k_h^+(v) = \mathbb{1}(v \geq 0)k_h(v)$.

The share of always-assigned units among all units just to the right of the cutoff is estimated as:

$$\widehat{\tau} = \max\left\{1 - \widehat{f}^-/\widehat{f}^+, 0\right\},$$

where $\widehat{f}^-$ and $\widehat{f}^+$ are estimators of $f_X(0^-)$ and $f_X(0^+)$, respectively. In the notation from the main text, $\widehat{\eta} = 1 - \widehat{\tau} = \min\{\widehat{f}^-/\widehat{f}^+, 1\}$. The density limits can be estimated using, e.g., the 'linear' boundary kernel (Jones, 1993) as

$$\widehat{f}^+ = \frac{1}{n}\sum_{i=1}^n k_b^+(X_i)\frac{\bar{\mu}_2 - \bar{\mu}_1|X_i/b|}{\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2} \text{ and } \widehat{f}^- = \frac{1}{n}\sum_{i=1}^n k_b^-(X_i)\frac{\bar{\mu}_2 - \bar{\mu}_1|X_i/b|}{\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2}. \tag{1.C.1}$$

To analyze this estimator, I impose smoothness assumptions on the density.

Assumption 1.7. *There exists $\epsilon > 0$ such that $f(\cdot)$ is twice continuously differentiable on $(-\epsilon, 0) \cup (0, \epsilon)$. Moreover, for the limits $\partial_x^j f_X(0^+)$ and $\partial_x^j f_X(0^-)$ exist for $j \in \{0, 1, 2\}$, and $f_X(0^+), f_X(0^-) > 0$.*

Lemma 1.C.1 yields an asymptotical linear representation of $\widehat{\eta}$.

Lemma 1.C.1. *Suppose that Assumptions 1.1, 1.4(a), and 1.7 hold. Moreover, $b \to 0$ and $nb \to \infty$. Then*

$$\frac{1}{\eta}\Big(\widehat{\eta} - \eta\Big) = \frac{\widehat{f}^- - f_X(0^-)}{f_X(0^-)} - \frac{\widehat{f}^+ - f_X(0^+)}{f_X(0^+)} + o(b^2) + o_p((nb)^{-1/2}).$$

I note that the asymptotic bias and variance of $\frac{1}{\eta}\Big(\widehat{\eta} - \eta\Big)$ are given by

$$A_\eta = \frac{1}{2}\bar{\mu}\left(\frac{f_X''(0^-)}{f_X(0^-)} - \frac{f_X''(0^+)}{f_X(0^+)}\right)b^2 + o(b^2) \text{ and } W_\eta = \bar{\kappa}\left(\frac{1}{f_X(0^+)} + \frac{1}{f_X(0^-)}\right).$$

These quantities appear in the asymptotic distribution of the bounds. The lemma implies that for bandwidths $b \asymp h$ this estimator satisfies Assumption 1.5.

Let $m(x) = \mathbb{E}[Y|X = x]$, $m^L(\eta, x) = \mathbb{E}[Y|X = x, Y \leq Q(\eta, x)]$, and $m^U(\eta, x) = \mathbb{E}[Y|X = x, Y \geq Q(1 - \eta, x)]$. The truncated conditional expectations $m^L(\eta, 0^+)$ and $m^U(\eta, 0^+)$ are estimated as

$$\widehat{m}_+^L(\widehat{\eta}) = e_1^\top \underset{\beta_0,\beta_1}{\arg\min}\sum_{i=1}^n k_h^+(X_i)(\psi_i^L(\widehat{\eta}, \widehat{Q}^{ll,+}(\widehat{\eta}, X_i; h)) - \beta_0 - \beta_1 X_i)^2,$$

$$\widehat{m}_+^U(\widehat{\eta}) = e_1^\top \underset{\beta_0,\beta_1}{\arg\min}\sum_{i=1}^n k_h^+(X_i)(\psi_i^U(\widehat{\eta}, \widehat{Q}^{ll,+}(1 - \widehat{\eta}, X_i; h)) - \beta_0 - \beta_1 X_i)^2,$$

where $\psi_i^L(u,q) = \psi_i(u,q)$ and $\psi_i^U(u,q) = \frac{1}{u}Y_i\mathbb{1}(q \leq Y_i) - \frac{1}{u}q(\mathbb{1}(q \leq Y_i) - u)$. The estimated quantile function $\widehat{Q}^{ll,+}$ is defined as in Section 1.2, except that it uses only observations to the right of the cutoff.

The conditional expectation $m(x_0^-)$ is estimated as

$$\widehat{m}_- = e_1^\top \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h^-(X_i)(Y_i - \beta_0 - \beta_1 X_i)^2.$$

The final estimators of the bounds on Γ are defined as

$$\widehat{\Gamma}^L = \widehat{m}_+^L(\widehat{\eta}) - \widehat{m}_- \text{ and } \widehat{\Gamma}^U = \widehat{m}_+^U(\widehat{\eta}) - \widehat{m}_-.$$

The asymptotic analysis requires some natural modifications of Assumptions 1.2 and 1.3 to analyze $\widehat{m}_L^+(\widehat{\eta})$ and $\widehat{m}_U^+(\widehat{\eta})$. Additionally, I impose standard assumption for the analysis of $\widehat{m}^-$.

Assumption 1.8. *For some $\epsilon > 0$ the following hold on $(-\epsilon, 0)$. (a) $m(x)$ is twice continuously differentiable in x, and $m(0^-)$, $m'(0^-)$ and $m''(0^-)$ exist; (b) $Var(Y|X = x)$ is continuous and $Var(Y|X = x^-)$ exists; (c) There exists $\xi > 0$ s.t $\mathbb{E}\Big[|Y|^{2+\xi}\Big|X = x\Big]$ is uniformly bounded.*

Proposition 1.C.1 establishes joint convergence of the bounds estimators.

Proposition 1.C.1. *Suppose that the Assumptions 1.1–1.4 and 1.6 hold, mutatis mutandis. Furthermore, Assumptions 1.7 and 1.8 hold, and $h/b \to \nu$. Then*

$$\sqrt{nh}\begin{bmatrix}\widehat{\Gamma}^L - \Gamma^L - (B_+^L(\eta) - B_-)\\ \widehat{\Gamma}^U - \Gamma^U - (B_+^U(\eta) - B_-)\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} V_+^L(\eta) + V_- & Cov_+(\eta) + V_- \\ Cov_+(\eta) + V_- & V_+^U(\eta) + V_-\end{bmatrix}\right),$$

where for $* \in \{L, U\}$

$$\begin{aligned} B_+^*(\eta) &= \frac{1}{2}\bar{\mu}\partial_x^2 m^*(\eta, 0^+)h^2 + o_p(h^2) + D_+^* A_\eta, & B_- &= \frac{1}{2}\bar{\mu}\partial_x^2 m(0^-)h^2 + o_p(h^2),\\ V_+^*(\eta) &= \frac{\bar{\kappa}}{f_X(0^+)}Var(\psi^*|X = 0^+) + \nu(D_+^*)^2 W_\eta, & V_- &= \frac{\bar{\kappa}}{f_X(0^-)}Var(Y|X = 0^-),\\ Cov_+(\eta) &= \frac{\bar{\kappa}}{f_X(0^+)}Cov(\psi^L, \psi^U|X = 0^+) + \nu D_L^+ D_U^+ W_\eta, \end{aligned}$$

with $\psi^L \equiv \psi^L(\eta, Q(\eta, X))$, $\psi^U \equiv \psi^U(\eta, Q(1-\eta, X))$, $D_+^L \equiv Q(\eta, 0^+) - m^L(\eta, 0^+)$, and $D_+^U \equiv Q(1-\eta, 0^+) - m^U(\eta, 0^+)$.

Since $\widehat{\eta}$ is obtained based only on realizations of the covariate, there is no asymptotic covariance between $\widehat{\eta}$ and the estimators of the three conditional expectations. The component in the asymptotic covariance due to estimation of η is negative since $D_+^L(\eta) > 0$ and $D_-^U(\eta) < 0$.

1.C.2. **Estimation of Conditional Lee Bounds.** The derivation follows the same steps as for regression discontinuity designs with a manipulated running variable. For $d \in \{0,1\}$, let $s_d(x) = \mathbb{P}(S = 1|D = d, X = x)$. The probability $s_d(x_0)$ can be estimated using the standard local linear estimator with the sample restricted to observations with $D_i = d$,

$$\widehat{s}_d(x_0) = e_1^\top \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h(X_i - x_0)(S_i - \beta_0 - \beta_1(X_i - x_0))^2 \mathbb{1}(D_i = d). \tag{1.C.2}$$

Let

$$\widehat{\eta}(x_0) = \frac{\widehat{s}_0(x_0)}{\widehat{s}_1(x_0)}.$$

To analyze the above estimator, I impose the following assumption.

Assumption 1.9. *(a) $s_d(x)$ is twice continuously differentiable for $d \in \{0,1\}$; (b) $\mathbb{E}[D|X = x]$ is continuous in x.*

Lemma 1.C.2. *Suppose that Assumptions 1.1, 1.4(a), and 1.9 hold. Moreover, $b \to 0$ and $nb \to \infty$. Then*

$$\frac{1}{\eta}\Big(\widehat{\eta}(x_0) - \eta(x_0)\Big) = \frac{\widehat{s}_0(x_0) - s_0(x_0)}{s_0(x_0)} - \frac{\widehat{s}_1(x_0) - s_1(x_0)}{s_1(x_0)} + o_p(b^2 + (nb)^{-1/2}).$$

I note that the asymptotic bias and variance of $\frac{1}{\eta}\Big(\widehat{\eta} - \eta\Big)$ are given by

$$\begin{aligned} A_\eta^{Lee} &= \frac{1}{2}\mu\Big\{\frac{s_0''(x_0)}{s_0(x_0)} - \frac{s_1''(x_0)}{s_1(x_0)}\Big\}b^2 + o_p(b^2) \\ W_\eta^{Lee} &= \frac{\kappa}{f_X(x_0)}\Big\{\frac{s_0(x_0)(1 - s_0(x_0))}{\mathbb{P}(D = 0|X = x_0)s_0(x_0)^2} + \frac{s_1(x_0)(1 - s_1(x_0))}{\mathbb{P}(D = 1|X = x_0)s_1(x_0)^2}\Big\}. \end{aligned}$$

These quantities appear in the asymptotic distribution of the bounds.

Recall that $Q_{DS}(u, x)$ denotes the u-quantile of Y conditional on $D = 1$, $S = 1$, and $X = x$. Further, let $m_{Lee}^L(\eta, x) = \mathbb{E}[Y|X = x, Y \leq Q_{DS}(\eta, x), D = 1, S = 1]$, $m_{Lee}^U(\eta, x) = \mathbb{E}[Y|X = x, Y \geq Q_{DS}(1 - \eta, x), D = 1, S = 1]$, and $m_{Lee}(x) = \mathbb{E}[Y|X = x, D = 0, S = 1]$.

The truncated conditional expectations $m_L^{Lee}(\eta, x_0^+)$ and $m_U^{Lee}(\eta, x_0^+)$ are estimated as

$$\begin{aligned} &\widehat{m}_{Lee}^L(x_0) \\ &\quad = e_1^\top \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h(X_i - x_0)S_iD_i(\psi_i^L(\widehat{\eta}(x_0), \widehat{Q}_{DS}^{ll}(\widehat{\eta}(x_0), X_i; x_0, h)) - \beta_0 - \beta_1(X_i - x_0))^2, \\ &\widehat{m}_{Lee}^U(x_0) \\ &\quad = e_1^\top \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h(X_i - x_0)S_iD_i(\psi_i^U(\widehat{\eta}(x_0), \widehat{Q}_{DS}^{ll}(1 - \widehat{\eta}(x_0), X_i; x_0, h)) - \beta_0 - \beta_1(X_i - x_0))^2, \end{aligned}$$

where $\psi_i^L(u, q) = \psi_i(u, q)$ and $\psi_i^U(u, q) = \frac{1}{u}Y_i\mathbb{1}(q \leq Y_i) - \frac{1}{u}q(\mathbb{1}(q \leq Y_i) - u)$.

The conditional expectation $m_{Lee}(x_0)$ is estimated as

$$\widehat{m}_{Lee}(x_0) = e_1^\top \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} k_h(X_i - x_0) S_i (1 - D_i)(Y_i - \beta_0 - \beta_1 (X_i - x_0))^2.$$

The final estimators of the bounds on the conditional average treatment effect are defined as

$$\begin{aligned} \widehat{\Delta}^L(x_0) &= \widehat{m}^L_{Lee}(x_0) - \widehat{m}_{Lee}(x_0), \\ \widehat{\Delta}^U(x_0) &= \widehat{m}^U_{Lee}(x_0) - \widehat{m}_{Lee}(x_0). \end{aligned}$$

I impose standard assumptions for the analysis of $\widehat{m}^{Lee}(x_0)$.

Assumption 1.10. *(a) $\partial_x^2 m_{Lee}(x)$ is continuous in x; (b) $Var(Y|X = x, D = 0, S = 1)$ is continuous in x; (c) $\mathbb{E}\left[|Y|^{2+\xi}|X = x, S = 1, D = 0\right]$ is bounded uniformly over $x \in \mathcal{X}$ for some $\xi > 0$.*

Proposition 1.C.2 establishes joint convergence of the bounds estimators. The dependence on x_0 is dropped to ease the notation.

Proposition 1.C.2. *Suppose that the Assumptions 1.1–1.4 and 1.6 hold, mutatis mutandis. Furthermore, Assumptions 1.9 and 1.10 hold, $h = O(n^{-1/5})$, and $h/b \to \nu$. Then*

$$\sqrt{nh} \begin{bmatrix} \widehat{\Delta}^L - \Delta^L - (B^L_{Lee} - B_{Lee}) \\ \widehat{\Delta}^U - \Delta^U - (B^U_{Lee} - B_{Lee}) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} V^L_{Lee} + V_{Lee} & Cov^{Lee} + V_{Lee} \\ Cov^{Lee} + V_{Lee} & V^U_{Lee} + V_{Lee} \end{bmatrix}\right),$$

where for $* \in \{L, U\}$

$$\begin{aligned} B^*_{Lee} &= \frac{1}{2}\mu \partial_x^2 m^*_{Lee}(\eta(x_0), x_0) h^2 + D^*_{Lee} A^{Lee}_\eta, \\ V^*_{Lee} &= \frac{\kappa}{f_X(x_0)\mathbb{E}[SD|X = x_0]} Var(\psi^*|X = x_0, S = 1, D = 1) + \nu (D^*_{Lee})^2 W^{Lee}_\eta, \\ Cov_{Lee} &= \frac{\kappa}{f_X(x_0)\mathbb{E}[SD|X = x_0]} Cov(\psi^L, \psi^U|X = x_0, S = 1, D = 1) + \nu D^L_{Lee} D^U_{Lee} W^{Lee}_\eta, \\ B_{Lee} &= \frac{1}{2}\mu \partial_x^2 m^{Lee}(x) h^2 + o_p(h^2), \\ V_{Lee} &= \frac{\kappa}{f_X(x_0)\mathbb{E}[S(1-D)|X = x_0]} Var(Y|X = x_0, S = 1, D = 0) \end{aligned}$$

with $\psi^L \equiv \psi^L(\eta(X), Q_{DS}(\eta(X), X))$, $\psi^U \equiv \psi^U(\eta(X), Q_{DS}(1 - \eta(X), X))$, $D^L_{Lee} \equiv Q_{DS}(\eta(x_0), x_0) - m^L_{Lee}(\eta(x_0), x_0)$, and $D^U_{Lee} \equiv Q_{DS}(1 - \eta(x_0), x_0) - m^U_{Lee}(\eta(x_0), x_0)$.

1.D. RULE OF THUMB FOR THE SMOOTHNESS CONSTANT

Armstrong and Kolesár (2020) propose a rule of thumb to calibrate the bound on the second derivative of the conditional expectation function. They run a quartic, global

Table 1.D.1: Coverage, average bandwidth, and average length of the 95% CI.

		Coverage			Bandwidth			CI length		
	Design for m_j:	1	2	3	1	2	3	1	2	3
Homoskedastic errors										
$\eta = 0.2$	Oracle $\widetilde{m}$	93.6	92.1	95.4	0.231	0.310	0.257	0.128	0.113	0.120
	Feasible $\widehat{m}$	93.4	92.2	95.7	0.227	0.307	0.260	0.128	0.113	0.119
$\eta = 0.5$	Oracle $\widetilde{m}$	95.0	93.1	96.0	0.207	0.279	0.231	0.104	0.091	0.098
	Feasible $\widehat{m}$	94.9	93.3	96.1	0.204	0.277	0.233	0.104	0.092	0.098
$\eta = 0.8$	Oracle $\widetilde{m}$	95.7	94.0	96.2	0.197	0.266	0.222	0.095	0.083	0.089
	Feasible $\widehat{m}$	95.7	94.0	96.4	0.196	0.265	0.222	0.095	0.084	0.089
Heteroskedastic errors										
$\eta = 0.2$	Oracle $\widetilde{m}$	93.4	92.6	95.6	0.239	0.310	0.250	0.129	0.115	0.123
	Feasible $\widehat{m}$	93.5	92.9	95.8	0.235	0.307	0.254	0.129	0.116	0.122
$\eta = 0.5$	Oracle $\widetilde{m}$	95.0	93.6	96.5	0.213	0.277	0.225	0.104	0.093	0.100
	Feasible $\widehat{m}$	95.1	93.7	96.5	0.210	0.276	0.227	0.105	0.094	0.100
$\eta = 0.8$	Oracle $\widetilde{m}$	95.7	94.3	96.6	0.202	0.264	0.215	0.095	0.085	0.091
	Feasible $\widehat{m}$	95.7	94.3	96.7	0.201	0.263	0.216	0.096	0.085	0.092

Notes: Estimators evaluated with their respective RMSE-optimal bandwidths. The sample size is $n = 1,000$, and the number of simulations is $S = 10,000$. The smoothness constant is selected using the rule of thumb discussed in Section 1.D.

regression, and estimate the maximal second derivative based on it. I adapt this approach to calibrate the bound on $\partial_x^2 m(\eta, x)$. In the first stage, I run a global, quartic quantile regression. I denote the resulting estimator as $\widehat{Q}^{glob}(\eta, X_i)$. In the second stage, I run a global quartic regression with $\psi_i(\eta, \widehat{Q}^{glob}(\eta, X_i))$ as the outcome variable.

I investigate the performance of this procedure in the setting from Section 1.5. The results are presented in Table 1.D.1. In this example, the rule of thumb leads to CIs with good coverage properties. This is consistent with the findings of Armstrong and Kolesár (2020).

1.E. PROOFS OF THE RESULTS IN THE MAIN TEXT

I define additional, shorthand notation. Let $q_0(\eta) = Q(\eta, x_0)$, $q_1(\eta) = \partial_x^1 Q(\eta, x_0)$, $\widehat{q}_0(\eta; a) = \widehat{q}_0(\eta, x_0; a)$, $\widehat{q}_1(\eta; a) = \widehat{q}_1(\eta, x_0; a)$, $\widehat{Q}(\eta, x; a) = \widehat{Q}^{ll}(\eta, x; x_0, a)$, $k_{h,i} = k_h(X_i - x_0)$, $X_{h,i} = (X_i - x_0)/h$, $\widetilde{X}_{h,i} = (1, X_{h,i})^\top$, $Q^*(\eta, x) = q_0(\eta) + q_1(\eta)(x - x_0)$, $L_i(b) = b_0 + b_1(X_i - x_0)$, $Y_i'(b) = Y_i - L_i(b)$, and $\mathcal{X}_h = \mathcal{X}(x_0, h)$. I put $C_f \equiv \sup\{|f_{Y|X}(y, x)| : x \in \mathcal{X} \text{ and } y \in [Q(\eta, x) \pm \epsilon]\} < \infty$, where ϵ is as in Assumption 1.2. Two-dimensional vectors are indexed starting with zero, so that, e.g., $b = (b_0, b_1)$, $q(\eta) = (q_0(\eta), q_1(\eta))$.

1.E.1. **Basic lemmas.** I state some auxiliary results which are used throughout the proofs.

Lemma 1.E.1. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then for $j \in \mathbb{N}$ it holds* *that*

$$S_{n,j} \equiv \frac{1}{n}\sum_{i=1}^{n} k_{h,i} X_{h,i}^{j} = \mu_j f_X(x_0) + o_p(1).$$

If additionally x_0 is an interior point, $f_X(x)$ is continuously differentiable, and j is odd, then $S_{n,j} = O_p(h + (nh)^{-1/2})$.

Proof. Standard kernel calculations. □

Lemma 1.E.2. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then for $j \in \mathbb{N}$ it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} k_{h,i} X_{h,i}^{j} \{\mathbb{1}(Y_i \leq Q(\eta, X_i)) - \eta\} = O_p((nh)^{-1/2}).$$

Proof. Standard kernel calculations. □

Lemma 1.E.3. *Suppose that Assumptions 1.1–1.4 hold. Then for $j \in \mathbb{N}$ it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} k_{h,i} X_{h,i}^{j} (Y_i - m(\eta, X_i))\mathbb{1}(Y_i \leq Q(\eta, X_i)) = O_p((nh)^{-1/2}).$$

Proof. Standard kernel calculations. □

Lemma 1.E.4. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then $a^j(\widehat{q}_j(\eta; a) - q_j(\eta)) = O_p(a^2 + (an)^{-1/2})$ for $j \in \{0, 1\}$.*

Proof. The lemma follows, e.g., from Theorem 2 of Fan et al. (1994). It also follows from the proof of Lemma 1.E.10, where I allow for the truncation quantile level to be estimated. □

Lemma 1.E.5. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then*

$$\sup_{x \in \mathcal{X}_h} |\widehat{Q}(\eta, x; a) - Q(\eta, x)| = O_p(w_n),$$

where $w_n = a^2 + h^2 + (a + h)(a^3 n)^{-1/2}$, as defined in Theorem 1.1.

Proof. Using a second-order Taylor expansion of $Q(\eta, x)$ in x with a mean-value form of the remainder and the triangle inequality, I obtain that

$$\begin{aligned}
&\sup_{x \in \mathcal{X}_h} |\widehat{Q}(\eta, x; a) - Q(\eta, x)| \\
&\leq |\widehat{q}_0(\eta; a) - q_0(\eta)| + \sup_{x \in \mathcal{X}_h} |(\widehat{q}_1(\eta; a) - q_1(\eta))(x - x_0)| + \sup_{x, \widetilde{x} \in \mathcal{X}_h} |\frac{1}{2}\partial_x^2 Q(\eta, \widetilde{x})(x - x_0)^2| \\
&= O_p(a^2 + (an)^{-1/2} + h(a + (a^3 n)^{-1/2}) + h^2).
\end{aligned}$$

□

Lemma 1.E.6. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold, and $\widetilde{Q}$ is a, possibly random, function such that $\sup_{x\in\mathcal{X}_h} |\widetilde{Q}(\eta,x) - Q(\eta,x)| = O_p(w_n)$. For $j \in \mathbb{N}$ it holds that:*

(i) $\displaystyle \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}(Y_i - Q(\eta,X_i))\{\mathbb{1}(Y_i \leq \widetilde{Q}(\eta,X_i)) - \mathbb{1}(Y_i \leq Q(\eta,X_i))\} = O_p\left(w_n^2\right)$,

(ii) $\displaystyle \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}(\widetilde{Q}(\eta,X_i) - Q(\eta,X_i))\{\mathbb{1}(Y_i \leq \widetilde{Q}(\eta,X_i)) - \mathbb{1}(Y_i \leq Q(\eta,X_i))\} = O_p\left(w_n^2\right)$,

(iii) $\displaystyle \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}(\mathbb{1}(Y_i \leq Q(\eta,X_i)) - \mathbb{1}(Y_i \leq \widetilde{Q}(\eta,X_i))) = O_p(w_n)$.

Proof. I prove only part (i). Parts (ii) and (iii) follow analogously. The proof is similar to the proof of Lemma A.3 of Kato (2012). For $l > 0$ let

$$\mathcal{M}_n(l) = \{g : \mathcal{X} \to \mathbb{R} \text{ s.t. } \sup_{x\in\mathcal{X}_h} |g(x) - Q(\eta,x)| \leq lw_n\}.$$

For a function $g : \mathcal{X} \to \mathbb{R}$, let

$$U_n(g) \equiv \left|\frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}(Y_i - Q(\eta,X_i))\Big\{\mathbb{1}(Y_i \leq g(X_i)) - \mathbb{1}(Y_i \leq Q(\eta,X_i))\Big\}\right|.$$

It suffices to show that for each fixed $l > 0$

$$\sup_{g\in\mathcal{M}_n(l)} U_n(g) = O_p(w_n^2). \tag{1.E.1}$$

It holds that

$$\begin{aligned} U_n(g) \leq& \frac{1}{n}\sum_{i=1}^{n} k_{h,i}|X_{h,i}^{j}|(Y_i - Q(\eta,X_i))\mathbb{1}(Q(\eta,X_i) < Y_i \leq g(X_i)) \\ &+ \frac{1}{n}\sum_{i=1}^{n} k_{h,i}|X_{h,i}^{j}|(Q(\eta,X_i) - Y_i)\mathbb{1}(g(X_i) < Y_i \leq Q(\eta,X_i)). \end{aligned}$$

Let $U_{n,1}(g)$ and $U_{n,2}(g)$ denote the first and the second element in the above sum, respectively. They are both nonnegative. It holds that

$$\sup_{g\in\mathcal{M}_n(l)} U_{n,1}(g) = \frac{1}{n}\sum_{i=1}^{n} k_{h,i}|X_{h,i}^{j}|(Y_i - Q(\eta,X_i))\mathbb{1}(Q(\eta,X_i) < Y_i \leq Q(\eta,X_i) + lw_n) \equiv \bar{U}_{n,1}.$$

Further,

$$\begin{aligned} E\left[\bar{U}_{n,1}\right] &\leq \mathbb{E}\left[k_h(X - x_0)|X_h^j|lw_n\mathbb{1}(Q(\eta,X) < Y \leq Q(\eta,X) + lw_n)\right] \\ &\leq C_f l^2 w_n^2 \int k_h(x - x_0)f(x)dx = O(w_n^2). \end{aligned}$$

Since $\bar{U}_{n,1}$ is nonnegative, it follows from Markov's inequality that $\bar{U}_{n,1} = O_p(w_n^2)$. Applying the same reasoning to $U_{n,2}(g)$ yields (1.E.1). □

1.E.2. **Proofs of Theorem 1.1 and Corollary 1.1.**

Proof of Theorem 1.1. It holds that

$$\widehat{m}(\eta, x_0; a, h) = \frac{S_{n,2}\Psi_{n,0}(a) - S_{n,1}\Psi_{n,1}(a)}{S_{n,2}S_{n,0} - S_{n,1}^2} \text{ and } \widetilde{m}(\eta, x_0; h) = \frac{S_{n,2}\widetilde{\Psi}_{n,0} - S_{n,1}\widetilde{\Psi}_{n,1}}{S_{n,2}S_{n,0} - S_{n,1}^2},$$

where $\Psi_{n,j}(a) = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\psi_i(\eta, \widehat{Q}(\eta, X_i; a))$, $\widetilde{\Psi}_{n,j} = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\psi_i(\eta, Q(\eta, X_i))$, and $S_{n,j}$ is defined in Lemma 1.E.1. Hence,

$$\widehat{m}(\eta, x_0; a, h) - \widetilde{m}(\eta, x_0; h) = \frac{S_{n,2}(\Psi_{n,0}(a) - \widetilde{\Psi}_{n,0}) - S_{n,1}(\Psi_{n,1}(a) - \widetilde{\Psi}_{n,1})}{S_{n,2}S_{n,0} - S_{n,1}^2}.$$

The denominator converges to a positive number. I consider the numerator. For $j \in \{0, 1\}$, it holds that

$$\begin{aligned}\Psi_{n,j}(a) - \widetilde{\Psi}_{n,j} = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\Big\{&\frac{1}{\eta}Y_i\{\mathbb{1}(Y_i \le \widehat{Q}(\eta, X_i; a)) - \mathbb{1}(Y_i \le Q(\eta, X_i))\} \\ &- \frac{1}{\eta}\widehat{Q}(\eta, X_i; a)\mathbb{1}(Y_i \le \widehat{Q}(\eta, X_i; a)) + \frac{1}{\eta}Q(\eta, X_i)\mathbb{1}(Y_i \le Q(\eta, X_i)) \\ &\pm \frac{1}{\eta}\widehat{Q}(\eta, X_i; a)\mathbb{1}(Y_i \le Q(\eta, X_i)) - (Q(\eta, X_i) - \widehat{Q}(\eta, X_i; a))\Big\} \\ =&\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\Big\{\frac{1}{\eta}(Q(\eta, X_i) - \widehat{Q}(\eta, X_i; a))\{\mathbb{1}(Y_i \le Q(\eta, X_i)) - \eta\}\Big\} + O_p(w_n^2),\end{aligned}$$

where the last equality follows from Lemma 1.E.6. Further,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\Big\{\frac{1}{\eta}(Q(\eta, X_i) - \widehat{Q}(\eta, X_i; a))\{\mathbb{1}(Y_i \le Q(\eta, X_i)) - \eta\}\Big\} \\ &\quad = \frac{1}{\eta}(q_0(\eta) - \widehat{q}_0(\eta; a))\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\{\mathbb{1}(Y_i \le Q(\eta, X_i)) - \eta\} \\ &\quad + \frac{1}{\eta}h(q_1(\eta) - \widehat{q}_1(\eta; a))\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^{j+1}\{\mathbb{1}(Y_i \le Q(\eta, X_i)) - \eta\} \\ &\quad + \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\frac{1}{\eta}(Q(\eta, X_i) - q_0(\eta) - q_1(\eta)(X_i - x_0))\{\mathbb{1}(Y_i \le Q(\eta, X_i)) - \eta\}.\end{aligned}$$

Let L_1, L_2, and L_3 denote the three terms above. By Lemmas 1.E.2 and 1.E.4, it holds that $L_1 = O_p(a^2 + (na)^{-1/2})O_p((nh)^{-1/2})$ and $L_2 = h/aO_p(a^2 + (na)^{-1/2})O_p((nh)^{-1/2})$. Moreover, $\mathbb{E}[L_3|\mathcal{X}_n] = 0$ and $\mathrm{Var}(L_3|\mathcal{X}_n) = O_p(h^4(nh)^{-1})$, which implies that $L_3 = O_p(h^2(nh)^{-1/2})$. In total,

$$\begin{aligned}\Psi_{n,j}(a) - \widetilde{\Psi}_{n,j} &= O_p(a^2 + (na)^{-1/2} + h(a + (a^3n)^{-1/2}) + h^2)O_p((nh)^{-1/2}) + O_p(w_n^2) \\ &= O_p(w_n(nh)^{-1/2} + w_n^2),\end{aligned}$$

which concludes the proof. □

Remark 1.1. In the proof of Theorem 1.1, I do not explicitly use the orthogonality condition, as stated in equation (1.1.5). However, this property is the reason why the terms with $\widehat{q}_0(\eta; a)$ and $\widehat{q}_1(\eta; a)$ are negligible in the expansion of $\Psi_{n,j}(a) - \widetilde{\Psi}_{n,j}$. Note that

$$\frac{d}{dg}\mathbb{E}[Y\mathbb{1}(Y \leq g) - g(\mathbb{1}(Y \leq g) - \eta)|X = x_0] = -\mathbb{E}[\mathbb{1}(Y \leq g) - \eta|X = x_0],$$

which evaluated at $g = Q(\eta, x_0)$ is zero.

Proof of Corollary 1.1. First, I show that under the assumptions made on the bandwidths, the remainder in Theorem 1.1 is of order $o_p(h^2 + (nh)^{-1/2})$. Recall that $w_n = a^2 + h^2 + (a + h)(a^3n)^{-1/2}$. By Assumption 1.4(b), it holds that

$$\begin{aligned}O_p(w_n(nh)^{-1/2} + w_n^2) &= O_p\left(w_n(nh)^{-1/2} + a^4 + h^4 + (a^2 + h^2)(a^3n)^{-1}\right)\\ &= O_p\left(h(a^3n)^{-1/2}(nh)^{-1/2} + a^4 + (an)^{-1} + h^2(a^3n)^{-1}\right) + o_p(h^2 + (nh)^{-1/2}).\end{aligned}$$

The following equivalence statements hold

$$\begin{aligned}h^2/(a^3n) \to 0 &\iff (nh)^{-1}h \prec a, & a^4/h^2 \to 0 &\iff a \prec \sqrt{h},\\ (nh)^{1/2}/(an) \to 0 &\iff (nh)^{-1/2}h \prec a, & (nh)^{1/2}h^2/(a^3n) \to 0 &\iff (nh)^{-1/6}h \prec a.\end{aligned}$$

The conditions on the right-hand sides hold under the assumptions made.

The lemma follows from standard theory applied to the infeasible estimator $\widetilde{m}(\eta, x_0; h)$. The variance is derived as follows

$$\begin{aligned}\mathrm{Var}(\psi(\eta, Q(\eta, X))|X = x_0) &= \mathbb{E}\left[(\psi(\eta, Q(\eta, X)) - m(\eta, x_0))^2 \,|X = x_0\right]\\ &= \mathbb{E}\left[\left(\frac{1}{\eta}(Y - m(\eta, X))\mathbb{1}(Y \leq Q(\eta, X))\right.\right.\\ &\qquad \left.\left.- \frac{1}{\eta}(Q(\eta, X) - m(\eta, X))(\eta - \mathbb{1}(Y \leq Q(\eta, X)))\right)^2 \middle| X = x_0\right]\\ &= \frac{1}{\eta}\mathrm{Var}(Y|Y \leq Q(\eta, X), X = x_0) + \frac{(1-\eta)}{\eta}(Q(\eta, x_0) - m(\eta, x_0))^2.\end{aligned}$$

□

1.E.3. **Proof of Theorem 1.2.** The main burden of the proof lies in studying the properties of the local linear quantile estimator with estimated quantile level, $\widehat{q}(\widehat{\eta}; h)$. In Lemmas 1.E.7 and 1.E.10, I show that, under the assumptions made, it has the same rate of convergence as the local linear quantile estimator with a known quantile level.

In the proof, I use the bandwidth-dependent estimand of the local linear quantile estimator defined as:

$$(q_0^*(u;h), q_1^*(u;h))^\top = \operatorname*{arg\,min}_{(b_0,b_1)\in\mathbb{R}^2} \mathbb{E}\left[\rho_u(Y_i - b_0 - b_1(X - x_0))k(X_h)\right]. \tag{1.E.2}$$

Further, I put $Q^*(u,x;h) = q_0^*(u;h) + q_1^*(u;h)(x - x_0)$.

Lemma 1.E.7. *Suppose that the assumptions of Theorem 1.2 hold. Then for $j \in \{0,1\}$,*

$$(i)\; h^j(q_j^*(\eta;h) - q_j(\eta)) = O(h^2) \text{ and } (ii)\; h^j(q_j^*(\widehat{\eta};h) - q_j^*(\eta;h)) = O(h^2) + O_p(v_n).$$

Proof. This lemma follows from derivations of Guerre and Sabbah (2012).[22] I outline only the main steps. It follows from the proof of their Theorem 1 that $h^j(q_j^*(u;h) - q_j(u)) = O(h^2)$ uniformly in u over some sufficiently small neighborhood of η. Part (i) follows.

Further, the first-order condition of the population minimization problem in 1.E.2 is

$$\mathbb{E}\left[k_h(X - x_0)\widetilde{X}_h\{\mathbb{1}(Y \le Q^*(u,X;h)) - u\}\right] = 0.$$

Using the implicit function theorem and continuity of $f_{Y|X}(y|x)$, it follows that $h^j q_j^*(u;h)$ is continuously differentiable with

$$\begin{bmatrix} \partial_u^1 q_0^*(u;h) \\ h\partial_u^1 q_1^*(u;h) \end{bmatrix} = \mathbb{E}\left[k_h(X - x_0) f_{Y|X}(Q^*(u,X;h)|X)\widetilde{X}_h\widetilde{X}_h^\top\right]^{-1} \mathbb{E}\left[k_h(X - x_0)\widetilde{X}_h\right],$$

which is bounded uniformly over u in a sufficiently small neighborhood of η. Hence, part (ii) follows using the mean value theorem. □

Next, I prove two equicontinuity results, which are then used to show convergence of the criterion function of the local linear quantile estimator with an estimated quantile level. I introduce the following additional notation. Let $v_n = (nh)^{-1/2}$, $\mathcal{M}_n(q,l) = \{b : |b_0 - q_0| \le l_0 v_n \text{ and } h|b_1 - q_1| \le l_1 v_n\}$. For a vector $l = (l_0, l_1)^\top$, I put $|l| \equiv ||l||_1 = |l_0| + |l_1|$.

Lemma 1.E.8. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Let $A_{i,n} = v_n\widetilde{X}_{h,i}^\top\theta$ for some θ and*

$$\begin{aligned} T(b) &= \sum_{i=1}^n k(X_{h,i})(Y_i'(b) - A_{i,n})\{\mathbb{1}(Y_i'(b) \le A_{i,n}) - \mathbb{1}(Y_i'(b) \le 0)\}, \\ \bar{T}(b) &= T(b) - \mathbb{E}[T(b)]. \end{aligned}$$

For any sequence $q_n \to q(\eta)$ and constant $M > 0$ it holds that

$$\sup_{b\in\mathcal{M}_n(q_n,M)} |\bar{T}(b)| = o_p(1).$$

[22] Their derivations are more involved as they provide convergence results uniform in the evaluation point, bandwidth, and quantile level. In my setting, x_0 is fixed, and h is a fixed sequence.

Proof. I will show that *(i)* $\bar{T}(q_n) = o_p(1)$ and *(ii)* $\sup_{b\in\mathcal{M}_n(q_n,M)} |\bar{T}(b) - \bar{T}(q_n)| = o_p(1)$.
Part (i). Note that

$$T(b) = \sum_{i=1}^{n} k(X_{h,i})(Y_i'(b) - A_{i,n})\{\mathbb{1}(0 < Y_i'(b) \leq A_{i,n}) - \mathbb{1}(A_{i,n} < Y_i'(b) \leq 0)\}.$$

Using the fact that $f_{Y|X}(y|x)$ is bounded over (x, y) in a sufficiently small neighborhood of $(x_0, Q(\eta, x_0))$, I obtain that

$$\mathrm{Var}(T(q_n)) \leq \sum_{i=1}^{n} \mathbb{E}\Big[k(X_{h,i})^2 A_{i,n}^2 \mathbb{1}(-|A_{i,n}| < Y_i'(q_n) \leq |A_{i,n}|)\Big] = O(nhv_n^3) = o(1).$$

Hence, $\bar{T}(q_n) = o_p(1)$.
Part (ii). I follow the lines of the proof of Lemma 4.1 of Bickel (1975). A similar claim has been shown by Ruppert and Carroll (1978, Lemma A.4). Let $\Delta_i(q, b) \equiv Y_i'(q) - Y_i'(b) = L_i(b - q)$. It holds that

$$\begin{aligned}
T(q) - T(b) = &\sum_{i=1}^{n} k(X_{h,i})\Big[(Y_i'(q) - Y_i'(b))\{\mathbb{1}(0 < Y_i'(q) \leq A_{i,n}) - \mathbb{1}(A_{i,n} < Y_i'(q) \leq 0)\} \\
&+ (Y_i'(b) - A_{i,n})\{\mathbb{1}(Y_i'(q) \leq A_{i,n}) - \mathbb{1}(Y_i'(q) \leq 0) - \mathbb{1}(Y_i'(b) \leq A_{i,n}) + \mathbb{1}(Y_i'(b) \leq 0)\}\Big] \\
= &\sum_{i=1}^{n} k(X_{h,i})\Big[\Delta_i(q, b)\{\mathbb{1}(0 < Y_i'(q) \leq A_{i,n}) - \mathbb{1}(A_{i,n} < Y_i'(q) \leq 0)\} \\
&+ (Y_i'(q) - A_{i,n} - \Delta_i(q, b)) \\
&\times \{\mathbb{1}(\Delta_i(q, b) < Y_i'(q) - A_{i,n} \leq 0) - \mathbb{1}(0 < Y_i'(q) - A_{i,n} \leq \Delta_i(q, b))\} \\
&+ (Y_i'(q) - A_{i,n} - \Delta_i(q, b))\{\mathbb{1}(0 < Y_i'(q) \leq \Delta_i(q, b)) - \mathbb{1}(\Delta_i(q, b) < Y_i'(q) \leq 0)\}\Big].
\end{aligned}$$

For $l = (l_0, l_1)$, let $b_{n,0}(l) = q_{n,0} + l_0 v_n$ and $b_{n,1}(l) = q_{n,1} + l_1 v_n/h$. Note that for $X_i \in \mathcal{X}_h$, it holds that $|\Delta_i(q_n, b_n(l))| \leq v_n|l|$. Therefore,

$$\begin{aligned}
\mathrm{Var}(T(b_n(l)) - T(q_n)) \leq &3\sum_{i=1}^{n} \mathbb{E}\Big[k(X_{h,i})^2(v_n|l|)^2\mathbb{1}(-|A_{i,n}| < Y_i'(q_n) \leq |A_{i,n}|) \\
&+ k(X_{h,i})^2(v_n|l|)^2\mathbb{1}(-v_n|l| < Y_i'(q_n) - A_{i,n} \leq v_n|l|) \\
&+ k(X_{h,i})^2(v_n|l| + |A_{i,n}|)^2\mathbb{1}(-v_n|l| < Y_i'(q_n) \leq v_n|l|)\Big] \\
= &O(nhv_n^3).
\end{aligned}$$

Hence, for any fixed l,

$$\bar{T}(b_n(l)) - \bar{T}(q_n) = o_p(1). \tag{1.E.3}$$

For a fixed $\delta > 0$ decompose $\mathcal{M}_n(q_n, M)$ as the union of cubes with vertices on the grid $J_n(\delta) = \{q_n + \delta M v_n(j_0, j_1/h)^\top : j_i \in \{0, \pm 1, ..., \pm\lceil 1/\delta\rceil\}$ for $i = 0, 1\}$, where $\lceil\cdot\rceil$ is the ceiling function. For $b \in \mathcal{M}_n(q_n, M)$, let $V_n(b)$ be the lowest vertex of the cube containing b. The result in (1.E.3) implies that

$$\max\left\{|\bar{T}(V_n(b)) - \bar{T}(q_n)| : b \in \mathcal{M}_n(q_n, M)\right\} = o_p(1).$$

Next, I consider the behavior on a cube. Note that for $X_i \in \mathcal{X}_h$, it holds that

$$\sup\{|\Delta_i(V_n(b), b)| : b \in \mathcal{M}_n(V_n(b), \delta M)\} = 2\delta M v_n.$$

Further,

$$\begin{aligned}|T(V_n(b)) - T(b)| \leq \sum_{i=1}^n k(X_{h,i})\{&2\delta M v_n \mathbb{1}(-|A_{i,n}| < Y_i'(V_n(b)) \leq |A_{i,n}|) \\ &+ 2\delta M v_n \{\mathbb{1}(-2\delta M v_n \leq Y_i'(V_n(b)) - A_{i,n} \leq 2\delta M v_n) \\ &+ (2\delta M v_n + |A_{i,n}|)\mathbb{1}(-2\delta M v_n \leq Y_i'(V_n(b)) \leq 2\delta M v_n)\} \\ \equiv \widetilde{T}&(V_n(b), \delta).\end{aligned}$$

The reasoning leading to (1.E.3) yields also that

$$\max_{b \in J_n(\delta)} |\widetilde{T}(b, \delta) - \mathbb{E}[\widetilde{T}(b, \delta)]| = o_p(1).$$

Moreover,

$$\max_{b \in J_n(\delta)} \mathbb{E}[\widetilde{T}(b, \delta)] \leq \delta O(1).$$

uniformly in $\delta \in (0, 1)$, which concludes the proof. □

Lemma 1.E.9. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Let*

$$\begin{aligned}S(b) &= \frac{1}{\sqrt{nh}} \sum_{i=1}^n k(X_{h,i}) X_{h,i}^j \mathbb{1}(Y_i'(b) \leq 0), \\ \bar{S}(b) &= S(b) - \mathbb{E}[S(b)].\end{aligned}$$

For any sequence $q_n \to q(\eta)$ and constant $M > 0$ it holds that

$$\sup_{b \in \mathcal{M}_n(q_n, M)} |\bar{S}(b) - \bar{S}(q_n)| = o_p(1) \text{ and } |\bar{S}(q_n) - \bar{S}(q(\eta))| = o_p(1).$$

Proof. The proof is similar to the proof of Lemma 1.E.8. I am using the notation defined therein. I note that

$$S(q) - S(b) = \frac{1}{\sqrt{nh}} \sum_{i=1}^n k(X_{h,i}) X_{h,i}^j \{\mathbb{1}(Y_i'(q) \leq \Delta_i(q, b)) - \mathbb{1}(Y_i'(q) \leq 0)\}.$$

It holds that $\mathrm{Var}(S(q_n) - S(q(\eta))) = o(1)$. Hence, the second claim follows.

For any fixed l it holds that

$$\mathrm{Var}(S(b_n(l)) - S(q_n)) = O_p(v_n) = o_p(1).$$

Hence,

$$\max\left\{|\bar{S}(V_n(b)) - \bar{S}(q_n)| : b \in \mathcal{M}_n(q_n, M)\right\} = o_p(1).$$

Moreover,

$$\begin{aligned}|S(V_n(b)) - S(b)| &\leq \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k(X_{h,i})|X_{h,i}|\mathbb{1}(-2\delta M v_n \leq Y_i'(V_n(b)) \leq 2\delta M v_n) \\ &\equiv \widetilde{S}(V_n(b), \delta)\end{aligned}$$

It holds

$$\max_{b \in J(\delta)} |\widetilde{S}(b, \delta) - \mathbb{E}[\widetilde{S}(b, \delta)]| = O_p(v_n) = o_p(1).$$

Finally,

$$\max_{b \in J(\delta)} \mathbb{E}[\widetilde{S}(b, \delta)] \leq \delta O_p(1).$$

uniformly in δ, which concludes the proof. □

Lemma 1.E.10. *Suppose that the assumptions of Theorem 1.2 hold. Then for $j \in \{0, 1\}$, it holds that $h^j(\widehat{q}_j(\widehat{\eta}; h) - q_j(\widehat{\eta}; h)) = O_p((nh)^{-1/2})$.*

Proof. Recall that $\rho_u(v) = v(u - \mathbb{1}(v \leq 0))$ and

$$\widehat{q}(u; h) = \underset{(b_0, b_1) \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \rho_u(Y_i - b_0 - b_1(X_i - x_0))k(X_{h,i}).$$

Let $\widehat{\theta}_n(u) = \sqrt{nh}(\widehat{q}_0(u; h) - q_0^*(u; h),\, h(\widehat{q}_1(u; h) - q_1^*(u; h)))^\top$. For a given u, the vector $\widehat{\theta}_n(u)$ minimizes the function

$$G_n(u, \theta) = \sum_{i=1}^{n} \left[\rho_u(Y_i^*(u; h) - v_n \theta^\top \widetilde{X}_{h,i}) - \rho_u(Y_i^*(u; h))\right] k(X_{h,i}),$$

where $Y_i^*(u; h) = Y_i - Q^*(u, X_i; h)$. Let

$$\begin{aligned}W_n(u) &= v_n \sum_{i=1}^{n} k(X_{h,i})\widetilde{X}_{h,i}\{u - \mathbb{1}(Y_i^*(u; h) \leq 0)\}, \\ T_n(u, \theta) &= -\sum_{i=1}^{n} k(X_{h,i})(Y_i^*(u; h) - v_n \theta^\top \widetilde{X}_{h,i}) \\ &\quad \times \left\{\mathbb{1}(Y_i^*(u; h) - v_n \theta^\top \widetilde{X}_{h,i} < 0) - \mathbb{1}(Y_i^*(u; h) < 0)\right\}.\end{aligned}$$

It holds that $G_n(u,\theta) = T_n(u,\theta) - \theta^\top W_n(u)$. Further,

$$\begin{aligned}
\mathbb{E}[T_n(u,\theta)|X_1, ..., X_n] &= -\sum_{i=1}^n k(X_{h,i}) \int_0^{v_n\theta^\top \widetilde{X}_{h,i}} \left(y - v_n\theta^\top \widetilde{X}_{h,i}\right) f_{Y^*(u)|X}(y|X_i)dy \\
&= \frac{1}{2}\sum_{i=1}^n k(X_{h,i}) f_{Y^*(u;h)|X}(\widetilde{z}_i(u)|X_i)(v_n\theta^\top \widetilde{X}_{h,i})^2 \\
&= \frac{1}{2n}\sum_{i=1}^n k_{h,i}(\theta^\top \widetilde{X}_{h,i})^2 (f_{Y|X}(q_0(u)|x_0) + \xi_{i,n}),
\end{aligned}$$

where $\widetilde{z}_i(u)$ lies between 0 and $v_n\theta^\top \widetilde{X}_{h,i}$, and $\xi_{i,n} = o(1)$ uniformly in $i \in \{1, ..., n\}$ and u in a sufficiently small neighborhood of η. Hence, it follows from Lemma 1.E.8 that

$$T_n(\widehat{\eta},\theta) = \theta^\top S\theta + o_p(1),$$

where

$$S = f_{Y|X}(q_0(\eta)|x_0) f_X(x_0) \begin{bmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{bmatrix}.$$

The convex, random function $\widehat{T}_n(\theta) \equiv T_n(\widehat{\eta},\theta)$ converges pointwise in θ to the convex function $\theta^\top S\theta$. By the convexity lemma (Pollard, 1991), this convergence is uniform on any compact set. The function $\frac{1}{2}\theta^\top S\theta - \theta^\top W_n(\widehat{\eta})$ is minimized at $S^{-1}W_n(\widehat{\eta})$. Since by construction $\mathbb{E}[W_n(u)] = 0$, Lemma 1.E.9 implies that

$$W_n(\widehat{\eta}) = W_n(\eta) + o_p(1) = O_p(1).$$

Using convexity again, the consistency argument of Pollard (1991) implies that $\widehat{\theta}_n(\widehat{\eta}) = S^{-1}W_n(\widehat{\eta}) + o_p(1)$, which concludes the proof. □

Proof of Theorem 1.2. Since $a \asymp h$, $w_n = h^2 + (nh)^{-1/2} \equiv r_n$. By Lemmas 1.E.7 and 1.E.10, $\widehat{q}(\widehat{\eta}; a)$ has the same rate of convergence as $\widehat{q}(\eta; a)$ has. Hence, the proof of Theorem 1.1 immediately implies that

$$\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \psi_i(\eta, \widehat{Q}(\widehat{\eta}, X_i; a)) - \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \psi_i(\eta, Q(\eta, X_i)) = O_p(r_n^2)$$

for $j \in \{0, 1\}$. Moreover,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \psi_i(\widehat{\eta}, \widehat{Q}(\widehat{\eta}, X_i; a)) - \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \psi_i(\eta, \widehat{Q}(\widehat{\eta}, X_i; a)) \\
&\qquad = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j (Y_i - \widehat{Q}(\widehat{\eta}, X_i; a))\mathbb{1}(Y_i \leq \widehat{Q}(\widehat{\eta}, X_i; a)) \left(\frac{1}{\widehat{\eta}} - \frac{1}{\eta}\right) \\
&\qquad = \left(\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j (Y_i - Q(\eta, X_i))\mathbb{1}(Y_i \leq Q(\eta, X_i)) + O_p(r_n)\right) \left(\frac{1}{\widehat{\eta}} - \frac{1}{\eta}\right),
\end{aligned}$$

where the second equality follows from Lemma 1.E.6 and the convergence rate of $\widehat{q}(\widehat{\eta}; a)$. Further, using the convergence rate of the local linear estimator, it follows that

$$\widehat{m}(\widehat{\eta}, x_0; a, h) = \widetilde{m}(\eta, x_0; h) + \eta(m(\eta, x_0) - Q(\eta, x_0) + O_p(r_n))\left(\frac{1}{\widehat{\eta}} - \frac{1}{\eta}\right).$$

The proof is concluded by noting that

$$\left(\frac{1}{\widehat{\eta}} - \frac{1}{\eta}\right) = -\frac{1}{\eta^2}(\widehat{\eta} - \eta) + O_p(r_n^2).$$

□

1.F. PROOFS OF THE RESULTS IN THE APPENDIX

1.F.1. **Proofs of Theorems 1.A.1 and 1.A.2.** These proofs are very similar to the proof of Theorem 1.1 and are therefore omitted.

1.F.2. **Proofs of Propositions 1.B.1 and 1.B.3.** Under the assumptions of these propositions, $a = h$, and hence $w_n = h^2 + (nh)^{-1/2} \equiv r_n$.

An essential result used to prove these two propositions, not required for the proof of Theorem 1.1, are the following approximate first-order conditions of the local linear quantile estimator.

Lemma 1.F.1. *Suppose that Assumptions 1.1 and 1.4 hold. Then for $j \in \{0, 1\}$ it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} k_{h,i} X_{h,i}^{j}(\eta - \mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h))) = O_p((nh)^{-1}).$$

Proof. Similar claims have been proven by Koenker and Bassett Jr (1978, Theorem 3.3) and Ruppert and Carroll (1980, Theorem 1). Let

$$G_n(b) = \frac{1}{n}\sum_{i=1}^{n} k_{h,i}\rho_\eta(Y_i'(b)),$$

where $\rho_\eta(v) = v[\eta - \mathbb{1}(v \leq 0)]$. It holds that $\partial_v^+ \rho_\eta(v) = \eta - \mathbb{1}(v < 0)$ and $\partial_v^- \rho_\eta(v) = \eta - \mathbb{1}(v \leq 0)$. Therefore, also the left and right derivatives of the criterion function exist. For $j \in \{0, 1\}$ it holds that

$$\partial_{b_j}^+ G_n(b) = \frac{h}{n}\sum_{i=1}^{n} k_{h,i} X_{h,i}^{j}\left((\mathbb{1}(Y_i'(b) < 0) - \eta)\mathbb{1}(X_{h,i}^{j} < 0) + (\mathbb{1}(Y_i'(b) \leq 0) - \eta)\mathbb{1}(0 < X_{h,i}^{j})\right),$$

$$\partial_{b_j}^- G_n(b) = \frac{h}{n}\sum_{i=1}^{n} k_{h,i} X_{h,i}^{j}\left((\mathbb{1}(Y_i'(b) \leq 0) - \eta)\mathbb{1}(X_{h,i}^{j} < 0) + (\mathbb{1}(Y_i(b) < 0) - \eta)\mathbb{1}(0 < X_{h,i}^{j})\right).$$

At the minimum, it holds that $\partial_{b_j}^- G_n(\widehat{q}(\eta)) \leq 0 \leq \partial_{b_j}^+ G_n(\widehat{q}(\eta))$. Using these inequalities, I obtain the following bounds on the expression of interest:

$$\begin{aligned}
0 &\leq \frac{h}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\Big\{\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) - \eta - \mathbb{1}(Y_i = \widehat{Q}(\eta, X_i; h))\mathbb{1}(X_{h,i}^{j} < 0)\Big\} \\
&\leq \partial_{b_j}^{+} G_n(\widehat{q}(\eta)) - \partial_{b_j}^{-} G_n(\widehat{q}(\eta)) \\
&= \frac{h}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\Big\{\mathbb{1}(Y_i = \widehat{Q}(\eta, X_i; h))\mathbb{1}(0 \leq X_{h,i}^{j}) - \mathbb{1}(Y_i = \widehat{Q}(\eta, X_i; h))\mathbb{1}(X_{h,i}^{j} < 0)\Big\}.
\end{aligned}$$

The lemma follows from the facts that k is bounded with bounded support and

$$\sum_{i=1}^{n} \mathbb{1}(Y_i = \widehat{Q}(\eta, X_i; h)) \leq 2 \text{ w.p. } 1$$

because the probability of having three collinear points in a sample is equal zero. □

Proof of Proposition 1.B.1. It holds that

$$\widehat{m}^{NM}(\eta, x_0; h, h) - \widehat{m}(\eta, x_0; h, h) = \frac{S_{n,2}(T_{n,0} - \Psi_{n,0}(h)) - S_{n,1}(T_{n,1} - \Psi_{n,1}(h))}{S_{n,2}S_{n,0} - S_{n,1}^2}$$

where $T_{n,j} = \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}Y_i\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h))$, and $\Psi_{n,j}(h)$ is defined in the proof of Theorem 1.1. From Lemma 1.F.1 it immediately follows that

$$\begin{aligned}
T_{n,0} - \Psi_{n,0}(h) &= O_p((nh)^{-1}), \\
T_{n,1} - \Psi_{n,1}(h) &= \frac{1}{\eta}\widehat{q}_1(\eta; h)\frac{h}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^2(\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) - \eta) + O_p((nh)^{-1}) \\
&= \frac{1}{\eta}q_1(\eta)\frac{h}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^2(\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) - \eta) + O_p(r_n^2).
\end{aligned}$$

Hence,

$$\widehat{m}^{NM}(\eta, x_0; h, h) - \widehat{m}(\eta, x_0; h, h) = hS_{n,1}q_1(\eta)O_p(r_n) + O_p(r_n^2),$$

which, combined with Lemma 1.E.1, concludes the proof. □

Proof of Proposition 1.B.3. It holds that

$$\widehat{m}^{TS}(\eta, x_0; h, h) = \frac{\widehat{S}_{n,2}T_{n,0} - \widehat{S}_{n,1}T_{n,1}}{\widehat{S}_{n,2}\widehat{S}_{n,0} - \widehat{S}_{n,1}^2},$$

where $\widehat{S}_{n,j} = \frac{1}{\eta n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h))$, and $T_{n,j}$ is defined in the proof of Proposition 1.B.1. It holds that

$$\widehat{S}_{n,2}\widehat{S}_{n,0} - \widehat{S}_{n,1}^2 = S_{n,2}S_{n,0} - S_{n,1}^2 + O_p(r_n).$$

Let $m^*(\eta, x) = m(\eta, x_0) + \partial_x^1 m(\eta, x_0)(x - x_0)$. By plugging in the expression $Y_i =$

$m^*(\eta, X_i) + (Y_i - m^*(\eta, X_i))$ in the definition of $\widehat{m}^{TS}(\eta, x_0; h, h)$, it follows that

$$\widehat{m}^{TS}(\eta, x_0; h, h) = m(\eta, x_0) + \frac{\widehat{S}_{n,2}U_{n,0} - \widehat{S}_{n,1}U_{n,1}}{\widehat{S}_{n,2}\widehat{S}_{n,0} - \widehat{S}_{n,1}^2},$$

where $U_{n,j} = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \frac{1}{\eta}(Y_i - m^*(\eta, X_i))\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h))$.

Lemma 1.E.6 yields that for $j \in \{0, 1\}$

$$\begin{aligned} U_{n,j} &= \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \frac{1}{\eta}(Y_i - Q(\eta, X_i))\mathbb{1}(Y_i \leq Q(\eta, X_i)), \\ &+ \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \frac{1}{\eta}(Q(\eta, X_i) - m^*(\eta, X_i))\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) + O_p(r_n^2). \end{aligned}$$

Moreover, by Lemma 1.F.1 and a minor modification of Lemma 1.E.6 to handle $Q(\eta, X_i) - Q^*(\eta, X_i)$ it holds that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n k_{h,i}\frac{1}{\eta}(Q(\eta, X_i) - m^*(\eta, X_i))\{\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) - \eta\} = O_p(r_n^2), \\ &\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}\frac{1}{\eta}(Q(\eta, X_i) - m^*(\eta, X_i))\{\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) - \eta\} \\ &\quad = \frac{1}{\eta}h(\partial_x^1 Q(\eta, x_0) - \partial_x^1 m(\eta, x_0))\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^2\{\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; h)) - \eta\} + O_p(r_n^2) \\ &\quad = h(\partial_x^1 Q(\eta, x_0) - \partial_x^1 m(\eta, x_0))O_p(r_n) + O_p(r_n^2). \end{aligned}$$

Hence,

$$\begin{aligned} U_{n,0} &= \frac{1}{n}\sum_{i=1}^n k_{h,i}\frac{1}{\eta}(Y_i - Q(\eta, X_i))\mathbb{1}(Y_i \leq Q(\eta, X_i)) + Q(\eta, X_i) - m^*(\eta, X_i) + O_p(r_n^2), \\ U_{n,1} &= \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}\frac{1}{\eta}(Y_i - Q(\eta, X_i))\mathbb{1}(Y_i \leq Q(\eta, X_i)) + Q(\eta, X_i) - m^*(\eta, X_i) \\ &\quad + h(\partial_x^1 Q(\eta, x_0) - \partial_x^1 m(\eta, x_0))O_p(r_n) + O_p(r_n^2). \end{aligned}$$

In particular, $U_{n,j} = O_p(r_n)$, and hence

$$\begin{aligned} \widehat{m}^{TS}(\eta, x_0; h, h) &= m(\eta, x_0) + \frac{S_{n,2}U_{n,0} - S_{n,1}U_{n,1}}{S_{n,2}S_{n,0} - S_{n,1}^2} \\ &= \widetilde{m}(\eta, x_0; h) + hS_{n,1}(\partial_x^1 Q(\eta, x_0) - \partial_x^1 m(\eta, x_0))O_p(r_n) + O_p(r_n^2), \end{aligned}$$

which, combined with Lemma 1.E.1, concludes the proof. □

1.F.3. **Proofs of Propositions 1.B.2 and 1.B.4.** To prove these propositions, I need an explicit expansion of the estimators in the coefficients defining the trimming function.

Lemma 1.F.2. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then*

$$\widehat{q}_0(\eta;a) - q_0(\eta) = \frac{1}{2}\mu\partial_x^2 Q(\eta,x_0)a^2 + \frac{\frac{1}{n}\sum_{i=1}^n k_{a,i}(\mu_2 - \mu_1 X_{a,i})[\eta - \mathbb{1}(Y_i \leq Q(\eta,X_i))]}{f_{Y|X}(q_0(\eta)|x_0)f(x_0)(\mu_2\mu_0 - \mu_1^2)}$$
$$+ o(a^2) + o_p((na^{-1/2})).$$

Proof. This representation follows from the proof of Theorem 2 of Fan et al. (1994). □

Lemma 1.F.3. *Suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then for $j \in \mathbb{N}$ it holds that*

$$\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \mathbb{1}(Y_i \leq \widehat{Q}(\eta,X_i;a)) = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \mathbb{1}(Y_i \leq Q^*(\eta,X_i))$$
$$+ \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j f_{Y|X}(Q(\eta,x_0)|x_0)\{\widehat{q}_0(\eta;a) - q_0(\eta) + (\widehat{q}_1(\eta;a) - q_1(\eta))(X_i - x_0)\} + o_p(r_n).$$

Proof. A conditional version of Lemma 1.E.9 implies that

$$\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \Big\{\mathbb{1}(Y_i \leq \widehat{Q}(\eta,X_i)) - \mathbb{E}[\mathbb{1}(Y \leq L_i(b)|X = X_i]\Big|_{b=\widehat{q}(\eta)}\Big\}$$
$$= \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j \Big\{\mathbb{1}(Y_i \leq Q^*(\eta,X_i)) - \mathbb{E}[\mathbb{1}(Y \leq Q^*(\eta,X))|X = X_i]\Big\} + o_p(r_n)$$

The result follows by a Taylor expansion using continuity of $f_{Y|X}(y|x)$. □

Proof of Proposition 1.B.2. Part (i). The result is an application of standard asymptotic theory for local linear estimation, using the fact that

$$\mathbb{E}\left[\left(\frac{1}{\eta}Y\mathbb{1}(Y \leq Q(\eta,X)) - m(\eta,X)\right)^2 |X = x_0\right]$$
$$= \mathbb{E}\left[\left(\frac{1}{\eta}(Y - m(\eta,X))\mathbb{1}(Y \leq Q(\eta,X)) - \frac{1}{\eta}m(\eta,X)(\eta - \mathbb{1}(Y \leq Q(\eta,X)))\right)^2 |X = x_0\right]$$
$$= \frac{1}{\eta}\mathrm{Var}(Y|Y \leq Q(\eta,X), X = x_0) + \frac{(1-\eta)}{\eta}m(\eta,x_0)^2.$$

Part (ii). It holds that

$$\widehat{m}^{NM}(\eta,x_0;a,h) = \frac{S_{n,2}T_{n,0}(a) - S_{n,1}T_{n,1}(a)}{S_{n,2}S_{n,0} - S_{n,1}^2}$$

where $T_{n,j}(a) = \frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\frac{1}{\eta}Y_i\mathbb{1}(Y_i \leq \widehat{Q}(\eta,X_i;a))$. I consider the numerator. First, note that:

$$\begin{aligned}T_{n,j}(a) &= \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}(Y_i - Q^*(\eta, X_i))\mathbb{1}(Y_i \leq Q^*(\eta, X_i)) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}Q^*(\eta, X_i)\mathbb{1}(Y_i \leq \widehat{Q}(\eta, X_i; a)) + O_p(r_n^2)\end{aligned}$$

Further, using Lemma 1.E.9,

$$\begin{aligned}T_{n,j}(a) &= \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}Y_i\mathbb{1}(Y_i \leq Q^*(\eta, X_i)) \\ &+ \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j} f_{Y|X}(q_0(\eta)|x_0)\frac{1}{\eta}q_0(\eta)\{\widehat{q}_0(\eta; a) - q_0(\eta) + (\widehat{q}_1(\eta; a) - q_1(\eta))(X_i - x_0)\} \\ &+ \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j} f_{Y|X}(q_0(\eta)|x_0)\frac{1}{\eta}q_1(\eta)(X_i - x_0) \\ &\times \{\widehat{q}_0(\eta; a) - q_0(\eta) + (\widehat{q}_1(\eta; a) - q_1(\eta))(X_i - x_0)\} + o_p(r_n).\end{aligned}$$

The last term is of order $O_p(r_n h)$. Let $u_i^*(\eta) = \frac{1}{\eta}Y_i\mathbb{1}(Y_i \leq Q^*(\eta, X_i)) - m^*(\eta, X_i)$, $e_i^*(\eta) = \frac{1}{\eta}\{\eta - \mathbb{1}(Y_i \leq Q^*(\eta, X_i))\}$, and

$$E_{n,j}(a, h) = \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}u_i^*(\eta) + \frac{1}{n}\sum_{i=1}^{n} k_{a,i}X_{a,i}^{j}\frac{1}{\eta}q_0(\eta)e_i^*(\eta).$$

It follows that

$$\widehat{m}^{NM}(\eta, x_0; a, h) = m(\eta, x_0) + \frac{\mu_2 E_{n,0}(a, h) - \mu_1 E_{n,1}(a, h)}{(\mu_2\mu_0 - \mu_1^2)f(x_0)} + o_p(r_n).$$

Asymptotic normality follows from standard results. The bias expression follows from:

$$\begin{aligned}\frac{d^2}{dx^2}\mathbb{E}[u^*(\eta)|X = x]|_{x=x_0} &= \partial_x^2 m(\eta, x_0) - \frac{1}{\eta}f_{Y|X}(q_0(\eta)|x_0)q_0(\eta)\partial_x^2 Q(\eta, x_0), \\ \frac{d^2}{dx^2}\mathbb{E}[e^*(\eta)|X = x]|_{x=x_0} &= \frac{1}{\eta}f_{Y|X}(q_0(\eta)|x_0)q_0(\eta)\partial_x^2 Q(\eta, x_0).\end{aligned}$$

The variance is calculated as follows. Recall that $h/a \to \rho$. It holds that

$$\begin{aligned}\mathrm{Var}(u^*(\eta)|X = x_0) &= \frac{1}{\eta}\mathrm{Var}(Y|Y \leq Q(\eta, X), X = x_0) + \frac{1-\eta}{\eta}m(\eta, x_0)^2, \\ \mathrm{Var}(e^*(\eta)|X = x_0) &= \frac{1-\eta}{\eta}.\end{aligned}$$

where the first line is derived in part (i) above. Furthermore,

$$
\begin{aligned}
&\mathrm{Var}\left(k_h(X_h)(\mu_2-\mu_1X_h)\frac{1}{\eta}m(\eta,x_0)e^*(\eta)+k_a(X_a)(\mu_2-\mu_1X_a)\frac{1}{\eta}Q(\eta,x_0)e^*(\eta)\right)\\
&=\int\left[\frac{1}{h}k\Big(\frac{x-x_0}{h}\Big)\Big(\mu_2-\mu_1\frac{x-x_0}{h}\Big)\frac{1}{\eta}m(\eta,x_0)+\frac{1}{a}k\Big(\frac{x-x_0}{a}\Big)\Big(\mu_2-\mu_1\frac{x-x_0}{a}\Big)\frac{1}{\eta}Q(\eta,x_0)\right]^2\\
&\quad\times\mathrm{Var}(e^*(\eta)|X=x)f_X(x)dx\\
&=\frac{1}{h}\int_{\mathcal{D}}\left[k(v)(\mu_2-\mu_1v)\frac{1}{\eta}m(\eta,x_0)+\rho k(v\rho)(\mu_2-\mu_1v\rho)\frac{1}{\eta}Q(\eta,x_0)\right]^2dv\\
&\quad\times\mathrm{Var}(e^*(\eta)|X=x_0)f_X(x_0)(1+o(1)),
\end{aligned}
$$

which concludes the proof. □

Proof of Proposition 1.B.4. Part (i). It holds that

$$\widetilde{m}^{TS}(\eta,x_0;h)=m(\eta,x_0)+\frac{\widetilde{S}_{n,2}\widetilde{U}_{n,0}-\widetilde{S}_{n,1}\widetilde{U}_{n,1}}{\widetilde{S}_{n,2}\widetilde{S}_{n,0}-\widetilde{S}_{n,1}^2},$$

where

$$
\begin{aligned}
\widetilde{U}_{n,j}&\equiv\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\frac{1}{\eta}(Y_i-m^*(\eta,X_i))\mathbb{1}(Y_i\le Q(\eta,X_i))=O_p(r_n),\\
\widetilde{S}_{n,j}&\equiv\frac{1}{\eta n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\mathbb{1}(Y_i\le Q(\eta,X_i))=\mu_jf_X(x_0)+o_p(1).
\end{aligned}
$$

The result follows from standard theory of local linear estimation.
Part (ii). It holds that

$$\widehat{m}^{TS}(\eta,x_0;h,h)=m(\eta,x_0)+\frac{\widehat{S}_{n,2}U_{n,0}(a,h)-\widehat{S}_{n,1}U_{n,1}(a,h)}{\widehat{S}_{n,2}\widehat{S}_{n,0}-\widehat{S}_{n,1}^2},$$

where

$$
\begin{aligned}
U_{n,j}(a,h)&=\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\frac{1}{\eta}(Y_i-m^*(\eta,X_i))\mathbb{1}(Y_i\le\widehat{Q}(\eta,X_i;a)),\\
\widehat{S}_{n,j}(a)&=\frac{1}{\eta n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\mathbb{1}(Y_i\le\widehat{Q}(\eta,X_i;a))=\mu_jf_X(x_0)+o_p(1).
\end{aligned}
$$

Lemma 1.E.6 yields that

$$
\begin{aligned}
U_{n,j}(a,h)=&\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\frac{1}{\eta}(Y_i-Q^*(\eta,X_i))\mathbb{1}(Y_i\le Q^*(\eta,X_i))\\
&+\frac{1}{n}\sum_{i=1}^n k_{h,i}X_{h,i}^j\frac{1}{\eta}(Q^*(\eta,X_i)-m^*(\eta,X_i))\mathbb{1}(Y_i\le\widehat{Q}(\eta,X_i;h))+O_p(r_n^2).
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
U_{n,j}(a,h) &= \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}(Y_i - Q^*(\eta, X_i))\mathbb{1}(Y_i \le Q^*(\eta, X_i)) \\
&+ \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}(Q^*(\eta, X_i) - m^*(\eta, X_i))\mathbb{1}(Y_i \le Q^*(\eta, X_i)) \\
&+ \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}\frac{1}{\eta}(Q^*(\eta, X_i) - m^*(\eta, X_i))f_{Y|X}(Q(\eta, x_0)|x_0) \\
&\times \{\widehat{q}_0(\eta; a) - q_0(\eta) + (\widehat{q}_1(\eta; a) - q_1(\eta))(X_i - x_0)\} + o_p(r_n).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\widehat{m}^{TS}(\eta, x_0; h, h) &= m(\eta, x_0) + \frac{\widehat{S}_{n,2}U^*_{n,0}(a,h) - \widehat{S}_{n,1}U^*_{n,1}(a,h)}{\widehat{S}_{n,2}\widehat{S}_{n,0} - \widehat{S}^2_{n,1}} \\
&+ \frac{1}{\eta}(Q^*(\eta, x_0) - m^*(\eta, x_0))f_{Y|X}(Q(\eta, x_0)|x_0)(\widehat{q}_0(\eta; a) - q_0(\eta)),
\end{aligned}
$$

where $U^*_{n,j}(h) = \frac{1}{n}\sum_{i=1}^{n} k_{h,i}X_{h,i}^{j}u_i^{**}(\eta)$ with $u_i^{**}(\eta) = \frac{1}{\eta}(Y_i - m^*(\eta, X_i))\mathbb{1}(Y_i \le Q^*(\eta, X_i))$.

The variance expression follows from the calculations in the proof of Proposition 1.B.2, and the bias from the fact that

$$
\frac{d^2}{dx^2}\mathbb{E}[u^{**}(\eta)|X = x]|_{x=x_0} = \partial_x^2 m(\eta, x_0) - \frac{1}{\eta}f_{Y|X}(q_0(\eta)|x_0)(q_0(\eta) - m(\eta, x_0))\partial_x^2 Q(\eta, x_0).
$$

□

1.F.4. **Proof of Proposition 1.B.5.**

Proof. Note that

$$
l(x) \equiv \mathbb{E}[\psi(\eta, Q(\eta, X)) - \psi(\eta, Q(\eta, x_0))|X = x] = \frac{1}{\eta}\int_{Q(\eta,x_0)}^{Q(\eta,x)} (y - Q(\eta, x_0))f_{Y|X}(y|x)dy.
$$

By the Leibniz integral rule, it holds that

$$
\begin{aligned}
l'(x) &= \frac{1}{\eta}\partial_x^1 Q(\eta, x)(Q(\eta, x) - Q(\eta, x_0))f_{Y|X}(Q(\eta, x)|x) \\
&+ \frac{1}{\eta}\int_{Q(\eta,x_0)}^{Q(\eta,x)} (y - Q(\eta, x_0))\partial_x f_{Y|X}(y|x)dy.
\end{aligned}
$$

Furthermore,

$$
l''(x_0) = \frac{1}{\eta}\Big(\partial_x^1 Q(\eta, x_0)\Big)^2 f_{Y|X}(Q(\eta, x_0)|x_0),
$$

which concludes the proof. □

CHAPTER 2

Simple Inference in Fuzzy Regression Discontinuity Designs with a Manipulated Running Variable

Joint work with Christoph Rothe.

2.1. INTRODUCTION

In a regression discontinuity (RD) design, units are assigned a treatment if and only if their running variable exceeds a fixed cutoff value. If units fully comply with their treatment assignment, the design is called sharp. Otherwise, the design is called fuzzy. The standard identification argument in RD designs relies on the assumption that units just to the left and just to the right of the cutoff are very similar in all aspects except for the treatment assignment status. This assumption, however, is often questionable if the running variable is not exogenously determined. For example, suppose that some units can misreport the value of their running variable and ensure that it falls above the cutoff. In that case, then the units observed on different sides of the cutoff are no longer be comparable.

To analyze settings where the standard RD assumptions fail, Gerard, Rokkanen, and Rothe (2020), henceforth GRR, develop a framework where there are two unobservable types of units: *always-assigned* units, for which the realization of the running variable is always to the right of the cutoff, and hence they are assigned the treatment; and *potentially-assigned* units, whose density of the running variable is smooth around the cutoff, and hence they satisfy the standard assumptions of an RD design. GRR show that the average treatment effect for the potentially-assigned units at the cutoff is partially identified under this model. In fuzzy RD designs, however, the sharp bounds take a complicated form, which makes it very difficult to derive the asymptotic distribution of the plug-in estimator proposed by GRR. In consequence, GRR employ a computationally-intensive bootstrap procedure to conduct inference.

In this paper, we consider the manipulation framework of GRR and derive alternative bounds on the treatment effect of interest. The proposed bounds are not sharp, but they take a very simple form. We develop an estimator of these bounds and construct confidence sets for the partially identified parameter. Our procedure can be easily combined with inference methods available in the nonparametric literature, such as robust bias corrections (Calonico et al., 2014) or bias-aware inference (Armstrong and Kolesár, 2020).

Our point of departure is the Wald-ratio representation of the local average treatment effect for the potentially-assigned units. It is given by the ratio of the jumps in the conditional expectation of the outcome variable and conditional treatment probability at the cutoff, both of which are calculated for the subpopulation of potentially-assigned units. In the considered model, the proportion of the always-assigned units among all units just to the right of the cutoff is identified by the jump in the density of the running variable at the cutoff. This information can be used to bound the denominator and numerator of the Wald ratio by considering the extreme scenarios in which the always-assigned units have the highest or the lowest outcomes among the units just to the right of the cutoff and that they are all treated or all untreated. Given the identified sets for the denominator and the numerator, we obtain the set of possible values for the treatment effect. This derivation resembles the construction of Anderson-Rubin confidence sets, where the null hypothesis is reformulated in such a way that it does not involve a ratio (Anderson et al., 1949); see also Noack and Rothe (2021) for an application of Anderson-Rubin-type confidence sets to fuzzy RD designs. Confidence sets are then constructed by test inversion, where we test the null hypothesis that a candidate parameter value belongs to the identified set. The definition of the identified set involves truncated conditional expectations, which we estimate using the estimator proposed in Chapter 1 of this thesis.

The remainder of this chapter is organized as follows. In Section 2.2, we outline the manipulation framework of GRR. In Section 2.3, we present a partial identification result. In Section 2.4, we propose an estimator of the bounds. Confidence sets are proposed in Section 2.5. In Section 2.6, we presents a simulation study. Section 2.7 concludes.

Throughout the chapter, we use the following notation. For a generic function $f(x)$, we write $f(0^+) = \lim_{x \downarrow 0} f(x)$ and $f(0^-) = \lim_{x \uparrow 0} f(x)$ for the right and left limit of the function $f(x)$ at zero, respectively. We also implicitly assume that whenever we take a limit or an expectation, it exists and is finite.

2.2. MODEL AND OBJECT OF INTEREST

In this section, we briefly outline the manipulation framework of GRR and introduce the parameter of interest. In the original paper, GRR provide an extensive discussion of applicability of their model, which we do not repeat here for the sake of brevity.

We consider a fuzzy RD design, where a unit is assigned the treatment if and only if their running variable, denoted by X_i, exceeds a fixed cutoff value, which we normalize to zero. Each unit belongs to one of two groups with the membership indicated by an unobservable dummy variable $M_i \in \{0,1\}$. Units with $M_i = 0$, called potentially-assigned, satisfy the assumptions of a valid RD design; see Assumption 2.1 for the precise statement of these conditions. Units with $M_i = 1$, called always-assigned, have realization of the running variable to the right of the cutoff, and hence they are assigned the treatment.

Potentially-assigned units have potential outcomes in the absence and in the presence of treatment, denoted by $Y_i(d)$ for $d \in \{0,1\}$, and potential treatment statuses, denoted by $D_i(x)$ for x in the support of X_i. We put $D_i^+ = D_i(0^+)$ and $D_i^- = D_i(0^-)$. The observed treatment status is denoted by the indicator D_i, and Y_i is the observed outcome variable. In Assumption 2.1, we restate the assumptions of the model developed by GRR; cf. their Assumptions 1–3. We drop only the first condition in part (iii) of their Assumption 1, as it is not relevant for our analysis.

Assumption 2.1. *(i)* $\mathbb{P}[D_i = 1|X_i = 0^+, M_i = 0] > \mathbb{P}[D_i = 1|X_i = 0^-, M_i = 0]$*; (ii)* $\mathbb{P}[D_i^+ \geq D_i^-|X_i = 0, M_i = 0] = 1$*; (iii)* $\mathbb{E}[Y_i(d)|D_i^+ = d^1, D_i^- = d^0, X_i = x, M_i = 0]$ *and* $\mathbb{P}[D_i^+ = d^1, D_i^- = d^0|X_i = x, M_i = 0]$ *are continuous in x at 0 for d,* d^0*,* $d^1 \in \{0,1\}$*; (iv)* $F_{X|M=0}(x)$ *is continuously differentiable in x at 0, and the derivative is strictly positive; (v)* $\mathbb{P}[X_i \geq 0|M_i = 1] = 1$*; (vi)* $F_{X|M=1}(x)$ *is right-differentiable in x at 0.*

Parts (i)–(iv) impose standard RD assumptions for the subpopulation of potentially-assigned units. We assume that treatment assignment has a weakly positive effect on the treatment take-up for all units, and there are at least some units that take up the treatment if they are assigned, and not otherwise. We further assume that the distributions of potential outcomes and potential treatment statuses evolve continuously through the cutoff. Lastly, the running variable is continuously distributed at the cutoff with positive and continuous density. Parts (v) and (vi) concern the always-assigned units. Their running variable takes on values only to the right of the cutoff, which is the defining feature of this subpopulation. Moreover, it does not have a mass point at the cutoff. If that was the case, these observations could be simply removed from the dataset. Hence, this assumption is not restrictive. Under Assumption 2.1, the running variable in the whole population is continuously distributed, but its density, which we denote by f_X, can be discontinuous at the cutoff.

Following GRR, we focus on the local average treatment effect among the potentially-assigned compliers at the cutoff. These are the units who would be treated if their running variable was just to the right of the cutoff and would not be treated if it was just to the left of the cutoff. Formally, the object of interest is defined as:

$$\Gamma = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = 0, D_i^+ > D_i^-, M_i = 0]. \tag{2.2.1}$$

This quantity is analogous to the standard estimand in fuzzy RD designs, where the average treatment effect is identified only for the subpopulation of compliers. The non-standard feature of the above definition is, however, that it involves conditioning on the unobserved indicator M_i. For that reason, Γ is not point-identified.

2.3. PARTIAL IDENTIFICATION

In this section, we develop a partial identification result for Γ. We first present a preliminary analysis of the Wald-ratio-type representation of the estimand and then characterize a (non-sharp) identified set. We conclude by discussing the relation to the sharp bounds obtained by GRR.

2.3.1. **Preliminaries.** Potentially-assigned units satisfy the assumptions of the classic fuzzy RD framework. We can therefore use standard derivations to represent the estimand as the ratio of the jump in the conditional expectation of the outcome variable and the change in the conditional treatment probability among the potentially-assigned units at the cutoff (see, e.g., Lee and Lemieux, 2010):[1]

$$\Gamma = \frac{\Delta}{\Psi}, \tag{2.3.1}$$

where

$$\Delta = \mathbb{E}[Y_i|X_i = 0^+, M_i = 0] - \mathbb{E}[Y_i|X_i = 0^-], \tag{2.3.2}$$

$$\Psi = \mathbb{E}[D_i|X_i = 0^+, M_i = 0] - \mathbb{E}[D_i|X_i = 0^-]. \tag{2.3.3}$$

In the above definitions, we do not condition on $M_i = 0$ when the running variable falls to the left of the cutoff because, by definition, there are only potentially-assigned units on this side of the cutoff. The conditional expectations $\mathbb{E}[Y_i|X_i = 0^+, M_i = 0]$ and $\mathbb{E}[D_i|X_i = 0^+, M_i = 0]$ are not point identified because the indicator M_i is not observed, but they can be bounded.

The first step in our partial identification analysis is to note that the proportion of always-assigned units among all units just to the right of the cutoff is identified from the size of the discontinuity in the observed density f_X at the cutoff. Specifically, under the

[1] GRR use this representation in a working-paper version of their article (Gerard et al., 2016).

model outlined in Section 2.2, GRR show that

$$\tau \equiv \mathbb{P}[M_i = 1 | X_i = 0^+] = 1 - \frac{f_X(0^-)}{f_X(0^+)}. \tag{2.3.4}$$

For any given extent of manipulation, measured by τ, the numerator and denominator of the ratio in (2.3.1) can be bounded using the trimming approach that GRR apply to analyze sharp RD designs. The bounds on Δ are obtained by considering the extreme scenarios in which the always-assigned units have the highest or the lowest $\tau \cdot 100\%$ outcomes among the units just to the right of the cutoff. If the conditional distribution of the outcome variable just to the right of the cutoff is continuous, then Δ is bounded from below and from above, respectively, by

$$\begin{aligned}\Delta^L &= \Delta^L(\tau) = \mathbb{E}[Y_i | X_i = 0^+, Y_i \leq Q_{Y|X=0^+}(1-\tau)] - \mathbb{E}[Y_i | X_i = 0^-],\\ \Delta^U &= \Delta^U(\tau) = \mathbb{E}[Y_i | X_i = 0^+, Y_i \geq Q_{Y|X=0^+}(\tau)] - \mathbb{E}[Y_i | X_i = 0^-].\end{aligned}$$

Similarly, the bounds on Ψ can be obtained by considering the scenarios in which the always-assigned units are either all treated or all untreated. In some cases, these extreme scenarios might be inconsistent with the observed treatment probabilities, but Ψ is always bounded from below and from above, respectively, by

$$\begin{aligned}\Psi^L &= \Psi^L(\tau) = \frac{\mathbb{E}[D_i | X_i = 0^+] - \tau}{1-\tau} - \mathbb{E}[D_i | X_i = 0^-],\\ \Psi^U &= \Psi^U(\tau) = \frac{\mathbb{E}[D_i | X_i = 0^+]}{1-\tau} - \mathbb{E}[D_i | X_i = 0^-].\end{aligned}$$

2.3.2. **Identified Set.** Based on the identity $\Gamma = \Delta/\Psi$, the bounds on Δ and Ψ translate into bounds on Γ, which are given in Table 2.1. These bounds, however, involve multiple case distinctions, which generally poses a problem for conducting inference. This issue, however, can be circumvented by using an equivalent formulation of the parameter Γ. Specifically, we exploit the following identity:

$$\Delta - \Gamma\Psi = 0. \tag{2.3.5}$$

This reformulation is in the spirit of 'identification-robust' Anderson-Rubin test in the weak instrumental variables literature.

The relation in (2.3.5) and the bounds on Δ and Ψ yield two conditional moment inequalities that the true value of the parameter Γ satisfies. They are obtained by considering the lowest and the highest value that the expression $\Delta - \Gamma\Psi$ can take given that $\Delta \in [\Delta^L, \Delta^U]$ and $\Psi \in [\Psi^L, \Psi^U]$. Their form differs slightly depending on whether Γ

is positive or negative. If $\Gamma \geq 0$, then

$$\Delta^L - \Gamma\Psi^U \leq 0 \leq \Delta^U - \Gamma\Psi^L. \tag{2.3.6}$$

If $\Gamma < 0$, then

$$\Delta^L - \Gamma\Psi^L \leq 0 \leq \Delta^U - \Gamma\Psi^U. \tag{2.3.7}$$

We will now consider the set of candidate values γ for Γ that are consistent with the above inequalities. To express this set concisely, we define the following combinations of the bounds on Δ and Ψ. For $k_1, k_2 \in \{L, U\}$, we put

$$g^{k_1k_2}(\tau,\gamma) = \Delta^{k_1}(\tau) - \gamma\Psi^{k_2}(\tau), \tag{2.3.8}$$

where we emphasize the dependence on τ for consistency of notation with the corresponding estimators introduced in the next section. Theorem 2.1 states a partial identification result, which is based on the restrictions implied by inequalities (2.3.6) and (2.3.7).

Theorem 2.1. *Suppose that Assumption 2.1 holds and $F_{Y|X=0^+}(y)$ is continuous in y. Then*

$$\Gamma \in \Gamma_I(\tau) \equiv \Big\{\gamma \geq 0 : g^{LU}(\tau,\gamma) \leq 0 \leq g^{UL}(\tau,\gamma)\Big\} \cup \Big\{\gamma < 0 : g^{LL}(\tau,\gamma) \leq 0 \leq g^{UU}(\tau,\gamma)\Big\}.$$

The inequalities in Theorem 2.1 can be explicitly solved for c. In general, $\Gamma_I(\tau)$ can be either a finite interval, the whole real line, a half-line, or the sum of two half-lines. We present this case distinction in Table 2.1. The finite ends are expressed as ratios of bounds on the numerator and the denominator. The particular form depends on the signs of the bounds on Δ and Ψ. This case distinction as well as the possibility of Ψ^L and Ψ^U being close to zero, would complicate inference based on this representation of the identified set. The possible shapes of the identified sets are analogous to Anderson-Rubin confidence sets in the weak instrumental variables literature.[2] In practice, the proposed form of the identified set $\Gamma_I(\tau)$ is best suited for settings in which $\Psi^L > 0$, which is equivalent to

$$\tau < \frac{\mathbb{E}[D_i|X_i = 0^+] - \mathbb{E}[D_i|X_i = 0^-]}{1 - \mathbb{E}[D_i|X_i = 0^-]}.$$

The above condition holds if τ is smaller than the observed jump in the treatment probability at the cutoff. This should be the case in most empirical scenarios in which the eligibility has a strong effect on the treatment take-up rate, and one considers manipulation as a small deviation from a valid RD design.

[2] Anderson-Rubin confidence sets in instrumental variables models can theoretically take the forms as in the second row of Table 2.2, but this event happens with probability zero.

Table 2.1: Identified set $\Gamma_I(\tau)$.

	$0 < \Delta^L$	$\Delta^L \leq 0 \leq \Delta^U$	$\Delta^U < 0$
$0 < \Psi^L$	$\left[\frac{\Delta^L}{\Psi^U}, \frac{\Delta^U}{\Psi^L}\right]$	$\left[\frac{\Delta^L}{\Psi^L}, \frac{\Delta^U}{\Psi^L}\right]$	$\left[\frac{\Delta^L}{\Psi^L}, \frac{\Delta^U}{\Psi^U}\right]$
$\Psi^L = 0$	$\left[\frac{\Delta^L}{\Psi^U}, \infty\right)$	$(-\infty, \infty)$	$\left(-\infty, \frac{\Delta^U}{\Psi^U}\right]$
$\Psi^L < 0$	$\left(-\infty, \frac{\Delta^U}{\Psi^L}\right] \cup \left[\frac{\Delta^L}{\Psi^U}, \infty\right)$	$(-\infty, \infty)$	$\left(-\infty, \frac{\Delta^U}{\Psi^U}\right] \cup \left[\frac{\Delta^L}{\Psi^L}, \infty\right)$

Note: In the last row, we use that fact that $0 < \Psi^U$ by assumption.

2.3.3. **Discussion.** The proposed bounds are not sharp for a number of reasons. In some cases, we could infer that the extreme scenarios of location of always-assigned units are not consistent with the data. For example, using the fact that the outcome distribution of the potentially-assigned always-takers at the cutoff is identified, we could exclude the appropriate portion of the observed outcome distribution just to the right of the cutoff before considering the extreme scenarios of location of the always-assigned units. Moreover, in our procedure, different units can be trimmed when constructing bounds on the numerator and the denominator of Γ, which indicates that these extreme scenarios are not empirically feasible. Nevertheless, the proposed bounds are valid, and they can be used as a basis for constructing confidence sets.

2.4. ESTIMATION

In this section, we introduce estimators of the conditional moment functions $g^*(\tau, \gamma)$ appearing in the definition of the identified set $\Gamma_I(\tau)$ in Theorem 2.1 and discuss their asymptotic properties. These results form the key input for constructing confidence sets in Section 2.5.

Throughout this section, we use the following shorthand notation. For a generic random variable W_i, we write $m_W(x) = \mathbb{E}[W_i|X_i = x]$. Further, $m_Y^L(\tau, x) = \mathbb{E}[Y_i|X_i = x, Y_i \leq Q_{Y|X=x}(1-\tau)]$ and $m_Y^U(\tau, x) = \mathbb{E}[Y_i|X_i = x, Y_i \geq Q_{Y|X=x}(\tau)]$. One-sided limits at the cutoff are denoted by: $m_{W-} = m_W(0^-)$, $m_{W+} = m_W(0^+)$, $m_{Y+}^L(\tau) = m_Y^L(\tau, 0^+)$, $m_{Y+}^U(\tau) = m_Y^U(\tau, 0^+)$, and $Q_{Y+}(u) = Q_{Y|X=0^+}(u)$ for $u \in (0, 1)$.

2.4.1. **Estimators.** We propose estimators of the moment functions $g^*(\tau, \gamma)$ based on kernel methods. Let $k(\,\cdot\,)$ be a kernel function and h a bandwidth, which governs the size of the estimation window. For ease of exposition, we use the same bandwidth in all the estimation steps, but it is straightforward to allow for different bandwidths in different steps. We define $k_h(v) = k(v/h)/h$, $k_h^-(v) = k_h(v)\mathbb{1}(v < 0)$, and $k_h^+(v) = k_h(v)\mathbb{1}(v \geq 0)$.

Proportion of Always-Assigned Units. We estimate the one-sided limits of the density of the running variable at the cutoff using 'linear' boundary kernels (Jones, 1993). Let

$$\widehat{f}^- = \frac{1}{n}\sum_{i=1}^n k_h^-(X_i)\frac{\bar{\mu}_2 - \bar{\mu}_1|X_i/h|}{\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2} \text{ and } \widehat{f}^+ = \frac{1}{n}\sum_{i=1}^n k_h^+(X_i)\frac{\bar{\mu}_2 - \bar{\mu}_1|X_i/h|}{\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2}, \tag{2.4.1}$$

where $\bar{\mu}_j = \int_0^\infty v^j k(v)dv$. The proportion τ of always-assigned units among all the units just to the right of the cutoff is estimated as:

$$\widehat{\tau} = \max\left\{1 - \widehat{f}^-/\widehat{f}^+, 0\right\}.$$

Conditional Expectations. The conditional expectations m_{D-}, m_{D+}, and m_{Y-} are estimated using the local linear estimator. For a generic outcome variable W_i, we estimate m_{W+} and m_{W-} by:

$$\widehat{m}_{W+} = \widehat{\mathbb{E}}[W_i|X_i = 0^+] \equiv e_1^\top \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h^+(X_i)(W_i - \beta_0 - \beta_1 X_i)^2,$$
$$\widehat{m}_{W-} = \widehat{\mathbb{E}}[W_i|X_i = 0^-] \equiv e_1^\top \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h^-(X_i)(W_i - \beta_0 - \beta_1 X_i)^2.$$

Truncated Conditional Expectations. The truncated conditional expectations of the outcome variable, $m_{Y+}^L(\tau)$ and $m_{Y+}^U(\tau)$, are estimated using the two-stage, local-linear-type estimator proposed in Chapter 1. In the first stage, we run a local linear quantile regression. The intercept and slope of the conditional u-quantile function just to the right of the cutoff, $Q_{Y|X=0^+}(u)$ and $\partial_x^1 Q_{Y|X=x}(u)|_{x=0^+}$, are estimated as:

$$(\widehat{q}_0(u), \widehat{q}_1(u))^\top = \operatorname*{arg\,min}_{\beta_0,\beta_1} \sum_{i=1}^n k_h^+(X_i)\rho_u(Y_i - \beta_0 - \beta_1 X_i),$$

where $\rho_u(v) = v(u - \mathbb{1}(v \leq 0))$ is the 'check' function. For $x \geq 0$, we put

$$\widehat{Q}_{Y|X=x}(u) = \widehat{q}_0(u) + \widehat{q}_1(u)x.$$

In the second stage, we run local linear regressions with generated outcome variables. For a function $q : (0,1) \times \text{supp}(X_i) \to \mathbb{R}$, we define

$$\psi_i^L(\tau, q) = \frac{1}{1-\tau}Y_i\mathbb{1}(Y_i \leq q(\tau, X_i)) - \frac{1}{1-\tau}q(\tau, X_i)(\mathbb{1}(Y_i \leq q(\tau, X_i)) - (1-\tau)),$$
$$\psi_i^U(\tau, q) = \frac{1}{1-\tau}Y_i\mathbb{1}(q(1-\tau, X_i) \leq Y_i) - \frac{1}{1-\tau}q(1-\tau, X_i)(\mathbb{1}(q(1-\tau, X_i) \leq Y_i) - (1-\tau)).$$

We estimate $m^L_{Y+}(\tau)$ and $m^U_{Y+}(\tau)$ using local linear estimators with generated outcome variables $\psi^L_i(\tau, \widehat{Q}_{Y|X})$ and $\psi^U_i(\tau, \widehat{Q}_{Y|X})$, respectively:

$$\widehat{m}^L_{Y+}(\tau) = \widehat{\mathbb{E}}\Big[\psi^L_i\Big(\tau, \widehat{Q}_{Y|X}\Big)\Big|X_i = 0^+\Big] \text{ and } \widehat{m}^U_{Y+}(\tau) = \widehat{\mathbb{E}}\Big[\psi^U_i\Big(\tau, \widehat{Q}_{Y|X}\Big)\Big|X_i = 0^+\Big].$$

Final Estimators. By linearity of the local linear estimator, the estimators of the (truncated) conditional expectations can be combined into one estimator on each side of the cutoff. For $* \in \{LU, UL, LL, UU\}$, $g^*(\tau, \gamma)$ is estimated by

$$\widehat{g}^*(\tau, \gamma) = \widehat{\mathbb{E}}\Big[G^*_{+,i}\Big(\tau, \widehat{Q}_{Y|X}, \gamma\Big)\Big|X_i = 0^+\Big] - \widehat{\mathbb{E}}[G_{-,i}(\gamma)|X_i = 0^-]. \tag{2.4.2}$$

where $G_{-,i}(\gamma) = Y_i - \gamma D_i$ and

$$\begin{aligned} G^{LU}_{+,i}(\tau, q, \gamma) &= \psi^L_i(\tau, q) - \gamma\frac{D_i}{1-\tau}, & G^{UL}_{+,i}(\tau, q, \gamma) &= \psi^U_i(\tau, q) - \gamma\frac{D_i - \tau}{1-\tau}, \\ G^{LL}_{+,i}(\tau, q, \gamma) &= \psi^L_i(\tau, q) - \gamma\frac{D_i - \tau}{1-\tau}, & G^{UU}_{+,i}(\tau, q, \gamma) &= \psi^U_i(\tau, q) - \gamma\frac{D_i}{1-\tau}. \end{aligned}$$

2.4.2. **Asymptotic Results.** We analyze the asymptotic distribution of the estimators of $g^*(\tau, \gamma)$ in two versions. The first one concerns analysis where the trimming proportion τ is fixed, and the second one allows for an estimated $\widehat{\tau}$.

To state the assumptions for our asymptotic analysis, we introduce some additional notation. For $\epsilon > 0$, let $\mathcal{X}^\epsilon = (-\epsilon, \epsilon)$, $\mathcal{X}^\epsilon_- = (-\epsilon, 0)$, and $\mathcal{X}^\epsilon_+ = [0, \epsilon)$. For an interval I, let $\mathcal{C}^j(I)$ denote the class of functions whose jth derivative is continuous on the interior of I (with $j = 0$ corresponding to continuous functions). Further, let $\mathcal{C}^j_+(I)$ denote the class of functions that belong to $\mathcal{C}^j(I)$ and are bounded away from zero on I. Throughout the rest of the paper, we implicitly assume that if a function is continuous on an open interval I, then also its limits at the boundary points of I exist and are finite.

Assumption 2.2. *There exists $\epsilon > 0$ such that the following conditions hold. (i) The data $\{(Y_i, X_i, D_i)^n_{i=1}\}$ are an i.i.d. sample from a fixed population; (ii) $f_X(\,\cdot\,) \in \mathcal{C}^2_+(\mathcal{X}^\epsilon_-) \cap \mathcal{C}^2_+(\mathcal{X}^\epsilon_+)$; (iii) $f_{Y|X=x}(y)$ is continuous in x and y on $\mathcal{X}^\epsilon_+ \times \mathbb{R}$; (iv) $m_Y(\,\cdot\,), m_D(\,\cdot\,) \in \mathcal{C}^2(\mathcal{X}^\epsilon_-)$; (v) $m^L_Y(u, \cdot\,)$, $m^U_Y(u, \cdot\,), Q_{Y|X=\cdot}(u), m_D(\,\cdot\,) \in \mathcal{C}^2(\mathcal{X}^\epsilon_+)$ for all $u \in [0, 1)$; (vi) $\partial^2_x Q_{Y|X=x}(u)$ is continuous in u for $x \in \mathcal{X}^\epsilon_+$; (vii) $\mathbb{V}[G^*_{+,i}(\gamma, Q_{Y|X}, \tau)|X_i = \,\cdot\,] \in \mathcal{C}^0_+(\mathcal{X}^\epsilon_+)$ for all $\gamma \in \mathbb{R}$ and $* \in \{LU, UL, LL, UU\}$; (viii) $\mathbb{V}[G_{-,i}(\gamma)|X = \,\cdot\,] \in \mathcal{C}^0_+(\mathcal{X}^\epsilon_-)$ for every $\gamma \in \mathbb{R}$; (ix) $\mathbb{E}[Y^{2+\delta}_i|X_i = \,\cdot\,]$ is bounded on $\mathcal{X}$ for some $\delta > 0$; (x) The kernel k is a bounded and symmetric density function with compact support; (xi) As $n \to \infty$, $h \to 0$ and $nh \to 0$.*

Assumption 2.2 contains standard assumption for local-linear-type estimation of the density and conditional expectations, as well as the conditions for estimation of truncated conditional expectations developed in Chapter 1. In each case, the respective curve to be estimated has to be twice continuously differentiable to the left and to the right of the cutoff, but not necessarily at the cutoff. This requirement applies also to the conditional quantile function, which is a nuisance function when estimating the truncated conditional expectations. These smoothness conditions are complemented with standard moment conditions. The restrictions on the kernel and bandwidth are standard.

To present the asymptotic distribution, we define the following kernel constants:

$$\bar{\mu} = (\bar{\mu}_2^2 - \bar{\mu}_1\bar{\mu}_3)/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2) \text{ and } \bar{\kappa} = \int_0^\infty (k(v)(\bar{\mu}_1 v - \bar{\mu}_2))^2 dv/(\bar{\mu}_2\bar{\mu}_0 - \bar{\mu}_1^2)^2,$$

where, as defined after (2.4.1), $\bar{\mu}_j = \int_0^\infty v^j k(v)dv$. Theorem 2.1 states the asymptotic distribution of $\widehat{g}^*(\tau,\gamma)$ and $\widehat{g}^*(\widehat{\tau},\gamma)$.

Theorem 2.2. *Suppose that Assumptions 1 and 2 hold. For all $\gamma \in \mathbb{R}$ and $* \in \{LU, UL, LL, UU\}$, it holds that:*

(i) If $\tau \in [0,1)$, then

$$\sqrt{nh}\left(\widehat{g}^*(\tau,\gamma) - g^*(\tau,\gamma) - B^*(\tau,\gamma)h^2\right) \xrightarrow{d} \mathcal{N}(0, V^*(\tau,\gamma)),$$

where

$$B^*(\tau,\gamma) = \frac{1}{2}\bar{\mu}\left(\partial_x^2\mathbb{E}[G^*_{+,i}(\tau, Q_{Y|X},\gamma)|X_i = x]|_{x=0^+} - \partial_x^2\mathbb{E}[G_{-,i}(\gamma)|X_i = x]|_{x=0^-}\right) + o_p(1),$$
$$V^*(\tau,\gamma) = \frac{\bar{\kappa}}{f_X(0^+)}\mathbb{V}[G^*_{+,i}(\tau, Q_{Y|X},\gamma)|X_i = 0^+] + \frac{\bar{\kappa}}{f_X(0^-)}\mathbb{V}[G_{-,i}(\gamma)|X_i = 0^-].$$

Moreover, the pairs $\left(\widehat{g}^{LU}(\tau,\gamma), \widehat{g}^{UL}(\tau,\gamma)\right)$ and $\left(\widehat{g}^{LL}(\tau,\gamma), \widehat{g}^{UU}(\tau,\gamma)\right)$ are jointly asymptotically normally distributed if $\tau \in (0,1)$.

(ii) If $\tau \in (0,1)$, then

$$\sqrt{nh}\left(\widehat{g}^*(\widehat{\tau},\gamma) - g^*(\tau,\gamma) - (B^*(\tau,\gamma) + C^*(\tau,\gamma)B_\tau)\, h^2\right) \xrightarrow{d} \mathcal{N}(0, V^*(\tau,\gamma) + (C^*(\tau,\gamma))^2 V_\tau),$$

where

$$C^*(\tau,\gamma) = \partial_\tau\mathbb{E}[G^*_{+,i}(\tau, Q_{Y|X},\gamma)|X_i = 0^+],$$
$$B_\tau = \frac{1}{2}\bar{\mu}\,(1-\tau)\left(f_X''(0^+)/f_X(0^+) - f_X''(0^-)/f_X(0^-)\right) + o_p(1),$$
$$V_\tau = \bar{\kappa}(1-\tau)^2\left(1/f_X(0^+) + 1/f_X(0^-)\right).$$

Moreover, the pairs $\left(\widehat{g}^{LU}(\widehat{\tau},\gamma), \widehat{g}^{UL}(\widehat{\tau},\gamma)\right)$ and $\left(\widehat{g}^{LL}(\widehat{\tau},\gamma), \widehat{g}^{UU}(\widehat{\tau},\gamma)\right)$ are jointly asymptotically normally distributed.

The asymptotic distribution in part (i) resembles the standard results for local linear estimation. The key step to obtain it, is to show that the estimator using the estimated function $\widehat{Q}_{Y|X}$ is asymptotically equivalent to the estimator employing the true conditional quantile function $Q_{Y|X}$. This point is discussed extensively in Chapter 1. The bias expressions can be derived in a closed form. For $k \in \{L, U\}$, it holds that

$$\begin{aligned}
\partial_x^2 \mathbb{E}[G_{+,i}^{Lk}(\tau, Q_{Y|X}, \gamma)|X_i = x]|_{x=0^+} &= \partial_x^2 m_Y^L(\tau, 0^+) - \frac{\gamma}{1-\tau}\partial_x^2 m_D(0^+),\\
\partial_x^2 \mathbb{E}[G_{+,i}^{Uk}(\tau, Q_{Y|X}, \gamma)|X_i = x]|_{x=0^+} &= \partial_x^2 m_Y^U(\tau, 0^+) - \frac{\gamma}{1-\tau}\partial_x^2 m_D(0^+).
\end{aligned}$$

Moreover,

$$\partial_x^2 \mathbb{E}[G_{-,i}(\gamma)|X_i = x]|_{x=0^+} = \partial_x^2 m_Y(\tau, 0^+) - \gamma \partial_x^2 m_D(0^-).$$

The simple form of these bias expressions makes it possible to account for the smoothing bias when conducting inference using standard methods available in the literature.

In part (ii), the additional bias and variance terms are due to estimation of the trimming proportion τ. B_τ and V_τ represent the leading bias term and the asymptotic variance of $\widehat{\tau}$, respectively. These quantities appear in the asymptotic distribution of $\widehat{g}^*(\gamma, \widehat{\tau})$ scaled by the derivative of the estimand with respect to the trimming proportion. These derivatives take the following form:

$$\begin{aligned}
C^{LU}(\tau, \gamma) &= \frac{1}{1-\tau}\left(m_{Y+}^L(\tau) - Q_{Y+}(1-\tau)\right) - \frac{\gamma}{(1-\tau)^2} m_{D+},\\
C^{UL}(\tau, \gamma) &= \frac{1}{1-\tau}\left(m_{Y+}^U(\tau) - Q_{Y+}(\tau)\right) - \frac{\gamma}{(1-\tau)^2}(m_{D+} - 1),\\
C^{LL}(\tau, \gamma) &= \frac{1}{1-\tau}\left(m_{Y+}^L(\tau) - Q_{Y+}(1-\tau)\right) - \frac{\gamma}{(1-\tau)^2}(m_{D+} - 1),\\
C^{UU}(\tau, \gamma) &= \frac{1}{1-\tau}\left(m_{Y+}^U(\tau) - Q_{Y+}(\tau)\right) - \frac{\gamma}{(1-\tau)^2} m_{D+}.
\end{aligned}$$

2.5. CONFIDENCE SETS

In this section, we construct confidence sets (CSs) by inverting a test of the hypothesis that a candidate value γ belongs to the identified set $\Gamma_I(\tau)$. We discuss separately inference with a fixed and estimated manipulation level τ.

2.5.1. **Fixed Manipulation Level.** Suppose that the researcher presumes a certain proportion manipulation in the data. We construct a CS via test inversion, based on the following test statistics:

$$t^*(\tau, \gamma) = \frac{\widehat{g}^*(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^*(\tau, \gamma))},$$

where $* \in \{LU, UL, LL, UU\}$ and $\widehat{se}(\widehat{g}^*(\tau, \gamma))$ is some consistent standard error, which can be constructed based on the residuals from the local linear regressions in (2.4.2) under additional assumptions. We consider two ways of accounting for the asymptotic bias: undersmoothing and the bias-aware approach.

Undersmoothing. Suppose that the bias is asymptotically negligible in the sense that $nh^5 \to 0$, then

$$t^*(\tau, \gamma) \to \mathcal{N}(0, 1). \tag{2.5.1}$$

Moreover, the pairs $(t^{LU}(\tau, \gamma), t^{UL}(\tau, \gamma))$ and $(t^{LL}(\tau, \gamma), t^{UU}(\tau, \gamma))$ are jointly asymptotically normal. Critical values for this testing problem are motivated by the derivations of Imbens and Manski (2004) and Stoye (2009). Let

$$\widehat{\mathcal{D}}(\tau, \gamma) = \widehat{\Delta}^U(\tau) - \widehat{\Delta}^L(\tau) + |\gamma| \frac{\tau}{1 - \tau}. \tag{2.5.2}$$

$\widehat{\mathcal{D}}(\tau, \gamma)$ represents the difference between the moment functions used in the definition of $\Gamma_I(\tau)$; see Appendix 2.A.3. In the analysis of coverage of CSs for the partially identified parameter Γ, $\widehat{\mathcal{D}}(\tau, \gamma)$ plays the role of the length of the identified set in the derivations of Imbens and Manski (2004) and Stoye (2009).[3]

For $\gamma \geq 0$, we define $c_\alpha(\tau, \gamma)$ as the solution to the following equation:

$$\Phi\left(c_\alpha(\tau, \gamma) + \frac{\widehat{\mathcal{D}}(\tau, \gamma)}{\max\{\widehat{\mathrm{se}}(\widehat{g}^{LU}(\tau, \gamma)), \widehat{\mathrm{se}}(\widehat{g}^{UL}(\tau, \gamma))\}} \right) - \Phi\left(-c_\alpha(\tau, \gamma)\right) = 1 - \alpha. \tag{2.5.3}$$

For $\gamma < 0$, we define $c_\alpha(\tau, \gamma)$ as the solution to the following equation:

$$\Phi\left(c_\alpha(\tau, \gamma) + \frac{\widehat{\mathcal{D}}(\tau, \gamma)}{\max\{\widehat{\mathrm{se}}(\widehat{g}^{LL}(\tau, \gamma)), \widehat{\mathrm{se}}(\widehat{g}^{UU}(\tau, \gamma))\}} \right) - \Phi\left(-c_\alpha(\tau, \gamma)\right) = 1 - \alpha. \tag{2.5.4}$$

We build our CS by collecting the values γ for which the hypothesis $H_0 : \gamma \in \Gamma_I(\tau)$ is not rejected:

$$\begin{aligned} CS_\alpha(\tau) = &\{\gamma \geq 0 : -c_\alpha(\tau, \gamma) \leq t^{UL}(\tau, \gamma),\ t^{LU}(\tau, \gamma) \leq c_\alpha(\tau, \gamma)\} \\ &\cup \{\gamma < 0 : -c_\alpha(\tau, \gamma) \leq t^{UU}(\tau, \gamma),\ t^{LL}(\tau, \gamma) \leq c_\alpha(\tau, \gamma)\}. \end{aligned} \tag{2.5.5}$$

If $\widehat{\mathcal{D}}(\tau, \gamma)$ is large relative to the standard errors, then $c_\alpha(\tau, \gamma)$ is approximately the $1 - \alpha$ quantile of the standard normal distribution, as in one-sided confidence intervals. If $\widehat{\mathcal{D}}(\tau, \gamma) = 0$, then $c_\alpha(\gamma)$ equals the $1 - \alpha/2$ quantile of the standard normal distribution, as in two-sided confidence intervals.

Analogous CSs can be constructed using robust bias corrections of Calonico et al. (2014, 2018) because they also rely on the fact that the test statistic follows the standard normal distribution.

[3] In this paper, we do not provide any formal coverage guarantees uniformly in the data generating processes, but given the robustness properties of the CIs of Imbens and Manski (2004) and Stoye (2009), the proposed CIs can be expected to perform well in finite samples.

Bias-Aware Inference. Suppose that the second derivatives of the (truncated) conditional expectations involved in the definition of $\Gamma_I(\tau)$ are bounded by some known constants. Specifically, suppose that $|\partial_x^2 m_Y(0^-)| \leq \bar{B}_{Y-}$, $|\partial_x^2 m_Y^L(\tau, 0^+)| \leq \bar{B}_{Y+}^L(\tau)$, $|\partial_x^2 m_Y^U(\tau, 0^+)| \leq \bar{B}_{Y+}^U(\tau)$, $|\partial_x^2 m_D(0^-)| \leq \bar{B}_{D-}$, and $|\partial_x^2 m_D(0^+)| \leq \bar{B}_{D+}$. Based on the expressions given after Theorem 2.1, the leading bias terms of the estimators $\widehat{g}^*(\tau, \gamma)$ can bounded using these constants. For example,

$$\left|B^{LU}(\tau,\gamma)h^2\right| \leq \frac{1}{2}\bar{\mu}\left(\bar{B}_{Y+}^L(\tau) + \frac{c}{1-\tau}\bar{B}_D + \bar{B}_{Y-} + c\bar{B}_{D-}\right)h^2 \equiv \bar{b}^{LU}(\tau,\gamma).$$

Following this reasoning, for $* \in \{LL, UU, LU, UL\}$, we obtain $\bar{b}^*(\tau,\gamma)$ such that $|B^*(\tau,\gamma)| \leq \bar{b}^*(\tau,\gamma)$. Let $\widehat{r}^*(\tau,\gamma) = \bar{b}^*(\tau,\gamma)/\widehat{\text{se}}\left(\widehat{g}^*(\tau,\gamma)\right)$ be the maximal value of the ratio of the leading asymptotic bias to the standard error of $\widehat{g}^*(\tau,\gamma)$.

For $\gamma \geq 0$, we define critical values $c_\alpha^{BA}(\tau,\gamma)$ as the solution to the following equation:

$$\Phi\left(c_\alpha^{BA}(\tau,\gamma) + \frac{\widehat{\mathcal{D}}(\tau,\gamma)}{\max\{\widehat{\text{se}}(\widehat{g}^{LU}(\tau,\gamma)), \widehat{\text{se}}(\widehat{g}^{UL}(\tau,\gamma))\}} + \max\left\{\widehat{r}^{UL}(\tau,\gamma), \widehat{r}^{LU}(\tau,\gamma)\right\}\right)$$
$$-\Phi\left(-c_\alpha^{BA}(\tau,\gamma) + \max\left\{\widehat{r}^{UL}(\tau,\gamma), \widehat{r}^{LU}(\tau,\gamma)\right\}\right) = 1-\alpha.$$

For $\gamma < 0$, we define $c_\alpha(\tau,\gamma)$ as the solution to the following equation:

$$\Phi\left(c_\alpha^{BA}(\tau,\gamma) + \frac{\widehat{\mathcal{D}}(\tau,\gamma)}{\max\{\widehat{\text{se}}(\widehat{g}^{LL}(\tau,\gamma)), \widehat{\text{se}}(\widehat{g}^{UU}(\tau,\gamma))\}} + \max\left\{\widehat{r}^{LL}(\tau,\gamma), \widehat{r}^{UU}(\tau,\gamma)\right\}\right)$$
$$-\Phi\left(-c_\alpha^{BA}(\tau,\gamma) + \max\left\{\widehat{r}^{LL}(\tau,\gamma), \widehat{r}^{UU}(\tau,\gamma)\right\}\right) = 1-\alpha.$$

As in the case of undersmoothing, we build our CS by collecting the values γ for which the hypothesis $H_0 : \gamma \in \Gamma_I(\tau)$ is not rejected:

$$\begin{aligned} CS_\alpha^{BA}(\tau) =&\{\gamma \geq 0 : -c_\alpha^{BA}(\tau,\gamma) \leq t^{UL}(\tau,\gamma),\, t^{LU}(\tau,\gamma) \leq c_\alpha^{BA}(\tau,\gamma)\} \\ &\cup\{\gamma < 0 : -c_\alpha^{BA}(\tau,\gamma) \leq t^{UU}(\tau,\gamma),\, t^{LL}(\tau,\gamma) \leq c_\alpha^{BA}(\tau,\gamma)\}. \end{aligned} \tag{2.5.6}$$

If $\widehat{\mathcal{D}}(\tau,\gamma)$ is large relative to the standard errors, then $c_\alpha^{BA}(\tau,\gamma)$ is approximately the $1-\alpha$ quantile of $\mathcal{N}(\max\left\{\widehat{r}^{UL}(\tau,\gamma), \widehat{r}^{LU}(\tau,\gamma)\right\}, 1)$ for $\gamma \geq 0$, and similarly for $\gamma < 0$. This definition is similar to critical values for one-sided, bias-aware CIs of Armstrong and Kolesár (2020). If $\widehat{\mathcal{D}}(\tau,\gamma) = 0$, then $c_\alpha^{BA}(\gamma)$ equals the $1-\alpha/2$ quantile of the folded normal distribution $|\mathcal{N}(\max\left\{\widehat{r}^{UL}(\tau,\gamma), \widehat{r}^{LU}(\tau,\gamma)\right\}, 1)|$ for $\gamma \geq 0$, and similarly for $\gamma < 0$. This definition is similar to critical values for two-sided CIs of Armstrong and Kolesár (2020).

2.5.2. **Estimated Manipulation Level.** In settings where sizable manipulation clearly occurs in the data, we conduct inference with an estimated manipulation level analogously to the previous subsection, based on test statistics of the form:

$$t^*_{\text{est}}(\gamma) = \frac{\widehat{g}^*(\widehat{\tau}, \gamma)}{\widehat{\text{se}}(\widehat{g}^*(\widehat{\tau}, \gamma))}.$$

With undersmoothing, critical values can be obtained as in (2.5.3)–(2.5.4) using the standard error $\widehat{\text{se}}(\widehat{g}^*(\widehat{\tau}, \gamma))$. For bias-aware inference, the maximal bias-to-standard-error ratio has to take into account the additional bias $C^*(\tau)B_\tau h^2$ due to estimation of τ, which can be bounded if $|f''_X(0^+)|$ and $|f''_X(0^-)|$ are bounded by some known constants.

The asymptotic, normal approximation of the distribution of the estimator $\widehat{g}^*(\widehat{\tau}, \gamma)$, presented in Theorem 2.2, is reliable if τ is well separated from zero. Following the moment inequality literature (e.g. Andrews and Soares, 2010), we can establish whether this is the case by conducting a conservative test of the hypothesis of no manipulation with significance level slowly converging to zero as the sample size grows. If the manipulation is not clearly detectable in the data, one could design a conservative procedure in the spirit of the bootstrap procedure proposed by GRR by 'tilting' the estimator $\widehat{\tau}$ away from zero. We leave this for future research.

2.6. SIMULATIONS

In this section, we investigate the performance of the testing procedure used to construct our proposed CSs in a simulation study. We consider two settings: estimation with an estimated manipulation level when the true manipulation level is sizable and a sensitivity analysis for any level of manipulation.

The data is generated from a model with the two-group structure discussed in Section 2.2. Among the potentially assigned units, there is 10% of always-takers, who are treated regardless of the value of their running variable, 10% of never-takers, who are never treated, and 80% of compliers, who receive the treatment if and only if their running variable is above the cutoff. In each of these three groups, the running variable X_i is distributed uniformly on $[-1, 1]$. The running variable of always-assigned units is distributed uniformly on $[0, 1]$, and these units are always treated. The outcome variable Y_i is generated as follows:

$$Y_i = \mu_{T_i} + X_i \mathbb{1}(0 \leq X_i) - X_i \mathbb{1}(X_i < 0) + \frac{q}{2}(X_i^2 \mathbb{1}(0 \leq X_i) - X_i^2 \mathbb{1}(X_i < 0)) + \varepsilon_i,$$

where $\varepsilon_i \sim Unif([-1, 1])$ and $T_i \in \{AA, AT, NT, CO\}$ denotes the unit type, which corresponds to always-assigned (AA) units and potentially-assigned: always-takers (AT), never-takers (NT), and compliers (CO). We set $\mu_{NT} = -1$, $\mu_{AT} = 1$, and $\mu_{CO} = 0$. The compliers have the same outcome in the presence and in the absence of treatment, so that there is no treatment effect. The value of μ_{AA} varies across simulation settings. The model residuals are homoskedastic, so that the second derivative of truncated conditional expectation functions equals $q \cdot \text{sgn}(X_i)$ for all truncation quantile levels.

In the setting with a known manipulation level, we consider five different manipulation levels, $\tau \in \{0, 0.01, 0.1, 0.2, 0.3\}$, and the sample size is $n = 10,000$. In the setting with an estimated manipulation level, $\tau \in \{0.1, 0.2, 0.3\}$ and $n = 50,000$, which ensures that τ is sufficiently bounded away from zero to rely on the asymptotic results from Section 2.4. The same bandwidth $h = 0.5$ is used in all steps.

Table 2.2: Simulation results.

		Coverage					Critical values				
	τ:	0	0.01	0.1	0.2	0.3	0	0.01	0.1	0.2	0.3
A) Known manipulation level, $q = 0$											
	-4	94.9	94.5	94.7	95.0	94.9	1.96	1.67	1.64	1.64	1.64
	-2	95.4	95.3	96.8	97.8	97.5	1.96	1.69	1.64	1.64	1.64
μ_{AA}	0	95.5	96.5	100.0	100.0	100.0	1.96	1.70	1.64	1.64	1.64
	2	95.2	94.9	97.1	97.7	97.8	1.96	1.69	1.64	1.64	1.64
	4	95.3	94.5	95.8	95.3	94.6	1.96	1.67	1.64	1.64	1.64
B) Estimated manipulation level, $q = 0$											
	-4	-	-	96.5	95.6	95.4	-	-	1.65	1.64	1.64
	-2	-	-	98.1	98.4	98.6	-	-	1.65	1.64	1.64
μ_{AA}	0	-	-	100.0	100.0	100.0	-	-	1.65	1.64	1.64
	2	-	-	98.3	98.2	98.3	-	-	1.64	1.64	1.64
	4	-	-	96.4	95.6	94.9	-	-	1.65	1.64	1.64
C) Known manipulation level, $q = 1$											
	-4	94.4	94.5	94.7	95.0	94.9	2.23	2.18	2.07	2.01	1.96
	-2	94.7	95.3	96.8	97.8	97.4	2.23	2.19	2.14	2.11	2.07
μ_{AA}	0	95.1	98.1	100.0	100.0	100.0	2.22	2.20	2.20	2.19	2.18
	2	94.7	98.9	99.7	99.7	99.8	2.23	2.19	2.14	2.11	2.08
	4	94.8	99.2	99.5	99.1	98.5	2.23	2.18	2.07	2.01	1.96
D) Estimated manipulation level, $q = 1$											
	-4	-	-	98.3	96.7	95.9	-	-	2.13	2.06	2.03
	-2	-	-	98.8	98.6	98.6	-	-	2.22	2.19	2.20
μ_{AA}	0	-	-	100.0	100.0	100.0	-	-	2.33	2.38	2.43
	2	-	-	100.0	100.0	100.0	-	-	2.22	2.19	2.20
	4	-	-	99.5	99.0	98.5	-	-	2.13	2.06	2.03

Notes: The results are based on $S = 5,000$ simulations. The sample size equals $n = 10,000$ in Panels A and C, and $n = 50,000$ in Panels B and D.

We consider five different level shifts of the outcome distribution of the always-assigned units, $\mu_{AA} \in \{-4, -2, 0, 2, 4\}$. The value of μ_{AA} determines whether one of the extreme scenarios considered when bounding Δ corresponds to the true location of the always-assigned units. The outcomes of always-assigned units are ‘separated’ from the outcomes of the potentially-assigned units if $\mu_{AA} \in \{-4, 4\}$. In these scenarios, the true treatment effect, which equals zero, lies on the boundary of the identified set $\Gamma_I(\tau)$.

In Table 2.2, we report the proportion of draws in which the true null hypothesis is not rejected, i.e. the coverage of the proposed CSs, and the associated critical values. In Panel A, the coverage is close to 95% if $\mu \in \{-4, 4\}$. If zero is well in the interior of the identified set, then the test never rejects. The critical values adapt to the length of the identified set, which changes with the true manipulation level τ. If $\tau = 0$, then the critical values correspond to a two-sided test. If $\tau \geq 0.1$, they correspond to a one-sided test. The construction used ensures that they change smoothly between 1.64 and 1.96 as τ approaches zero. In Panel C, the critical values are inflated due to accounting for the smoothing bias. For $\tau = 0$, the coverage is maintained at 95%. For positive values of τ, the coverage is no longer symmetric in τ because the worst-case bias realizes only for one of the bounds. The results with an estimated manipulation level in Panels B and D exhibit similar patterns to the corresponding results in Panels A and C in terms of coverage. The critical values in Panel D are smaller than in Panel C because the ratio of the worst-case bias to the standard error is smaller in the former setting.[4]

2.7. CONCLUSIONS

In this paper, we propose a method of conducting inference on a meaningful local average treatment effect in fuzzy RD designs with a manipulated running variable. It combines simple bounds on the numerator and the denominator of the Wald ratio for the subpopulation of interest with Anderson-Rubin-type confidence sets.

[4] In this particular simulation setting, estimation of the density limits does not produce any additional bias, but it increases the variance despite the larger sample size.

Appendix

2.A. PROOFS

2.A.1. **Proof of Theorem 2.1.** The theorem follows directly from inequalities (2.3.6) and (2.3.7).

2.A.2. **Proof of Theorem 2.2.** Let $v_n = h^2 + (nh)^{-1/2}$.
Part (i). It follows from Theorem 1.1 in Chapter 1 that

$$\widehat{\mathbb{E}}\Big[G^*_{+,i}\Big(\tau, \widehat{Q}_{Y|X}, \gamma\Big)\Big|X_i = 0^+\Big] = \widehat{\mathbb{E}}\Big[G^*_{+,i}\Big(\tau, Q_{Y|X}, \gamma\Big)\Big|X_i = 0^+\Big] + o_p(v_n).$$

Asymptotic normality then follows from standard theory of local linear estimation.
Part (ii). Under the assumptions made, it holds that $\widehat{\tau} = \tau + O_p(v_n)$. By Theorem 1.2 in Chapter 1, for $k \in \{L, U\}$ it holds that

$$\widehat{\mathbb{E}}\Big[\psi^k_i\Big(\tau, \widehat{Q}_{Y|X}\Big)\Big|X_i = 0^+\Big] = \widehat{\mathbb{E}}\Big[\psi^k_i\Big(\tau, Q_{Y|X}\Big)\Big|X_i = 0^+\Big] + \partial_\tau m^k_{Y+}(\tau)(\widehat{\tau} - \tau) + o_p(v_n).$$

Moreover, we have that

$$\frac{1}{1-\widehat{\tau}} - \frac{1}{1-\tau} = \frac{\widehat{\tau}-\tau}{(1-\tau)^2} + o_p(v_n) \text{ and } \frac{\widehat{\tau}}{1-\widehat{\tau}} - \frac{\tau}{1-\tau} = \frac{\widehat{\tau}-\tau}{(1-\tau)^2} + o_p(v_n).$$

Hence, using linearity of the local linear estimator, we obtain that:

$$\begin{aligned}
\widehat{\mathbb{E}}\left[\frac{D_i - \widehat{\tau}}{1-\widehat{\tau}}\middle|X_i = 0^+\right] &= \widehat{\mathbb{E}}\left[\frac{D_i - \tau}{1-\tau}\middle|X_i = 0^+\right] + \frac{m_{D+} - 1}{(1-\tau)^2}(\widehat{\tau} - \tau) + o_p(v_n),\\
\widehat{\mathbb{E}}\left[\frac{D_i}{1-\widehat{\tau}}\middle|X_i = 0^+\right] &= \widehat{\mathbb{E}}\left[\frac{D_i}{1-\tau}\middle|X_i = 0^+\right] + \frac{m_{D+}}{(1-\tau)^2}(\widehat{\tau} - \tau) + o_p(v_n),
\end{aligned}$$

which concludes the proof.

2.A.3. **Validity of CSs.** We consider bias-aware CSs. Undersmoothing is obtained as a special case when $\widehat{r}^*(\tau, \gamma) = 0$. First, note that:

(i) For $\gamma \geq 0$, $\widehat{g}^{UL}(\tau, \gamma) - \widehat{g}^{LU}(\tau, \gamma) = \widehat{\Delta}^U(\tau) - \widehat{\Delta}^L(\tau) + \gamma\frac{\tau}{1-\tau} = \widehat{\mathcal{D}}(\tau, c)$,

(ii) For $\gamma < 0$, $\widehat{g}^{UU}(\tau, \gamma) - \widehat{g}^{LL}(\tau, \gamma) = \widehat{\Delta}^U(\tau) - \widehat{\Delta}^L(\tau) - \gamma\frac{\tau}{1-\tau} = \widehat{\mathcal{D}}(\tau, c)$.

where $\widehat{\mathcal{D}}(\tau, c) = \widehat{\Delta}^U(\tau) - \widehat{\Delta}^L(\tau) + |\gamma|\frac{\tau}{1-\tau}$, as defined in Section 2.5.

Suppose that $\gamma \geq 0$ and $g^{UL}(\tau, \gamma) = 0$. For any $a > 0$, it holds that

$$\begin{aligned}
&\mathbb{P}\Big(- a \leq t^{UL}(\tau, \gamma), t^{LU}(\tau, \gamma) \leq a\Big) \\
&= \mathbb{P}\left(- a \leq \frac{\widehat{g}^{UL}(\tau, \gamma) - g^{UL}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{UL}(\tau, \gamma))}, \frac{\widehat{g}^{UL}(\tau, \gamma) - g^{UL}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{LU}(\tau, \gamma))} \leq a + \frac{\widehat{g}^{UL}(\tau, \gamma) - \widehat{g}^{LU}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{LU}(\tau, \gamma))}\right) \\
&= \mathbb{P}\left(- a \leq \mathcal{Z} + \frac{B^{UL}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{UL}(\tau, \gamma))} \leq a + \frac{\widehat{\mathcal{D}}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{LU}(\tau, \gamma))}\right),
\end{aligned}$$

where $\mathcal{Z} \sim \mathcal{N}(0, 1)$ and $|B^{UL}(\gamma)|/\widehat{\mathrm{se}}(\widehat{g}^{UL}(\gamma)) \leq \widehat{r}^{UL}(\gamma)$. The last expression is the smallest when the normal distribution is "shifted" maximally to the left, so that

$$\begin{aligned}
&\mathbb{P}(-a \leq t^{UL}(\tau, \gamma), t^{LU}(\tau, \gamma) \leq a) \\
&\quad \geq \mathbb{P}\left(-a + \widehat{r}^{UL}(\tau, \gamma) \leq \mathcal{Z} \leq a + \frac{\widehat{\mathcal{D}}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{LU}(\tau, \gamma))} + \widehat{r}^{UL}(\tau, \gamma)\right) + o(1).
\end{aligned}$$

Similarly, if $\gamma \geq 0$ and $g^{LU}(\tau, \gamma) = 0$, then for any $a > 0$, it holds that

$$\begin{aligned}
&\mathbb{P}(-a \leq t^{UL}(\tau, \gamma), t^{LU}(\tau, \gamma) \leq a) \\
&\quad \geq \mathbb{P}\left(-a + \widehat{r}^{LU}(\tau, \gamma) \leq \mathcal{Z} \leq a + \frac{\widehat{\mathcal{D}}(\tau, \gamma)}{\widehat{\mathrm{se}}(\widehat{g}^{UL}(\tau, \gamma))} + \widehat{r}^{LU}(\tau, \gamma)\right) + o(1).
\end{aligned}$$

In our pointwise asymptotics, $\widehat{\mathcal{D}}(\tau, \gamma)/\widehat{\mathrm{se}}(\widehat{g}^*(\tau, \gamma))$ is either zero (if $\tau = 0$), or it diverges to infinity. The ratios $\widehat{r}^*(\tau, \gamma)$ converge to a finite constant or diverge to infinity. With the definition of the critical values in the main text, the test is asymptotically valid. If $g^{LU}(\tau, \gamma) < 0 < g^{UL}(\tau, \gamma)$, then the nonrejection probability converges to one. The same reasoning applies when $\gamma < 0$.

CHAPTER 3

Flexible Covariate Adjustments in Regression Discontinuity Designs

Joint work with Claudia Noack and Christoph Rothe.

3.1. INTRODUCTION

Regression discontinuity (RD) designs are widely used for estimating causal treatment effects from observational data in economics and other social sciences. In a sharp RD design, the treatment status is determined by whether the running variable exceeds a fixed cutoff value. Under standard assumptions, the average treatment effect at the cutoff is identified by the size of the jump in the conditional expectation of the outcome variable given the running variable at the cutoff. This parameter is typically estimated using local linear regression methods, and various inference procedures have been proposed in the literature; see, e.g., Imbens and Kalyanaraman (2012), Calonico et al. (2014), and Armstrong and Kolesár (2020).

The standard estimator of the average treatment effect in sharp RD designs is based solely on the outcome variable and the running variable, but in many empirical applications, researchers include additional, pretreatment covariates linearly in the RD regression to reduce the variance of the estimates (see Calonico et al., 2019). However, linear adjustments in general do not fully exploit the information contained in the covariates. The goal of this paper is to improve upon these methods.

We propose a novel class of covariate-adjusted RD estimators. They are constructed in two stages. In the first stage, we obtain adjustment terms, which aim at capturing the variation in the outcome variable near the cutoff that can be explained by the additional covariates. The adjustment terms are estimated using cross-fitting, which allows us to use a wide range of methods in the first stage under weak conditions. We generate our covariate-adjusted outcome variable by subtracting the adjustment terms from the original outcomes. In the second stage, we estimate the RD parameter in a local linear regression with our generated outcome variable.

Our proposed approach is based on the premise that in a valid RD design, the conditional distribution of the additional covariates given the running variable should evolve continuously through the cutoff. Such a condition is inherently related to the standard, behavioral identification arguments in RD designs, which postulate that the units just to the left and just to the right of the cutoff are very similar in all pretreatment characteristics.[1] Based on this feature, we can adjust our outcome variable by subtracting from it essentially any function of the additional covariates without changing the RD estimand. We can further choose the adjustment function that leads to the smallest variance of the RD estimator in the considered class of estimators. We find that the optimal adjustment function is given by the average of the conditional expectations of the outcome variable just to the left and just to the right of the cutoff given the additional covariates. This function is not known, and therefore we estimate it in the first stage.

An important feature of our proposed RD estimator is that it is very insensitive to the first-stage estimation error, which has the following important, practical and theoretical implications. First, we only require that the first-stage estimator concentrates, possibly very slowly, in a mean-squared-error-type sense around some deterministic sequence of functions. This condition is satisfied for a wide range of estimators, including parametric estimators, classic nonparametric methods, such as local linear and sieve estimators (Fan and Gijbels, 1996; Newey, 1997), as well as modern machine learning methods, such as lasso (Tibshirani, 1996), random forests (Breiman, 2001; Wager and Athey, 2018), and deep neural networks (Farrell et al., 2021). Importantly, our RD estimator is not very sensitive to the specific choice of the tuning parameters that are required for some of the above methods.

Second, in our asymptotic analysis, we can ignore the fact that the adjustment terms are estimated in the first stage. Our proposed RD estimator is asymptotically equivalent to an estimator employing the deterministic function around which the first-stage estimator concentrates. As a result, existing procedures for inference and bandwidth choice can be directly applied to the second-stage regression. Specifically, we obtain the standard error using the nearest-neighbors method. We also argue that one can choose the bandwidth and construct confidence intervals following the robust bias corrections approach of Calonico et al. (2014) or the bias-aware procedure of Armstrong and Kolesár (2020).

We further show that if the first-stage estimator consistently estimates the targeted

[1] Indeed, in empirical applications, testing continuity of the distribution of baseline covariates at the cutoff has become a standard way of assessing the validity of an RD design (Cattaneo et al., 2019).

conditional expectations, then our estimator is efficient in the considered class, but our asymptotic results remain valid whether or not this condition is satisfied. Our proposed covariate adjustments asymptotically lead to variance reductions compared to the standard RD estimator whenever the covariates have explanatory power for the outcome variable in a neighborhood of the cutoff.

Our proposed procedure is related to covariate adjustments used in randomized experiments to improve efficiency of the average treatment effect estimator (see, e.g., Wager et al., 2016). This analogy occurs because RD designs are similar in nature to randomized experiments. In randomized experiments, comparability of the treated and untreated units is ensured by random assignment, whereas in RD designs, it is ensured for units close to the cutoff by continuity of potential outcomes' distributions. Our proposed RD estimator has a very similar structure as the augmented inverse probability weighted estimator, which is widely used in randomized experiments. Accordingly, the minimal variance that our estimator can achieve resembles the efficiency bound for estimation of the average treatment effect under unconfoundedness (Hahn, 1998).

Literature. There exists an extensive literature on estimation in RD designs; see, e.g., Imbens and Lemieux (2008) and Cattaneo et al. (2019) for a textbook treatment. In general, existing methods do not require covariate information, but it is standard practice to incorporate covariates in order to reduce the variance of the estimates (see, e.g., Lee and Lemieux, 2010, Section 3.2.3). We contrast our approach with two papers that are most closely related to our approach.

Calonico et al. (2019) employ a local linear regression in the running variable with additional covariates included in a linear fashion. We allow for linear adjustments as a special case, but we cover a wide range of other, more flexible adjustments that improve efficiency compared to simple linear adjustments. We discuss the relation of our approach to that of Calonico et al. (2019) in more detail in Section 3.6.1.

Frölich and Huber (2019) propose a procedure using first-stage nonparametric predictions of the treatment effect conditional on the additional covariates at the cutoff, which achieves approximately the same variance as our estimator in some settings. However, their results rely on strong assumptions about the number of covariates and/or smoothness of the conditional expectation of the outcome variable given the covariates, which are not needed for our method.[2]

Our paper is also related to the literature on two-stage estimation with nuisance parameters (Andrews, 1994; Newey, 1994). The combination of locally-robust moment

[2] For example, Frölich and Huber (2019) allow for at most three continuous additional covariates and require that the bandwidth converges at a specific rate if the local linear estimator with a second-order kernel is used in the first stage.

conditions and cross-fitting has been used, e.g., by Belloni et al. (2017); Chernozhukov et al. (2018). Estimation of conditional treatment effects with orthogonal moments have been studied, e.g., by Kennedy et al. (2017); Kennedy (2020); Fan et al. (2020).

Plan of the Paper. The remainder of this paper is organized as follows. In Section 3.2, we introduce the setup. In Section 3.3, we present our proposed covariate-adjusted estimator. In Section 3.4, we present our main theoretical results under general conditions on the covariate adjustments used. We discuss implementation details in Section 3.5. In Section 3.6, we consider specific examples of covariate adjustments. We present a simulation study in Section 3.7. Section 3.8 concludes.

Notation. Throughout the paper, we use the following notation. For a generic function $f(x)$, we write $f(0^+) = \lim_{x\downarrow 0} f(x)$ and $f(0^-) = \lim_{x\uparrow 0} f(x)$ for the right and left limit of the function $f(x)$ at zero, respectively.

3.2. SETUP

In this section, we introduce the model and parameter of interest. Furthermore, we discuss estimation of the RD parameter based on local linear regression methods.

3.2.1. **Model and Parameter of Interest.** We consider a sharp RD design, in which the researcher investigates the causal effect of a binary treatment on some outcome variable of interest. The data $(W_i)_{i\in\{1,\ldots,n\}}$ are an i.i.d. sample of size n from the distribution of $W_i = (Y_i, X_i, Z_i)$. Here, $Y_i \in \mathbb{R}$ is the outcome variable, $X_i \in \mathbb{R}$ is the running variable, and $Z_i \in \mathbb{R}^d$ is a vector of additional covariates. Units receive the treatment if and only if the running variable exceeds some known threshold, which we normalize to zero without loss of generality. We denote the treatment indicator by T_i, so that $T_i = \mathbf{1}\{X_i \geq 0\}$.

Throughout the paper, we assume that the distribution of the running variable X_i is fixed, but we consider a sequence of conditional distributions of (Y_i, Z_i) given X_i that can change with n. In particular, we allow the dimension of Z_i to grow with n. For ease of notation, we leave the dependence on n implicit.

We denote the support of Z_i by $\mathcal{Z}$, and we let $\mathcal{X}$ be an open neighborhood of the cutoff that is contained in the support of the running variable. The density of the running variable is denoted by f_X, the conditional cumulative distribution function of Z_i given $X_i = x$ is denoted by $F_{Z|X}(z|x)$. If the corresponding conditional density exists, we denote it by $f_{Z|X}(z|x)$. Under standard assumptions (see, e.g., Lee and Lemieux, 2010) the average treatment effect at the cutoff is identified by the height of the jump in the conditional expectation of the observed outcome variable given the running variable at zero:

$$\tau = \mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0^-]. \tag{3.2.1}$$

We take this identification result as given and consider estimation of τ as defined above.

3.2.2. **Standard RD Estimator.**

In RD designs, the parameter of interest is typically estimated using local linear regression (see, e.g., Fan and Gijbels, 1996). The standard estimator is given by:

$$\widehat{\tau}(h) = e_1^\top \operatorname*{arg\,min}_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K(X_i/h)(Y_i - \beta^\top (T_i, X_i, T_i X_i, 1))^2,$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$, $h > 0$ is a bandwidth, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector. Using simple algebra, this estimator can be expressed as a weighted sum of the outcome variable:

$$\widehat{\tau}(h) = \sum_{i=1}^{n} w_i(h) Y_i,$$

where the weights $w_i(h)$ depend only on the realizations of the running variable. We give the explicit expressions for the weights in Appendix 3.C.1.

Under standard assumptions, the estimator $\widehat{\tau}(h)$ is asymptotically normally distributed. Its leading bias term is proportional to $\partial_x^2 \mathbb{E}[Y_i | X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i | X_i = x]|_{x=0^-}$, and it is of order h^2. The bias results from approximating the possibly non-linear conditional expectation function with a linear function. Its magnitude is determined by the degree of nonlinearity, measured by the value of the second derivative. The variance is of order $(nh)^{-1}$, and it is approximately proportional to $\mathbb{V}[Y_i | X_i = 0^+] + \mathbb{V}[Y_i | X_i = 0^-]$.

3.3. COVARIATE ADJUSTMENTS

In this section, we motivate our proposed estimation procedure, and we formally define the proposed covariate-adjusted RD estimator.

3.3.1. **Covariate-Adjusted Outcome Variable.**

We now introduce the key object of this paper. We consider a modified outcome variable of the following form:

$$M_i(\mu) = Y_i - \mu(Z_i), \tag{3.3.1}$$

where μ is a real-valued function of the additional covariates, which we refer to as the adjustment function.

For the further analysis, we impose a regularity condition on the admissible adjustment functions and require that $\mu(Z_i)$ is square integrable conditional on the running variable. We define the set of such functions as:

$$\mathcal{M}_n = \Big\{ \mu : \mathcal{Z} \to \mathbb{R} \text{ s.t. } \sup_{x \in \mathcal{X}} \mathbb{E}[\mu(Z_i)^2 | X_i = x] < \infty \Big\}.$$

The central premise for our proposed approach is that the conditional distribution of the additional covariates given the running variable evolves continuously through the cutoff.

Assumption 3.1. *For all $n \in \mathbb{N}$ and $\mu \in \mathcal{M}_n$, $\mathbb{E}\left[\mu(Z_i)|X_i = x\right]$ is continuous in x on $\mathcal{X}$.*

Assumption 3.1 requires that the conditional distribution of Z_i given $X_i = x$ converges weakly to the distribution of Z_i given $X_i = 0$, as x converges to 0.[3] Under this assumption, we can replace the outcome variable Y_i in the definition of τ in (3.2.1) with $M_i(\mu)$ without affecting the value of the estimand, that is:

$$\tau = \mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-] \tag{3.3.2}$$

for all $\mu \in \mathcal{M}_n$.

Motivated by the above representation, for any fixed $\mu \in \mathcal{M}_n$, the RD parameter τ could be estimated using the local linear RD estimator with $M_i(\mu)$ as the outcome variable, which we denote by:

$$\widehat{\tau}(h; \mu) = \sum_{i=1}^{n} w_i(h) M_i(\mu). \tag{3.3.3}$$

In practice, the adjustment function might be estimated from the data. However, in a sense made precise in the next sections, we can replace the deterministic adjustment function with its estimate without affecting the first-order asymptotic properties of the final estimator of the RD parameter. We therefore first determine the adjustment function that minimizes the variance of the RD estimator $\widehat{\tau}(h; \mu)$.

3.3.2. **Optimal Adjustment Function.** The RD estimator $\widehat{\tau}(h; \mu)$ has variance that is approximately proportional to $\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]$. We find that the adjustment function that minimizes this expression is given by

$$\mu_n(z) = \frac{1}{2}\left(\mu_n^+(z) + \mu_n^-(z)\right), \tag{3.3.4}$$

where $\mu_n^+(z) = \mathbb{E}[Y_i|X_i = 0^+, Z_i = z]$ and $\mu_n^-(z) = \mathbb{E}[Y_i|X_i = 0^-, Z_i = z]$. This result follows from simple derivations, which we outline below to present the intuition behind this result.

Under Assumption 3.1, if μ_n^-, μ_n^+, $\mu \in \mathcal{M}_n$, then

$$\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-] = \mathbb{V}[M_i(\mu_n^+)|X_i = 0^+] + \mathbb{V}[M_i(\mu_n^-)|X_i = 0^-] + \mathcal{V}(\mu),$$

where the first two terms on the right-hand side do not depend on μ, and

$$\mathcal{V}(\mu) = \mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0] + \mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0].$$

[3] This condition holds if $F_{Z|X}(z|x) \to F_{Z|X}(z|0)$, as $x \to 0$, for all continuity points of $F_{Z|X}(z|0)$.

Our goal is therefore to minimize $\mathcal{V}(\mu)$. Each component of $\mathcal{V}(\mu)$ could be set to zero separately if μ was chosen as μ_n^+ or μ_n^-, respectively. It turns out that the whole expression $\mathcal{V}(\mu)$ is minimized by the function μ_n, which can be seen by noting that:

$$\mathcal{V}(\mu) = \mathcal{V}(\mu_n) + 2\mathbb{V}[\mu_n(Z_i) - \mu(Z_i)|X_i = 0] \geq \mathcal{V}(\mu_n).$$

This reasoning shows that indeed the expression $\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]$ achieves the smallest value if $\mu = \mu_n$. The function μ_n is essentially a unique minimizer up to shifts by a constant.

3.3.3. **Estimator.** We estimate τ in a two-stage procedure. In the first stage, we estimate the function μ_n defined in (3.3.4), which involves estimating the limits of the conditional expectation of the outcome variable given the additional covariates as the running variable approaches the cutoff from the left and from the right.

Conditional expectations at boundary points are often estimated using local linear methods because of their good bias properties. However, for our purposes, essentially any procedure can be adapted to estimate these limits by restricting the data to observations with the running variable close to the cutoff.[4] For example, we can use parametric estimators, classic nonparametric methods such as series and spline estimators (Masry, 1996; Newey, 1997), as well as modern machine learning methods including the lasso (Tibshirani, 1996), random forests (Breiman, 2001; Wager and Athey, 2018), and deep neural networks (Farrell et al., 2021).

In order to allow for a variety of, possibly highly complex, first-stage estimators, we use cross-fitting (see, e.g., Chernozhukov et al., 2018).[5] We split the data randomly into S disjoint folds denoted I_s for $s \in [S] = \{1, ..., S\}$, where all folds have the same number of observations to the left of the cutoff, and similarly to the right of the cutoff.[6] For $s \in [S]$, we define the complement of fold I_s as $I_s^c = [n] \setminus I_s$. Further, let $s(i)$ denote the index of the fold containing observation i, so that $i \in s(i)$. Given a selected estimation procedure, we define $\widehat{\mu}_{n,s}(z) = \widehat{\mu}_n(z; (W_i)_{i \in I_s^c})$, which is an estimator of $\mu_n(z)$ that uses all observations except for the sth fold of the data.

In the second stage, we estimate the RD parameter using our covariate-adjusted outcome variable. For each observation, we generate the outcome using the first-stage

[4] In our asymptotic analysis, we require only that the first-stage estimator concentrates around some deterministic sequence.

[5] For simple first-stage estimators, such as linear adjustments, cross-fitting is not required, but it offers a unifying approach that is suitable for all considered types of adjustments.

[6] In simulations, we choose S to be a moderate number, e.g. 5. We assume that the number of observations both to the left and to the right of the cutoff is divisible by S to simplify the notation.

estimate based on data from other folds. The final estimator is defined as:

$$\widehat{\tau}_{CF}(h;\widehat{\mu}_n) = \sum_{i=1}^{n} w_i(h) M_i(\widehat{\mu}_{n,s(i)}), \tag{3.3.5}$$

where the subscript CF refers to cross-fitting.

3.4. THEORETICAL RESULTS

In this section, we formally study the properties of the estimator $\widehat{\tau}_{CF}(h;\widehat{\mu}_n)$ under high-level conditions on the first-stage estimator. We also propose a method to estimate its variance.

3.4.1. Assumptions.

The conditions we impose in this section consist of standard assumptions in RD designs without covariates as well as high-level assumptions on the first-stage estimator $\widehat{\mu}_n$. Low-level conditions, tailored to specific types of covariate adjustments, are discussed in Section 3.6. Throughout the paper, we implicitly assume that if a real-valued function f is continuous on $\mathcal{X} \setminus \{0\}$, then also the limits $f(0^-)$ and $f(0^+)$ exist and are finite.

Assumption 3.2. *(i) X_i is continuously distributed with density f_X, which is continuous and bounded away from zero uniformly over $x \in \mathcal{X}$; (ii) The kernel function K is a bounded and symmetric density function that is continuous on its support and equals zero outside some compact set, say $[-1,1]$; (iii) As $n \to \infty$, $h \to 0$ and $nh \to \infty$.*

Assumption 3.2 contains basic conditions for our asymptotic analysis. The assumptions on the density of the running variable, kernel, and bandwidth are standard in the literature.

The next two assumptions concern the first-stage estimator. By construction, its properties are relevant only for observations that are used in the second-stage local linear regression, i.e. the observations with $|X_i| \leq h$. We define $\mathcal{X}_h = \mathcal{X} \cap [-h,h]$ and $\mathcal{Z}_h = \operatorname{supp}(Z_i | X_i \in \mathcal{X}_h)$.

Assumption 3.3. *For all $n \in \mathbb{N}$, there exist a set $\mathcal{T}_n \subset \mathcal{M}_n$ and a function $\bar{\mu}_n \in \mathcal{T}_n$ such that: (i) $\widehat{\mu}_{n,s}$ belongs to $\mathcal{T}_n$ with probability approaching 1 for all $s \in [S]$; (ii) It holds that:*

$$\sup_{\mu \in \mathcal{T}_n} \sup_{x \in \mathcal{X}_h} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i = x\right] = o(1).$$

Assumption 3.3 specifies the required mode of convergence for the first-stage estimator. We require that it belongs with high probability to some realization set $\mathcal{T}_n \subset \mathcal{M}_n$, which contracts around a deterministic sequence of functions $(\bar{\mu}_n)_{n\in\mathbb{N}}$ in a mean-squared-error-type sense. This assumption is weak, as $\bar{\mu}_n$ can be any function, not necessarily the targeted, true function μ_n, and we do not require any specific rate at which $\mathcal{T}_n$ shrinks. In particular, we allow for $\widehat{\mu}_n$ to be based on a misspecified parametric model for the function

$\bar{\mu}_n$, or to have an arbitrarily slowly vanishing bias, as long as the estimator concentrates around some deterministic sequence.

Assumption 3.3 can be ensured in various ways. If the adjustment function is linear, then it follows from convergence of the estimated coefficients if the additional covariates have bounded conditional second moments. Assumption 3.3 is also satisfied if the difference $\widehat{\mu}_{n,s} - \bar{\mu}_n$ converges to zero in the supremum norm on $\mathcal{Z}_h$. Such results are available for example for classic nonparametric estimators in settings with a fixed dimension of the additional covariates. Assumption 3.3 follows also from the unconditional convergence in mean square under mild conditions on the conditional distribution of the additional covariates given the running variable, which can be used to verify this assumption for machine learning methods; see Section 3.6.4 and Appendix 3.A.1.

Assumption 3.4. *For all $n \in \mathbb{N}$, it holds that:*

(i) $\mathbb{E}\left[\mu(Z_i)|X_i = x\right]$ *is twice continuously differentiable in x on $\mathcal{X} \setminus \{0\}$ for all $\mu \in \mathcal{M}_n$;*

(ii) $\sup_{\mu \in \mathcal{T}_n} \sup_{x \in \mathcal{X}_h \setminus \{0\}} \left|\partial_x^1 \mathbb{E}\left[\mu(Z_i) - \bar{\mu}_n(Z_i)|X_i = x\right]\right| = o(1/h)$*;*

(iii) $\sup_{\mu \in \mathcal{T}_n} \sup_{x \in \mathcal{X}_h \setminus \{0\}} \left|\partial_x^2 \mathbb{E}\left[\mu(Z_i) - \bar{\mu}_n(Z_i)|X_i = x\right]\right| = o(1)$*.*

Part (i) strengthens Assumption 3.1 and requires that $\mathbb{E}[\mu(Z_i)|X_i = x]$ is twice continuously differentiable to the left and to the right of the cutoff. We emphasize that we do not require continuity of the derivatives of $\mathbb{E}[\mu(Z_i)|X_i = x]$ at the cutoff. This assumption is analogous to the assumptions of Calonico et al. (2019), who assume that $\mathbb{E}[Z_i|X_i = x]$ is (thrice in their case) continuously differentiable to the left and to the right of the cutoff but not necessarily at the cutoff.[7] If, however, $\partial_x^2 \mathbb{E}[\mu(Z_i)|X_i = x]$ is continuous at the cutoff, we can exploit this assumption to simplify our asymptotic results; see Corollary 3.1. Parts (ii) and (iii) impose high-level requirements on derivatives of $\mathbb{E}\left[\mu(Z_i) - \bar{\mu}_n(Z_i)|X_i = x\right]$ for $\mu \in \mathcal{T}_n$.

Assumption 3.4 follows from Assumption 3.3 under regularity conditions on the conditional distribution of the additional covariates given the running variable. Specific conditions may depend on the estimator used. If the adjustment function is linear, then it follows if each component of $\mathbb{E}[Z_i|X_i = x]$ is twice continuously differentiable on $\mathcal{X} \setminus \{0\}$. Assumption 3.4 also follows whenever the conditional density $f_{Z|X}(z|x)$ is bounded away from zero on its support and the partial derivatives $\partial_x^j f_{Z|X}(z|x)$ are L-Lipschitz continuous in x for all z and $j \in \{0, 1\}$. We discuss further, technical sufficient conditions for this assumption in Appendix 3.A.2.

[7] In their main analysis, Calonico et al. (2019) assume only that $\mathbb{E}[Z_i|X_i = x]$ is continuous also at the cutoff, which ensures consistency of the RD estimator. The higher-order smoothness assumptions ensure that standard theory of local linear estimation can be applied to their RD estimator.

Assumption 3.5. *There exist constants B and L such that the following conditions hold for all $n \in \mathbb{N}$. (i) $\mathbb{E}[M_i(\bar{\mu}_n)|X_i = x]$ is twice continuously differentiable on $\mathcal{X} \setminus \{0\}$ with L-Lipschitz continuous second derivative bounded by B; (ii) For all $x \in \mathcal{X}$ and some $q > 2$ $\mathbb{E}[(M_i(\bar{\mu}_n) - \mathbb{E}[M_i(\bar{\mu}_n)|X_i])^q|X_i = x]$ exists and is bounded by B; (iii) $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$ is L-Lipschitz continuous and bounded from below by $1/B$ for all $x \in \mathcal{X} \setminus \{0\}$.*

Assumption 3.5 is a translation of standard RD assumptions to the setting with $M_i(\bar{\mu}_n)$ as the outcome variable. We employ these conditions to show asymptotic normality of our proposed RD estimator and to characterize its bias. Part (i) requires that the conditional expectation of the outcome variable is twice continuously differentiable to the left and to the right of the cutoff. Parts (ii) and (iii) impose standard assumptions on conditional moments of the outcome variable.

3.4.2. **Main Asymptotic Results.** In this section, we study the asymptotic properties of our estimator. We define the following kernel constants: $\bar{\nu} = (\bar{\nu}_2^2 - \bar{\nu}_1\bar{\nu}_3)/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)$ and $\bar{\kappa} = \int_0^\infty (k(v)(\bar{\nu}_1 v - \bar{\nu}_2))^2 dv/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)^2$, where $\bar{\nu}_j = \int_0^\infty v^j k(v)dv$.

Theorem 3.1. *Suppose that Assumptions 3.1–3.4 hold.*

(i) It holds that

$$\widehat{\tau}_{CF}(h; \widehat{\mu}_n) = \widehat{\tau}(h; \bar{\mu}_n) + o_p(h^2 + (nh)^{-1/2}).$$

Suppose additionally that Assumption 3.5 holds.

(ii) It holds that

$$\sqrt{nh}\, V(\bar{\mu}_n)^{-1/2} \left(\widehat{\tau}_{CF}(h; \widehat{\mu}_n) - \tau - B(\bar{\mu}_n)h^2\right) \to \mathcal{N}(0, 1),$$

where for $\mu \in \mathcal{M}_n$

$$B(\mu) = \frac{1}{2}\bar{\nu} \left(\partial_x^2 \mathbb{E}[M_i(\mu)|X_i = x]\Big|_{x=0^+} - \partial_x^2 \mathbb{E}[M_i(\mu)|X_i = x]\Big|_{x=0^-}\right) + o_P(1),$$
$$V(\mu) = \frac{\bar{\kappa}}{f_X(0)} \left(\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]\right).$$

(iii) For all functions $\mu \in \mathcal{M}_n$, it holds that

$$V(\mu) \geq V(\mu_n) = \frac{\bar{\kappa}}{f_X(0)}\Big(\mathbb{E}\left[\mathbb{V}[Y_i|Z_i, X_i]|X_i = 0^+\right] + \mathbb{E}\left[\mathbb{V}[Y_i|Z_i, X_i]|X_i = 0^-\right]$$
$$+ \frac{1}{2}\mathbb{V}\left[\mu_n^+(Z_i) - \mu_n^-(Z_i)|X_i = 0\right]\Big).$$

Part (i) states the key technical result. It shows that the proposed estimator is asymptotically equivalent to its infeasible analog with the estimator $\widehat{\mu}_n$ replaced with the deterministic sequence $\bar{\mu}_n$. We emphasize that this equivalence holds even though the first-stage estimator can converge arbitrarily slowly. This high insensitivity is only possible because for all $k \in \mathbb{N}$

$$\partial_\mu^k \left(\mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-] \right) |_{\mu=\mu_n} = 0 \tag{3.4.1}$$

where ∂_μ^k is the k-th functional derivative with respect to the function μ. This property is in the spirit of Neyman orthogonality with respect to the adjustment function μ. We discuss it further in Appendix 3.B.3.

Based on the asymptotic equivalence result in part (i), the asymptotic normality shown in part (ii) follows from standard theory of local linear estimation. The approximate variance depends on the sequence $\bar{\mu}_n$ around which the first-stage estimator concentrates. If $\bar{\mu}_n = \mu_n$, then the variance expression is similar to the efficiency bound for estimation of the average treatment effect under unconfoundedness with a constant conditional probability of treatment equal to one half (Hahn, 1998). We discuss the analogy between the covariate adjustments used for randomized experiments and our approach in Appendix 3.B.2.

The proposed covariate adjustments lead to efficiency gains compared to the standard RD estimator in a very wide range of settings, even if $\bar{\mu}_n \neq \mu_n$. We show in Appendix 3.D that $V(\bar{\mu}_n) < V(0)$ if and only if $\mathbb{V}[\mu_n(Z_i) - \bar{\mu}_n(Z_i)|X_i = 0] < \mathbb{V}[\mu_n(Z_i)|X_i = 0]$, i.e. whenever $\bar{\mu}_n(Z_i)$ has some explanatory power for $\mu_n(Z_i)$. This condition is satisfied for example if $\bar{\mu}_n(Z_i)$ represents some nontrivial projection of Y_i on Z_i based on the data in a neighborhood of the cutoff.

The bias expression simplifies under an additional smoothness assumption. If the smoothness condition in Assumption 3.4(i) holds also at the cutoff, then the leading bias does not depend on the function $\bar{\mu}_n$. The simplified bias expression is convenient for conducting statistical inference on based the bias-aware approach; see Section 3.5.2.

Corollary 3.1. *Suppose that Assumptions 3.1–3.5 hold and $\partial_x^2 \mathbb{E}[\bar{\mu}_n(Z_i)|X_i = x]$ is continuous at the cutoff for all $n \in \mathbb{N}$. Then*

$$B(\bar{\mu}_n) = \frac{1}{2}\bar{\nu} \left(\partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^-} \right) + o_P(1).$$

Remark 3.1. It follows from the proof of Theorem 3.1 that our proposed estimator is asymptotically equivalent to the average of RD estimators run on different folds of the data.[8] We prefer our version because existing estimation and inference routines as well as bandwidth selectors can be readily applied to the modified data $(M_i(\widehat{\mu}_{n,s(i)}), X_i)_{i\in[n]}$; see Section 3.5.

[8] A similar point is made by Chernozhukov et al. (2018) in the context of the (unconditional) average treatment effect estimation; cf. their methods DML1 and DML2. Fan et al. (2020) average local linear estimators run on different folds of the data in a conditional average treatment effect estimation problem.

3.4.3. **Standard Error.** To estimate the variance of our estimator, we use a standard error of the form

$$\widehat{se}^2_{CF}(h;\widehat{\mu}_n) = \sum_{i=1}^{n} w_i^2(h)\widehat{\sigma}_i^2(\widehat{\mu}_{n,s(i)}),$$

where $\widehat{\sigma}_i^2(\widehat{\mu}_{n,s(i)})$ is an estimator of the variance $\sigma_i^2(\bar{\mu}_n) = \mathbb{V}[M_i(\bar{\mu}_n)|X_i]$. Following Noack and Rothe (2021), we consider a version of the nearest neighbor variance estimator of Abadie et al. (2014).[9] We choose some R, say $R = 5$, which determines the number of neighbors to be used in the variance estimation. Based on the realized running variable, for each unit i, we determine its R nearest neighbors that are on the same side of the cutoff and within the same fold as unit i. Our estimator $\widehat{\sigma}_i^2(\widehat{\mu}_{n,s(i)})$ is proportional to the squared difference between $M_i(\widehat{\mu}_{n,s(i)})$ and its best linear predictor given the running variable based on its R nearest neighbors. We give a formal definition of this estimator in Appendix 3.C.4.

Proposition 3.4.1. *Suppose that Assumptions 3.1–3.5 hold and that for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$, $\sup_{\mu \in \mathcal{T}_n} \mathbb{E}[(M_i(\mu) - \mathbb{E}[M_i(\mu)|X_i])^4|X_i = x]$ is bounded by B. Then*

$$nh\,\widehat{s}^2_{CF}(h;\widehat{\mu}_n) - V(\bar{\mu}_n) = o_P(1).$$

The additional assumption imposed in Proposition 3.4.1 strengthens Assumption 3.5(ii). Existence of conditional fourth moments of the outcome variable is often used for showing consistency of standard errors.

3.5. IMPLEMENTATION DETAILS

In this section, we address point estimation and inference. We also discuss how to incorporate different bandwidths on different sides of the cutoff in the second stage.

3.5.1. **Bandwidth Choice.** One of the key steps to implement our estimation procedure is to choose the bandwidth h for the local linear regression in the second stage. We consider two approaches used in the RD literature.

First, we can select the bandwidth that minimizes the asymptotic mean squared error (AMSE), which is defined as:

$$AMSE_n(h) = B(\bar{\mu}_n)^2 h^4 + \frac{1}{nh} V(\bar{\mu}_n).$$

The optimal bandwidth is then given by $h_{\text{opt}} = \left(V(\bar{\mu}_n)/(4B(\bar{\mu}_n)^2)\right)^{1/5} n^{-1/5}$. It can be estimated following the procedures proposed by Imbens and Kalyanaraman (2012) and

[9] Alternatively, one can use the Eicker-Huber-White (EHW) standard error, but it might be conservative in finite samples; see the discussion by Abadie et al. (2014) in the standard nonparametric regression context.

Calonico et al. (2014). These procedures require estimating $\partial_x^2 \mathbb{E}[M_i(\bar{\mu}_n)|X_i = x]$ to the left and to the right of the cutoff, which can be done using our generated outcome variable under additional smoothness assumptions.

Second, we can adapt the 'bias-aware' approach of Armstrong and Kolesár (2020). They select the bandwidth that minimizes the worst-case mean squared error over a function class formed by imposing a bound on the second derivatives of the considered function. Suppose that $|\partial_x^2 \mathbb{E}[M_i(\bar{\mu}_n)|X_i = x]|$ is bounded by constants B_{M-} and B_{M+} to the left and to the right of the cutoff, respectively, and let $B_M = B_{M-} + B_{M+}$. The leading bias term of our estimator is then bounded in absolute value by $\frac{1}{2}|\bar{\nu}|B_M h^2$. The bandwidth minimizing the corresponding worst-case asymptotic mean squared error is given by $h_{\text{opt}}^{BA} = \left(V(\bar{\mu}_n)/(\bar{\nu}B_M)^2\right)^{1/5} n^{-1/5}$. Implementation of this bandwidth selector requires choosing the smoothness constants B_{M-} and B_{M+}. See Armstrong and Kolesár (2020) and Noack and Rothe (2021) for discussions of the choice of smoothness constants. We note that under the smoothness assumption in Corollary 3.1, it suffices if the smoothness constants B_{M-} and B_{M+} are chosen so as to bound $|\partial_x^2 \mathbb{E}[Y_i|X_i = x]|$ to the left and to the right of the cutoff, respectively.

3.5.2. **Confidence Intervals.** We construct confidence intervals (CIs) for τ based on the asymptotic distribution obtained in part (ii) of Theorem 3.1. The variance $V_n(\bar{\mu}_n)$ can be estimated using the standard error $\widehat{s}_{CF}(h; \widehat{\mu}_n)$ proposed in Section 3.4.3. To account for the asymptotic bias, we can adapt standard methods available in the nonparametric literature.

First, we consider undersmoothing (US), which relies on selecting a bandwidth of order smaller than $n^{-1/5}$. In this case, the bias is asymptotically negligible, and an asymptotically valid $1 - \alpha$ CI can be formed as:

$$CI_\alpha^{US} = [\widehat{\tau}_{CF}(h; \widehat{\mu}_n) \pm z_{1-\alpha/2} \cdot \widehat{s}_{CF}(h; \widehat{\mu}_n)], \tag{3.5.1}$$

where z_u is the u-quantile of the standard normal distribution. The two further approaches allow for the optimal bandwidths discussed in the previous section, which are of order $n^{-1/5}$.

Second, the robust bias corrections (RBC) proposed by Calonico et al. (2014) can be easily adapted to our setting. In this approach, we subtract an estimate of the leading bias term and account for the additional variation in the bias-corrected estimator when forming a CI. These additional steps can be conducted using our generated outcome variable $M_i(\widehat{\mu}_{n,s(i)})$ instead of the original outcome Y_i under further regularity conditions. Let $\widehat{\tau}_{CF}^{RBC}(h; \widehat{\mu}_n)$ be the bias-corrected estimator and $\widehat{s}_{CF}^{RBC}(h; \widehat{\mu}_n)$ the corresponding standard error. The proposed CI is given by:

$$CI_\alpha^{RBC} = [\widehat{\tau}_{CF}^{RBC}(h;\widehat{\mu}_n) \pm z_{1-\alpha/2} \cdot \widehat{s}_{CF}^{RBC}(h;\widehat{\mu}_n)], \tag{3.5.2}$$

The third approach adapts the 'bias-aware' approach of Armstrong and Kolesár (2020). Under the assumption of bounded second derivatives discussed in the previous section, it follows that an asymptotically valid $1-\alpha$ confidence interval can be formed as:

$$CI_\alpha^{BA} = [\widehat{\tau}_{CF}(h;\widehat{\mu}_n) \pm \mathrm{cv}_{1-\alpha}(\widehat{r}(h)) \cdot \widehat{s}_{CF}(h;\widehat{\mu}_n)],$$

where $\widehat{r}(h) = \frac{1}{2}|\bar{\nu}|B_M h^2/\widehat{s}_{CF}(h)$ and $\mathrm{cv}_{1-\alpha}(t)$ is the $1-\alpha$ quantile of the folded normal distribution $|\mathcal{N}(t,1)|$. One can also account for the maximal bias of the infeasible estimator $\widehat{\tau}(h;\bar{\mu}_n)$ conditional on $\mathcal{X}_n$ instead of bounding only the leading bias term.

3.5.3. **Different Bandwidths.** Our estimation procedure introduced in Section 3.3.3 employs a single bandwidth in the second-stage local linear regression. In some empirical settings, however, it might be desirable to run two separate local linear regressions using different bandwidths on different sides of the cutoff. The reason for that might be, for example, that the curvature of the conditional expectation of the outcome variable or its conditional variance are different to the left and to the right of the cutoff. Another reason for choosing different bandwidths might be that the density of the running variable is very steep at the cutoff, so that the numbers of observations with the running variable in $(-h_{\text{opt}}, 0)$ and $(0, h_{\text{opt}})$ are substantially different.

It is straightforward to account for different bandwidths in the asymptotic distribution of our estimator, but the adjustment term based on μ_n is no longer optimal in such a case. We therefore generalize the optimality result in part (iii) of Theorem 3.1. When bandwidths h_- and h_+ are used to the left and to the right of the cutoff, respectively, then the variance of our estimator in large samples is approximately equal to:

$$\widetilde{V}(\bar{\mu}_n) = \omega_+ \mathbb{V}[M_i(\bar{\mu}_n)|X_i = 0^+] + \omega_- \mathbb{V}[M_i(\bar{\mu}_n)|X_i = 0^-],$$

where $\omega_- = \sum_{i=1}^n w_{i,-}(h_-)^2$ and $\omega_+ = \sum_{i=1}^n w_{i,+}(h_+)^2$ and the weights $w_{i,-}$ and $w_{i,+}$ correspond to the local linear estimators run using the data to the left and to the right of the cutoff, respectively. The explicit expressions are given in Appendix 3.C.1.[10] The weights ω_- and ω_+ capture the inverse of the effective sample size to the left and to the right of the cutoff, respectively.

[10] Apart from allowing for different bandwidths, $\widetilde{V}(\mu)$ differs from $V(\mu)$ in Theorem 3.1 in that it does not rely on kernel-weighted sums of X_i to converge to their limits. As such, $\widetilde{\mathcal{V}}(\bar{\mu}_n)$ may capture the finite-sample variance of our estimator more accurately. Still, this expression remains valid only asymptotic as we use $\mathbb{V}[M_i(\mu)|X_i = x]$ evaluated to the left and to the right of the cutoff, rather than for each X_i separately.

We show in Appendix 3.D that $\widetilde{V}(\mu)$ is minimized by the function

$$\mu_n^*(z) = \frac{\omega_-}{\omega_- + \omega_+}\mu_n^-(z) + \frac{\omega_+}{\omega_- + \omega_+}\mu_n^+(z) \tag{3.5.3}$$

in the sense that $\widetilde{V}(\mu_n^*) \leq \widetilde{V}(\mu)$ for all $\mu \in \mathcal{M}_n$. This result is consistent with Theorem 3.1 because $\omega_-/(\omega_- + \omega_+) \to 1/2$ under our assumptions if $h_- = h_+$.

We remark that for any given bandwidths the above weighting scheme puts more weight to the side of the cutoff where the effective sample size is smaller. The reason for that is apparent in the proof given in Appendix 3.D, where we show that minimization of $\widetilde{V}(\mu)$ is equivalent to minimization of $\widetilde{\mathcal{V}}(\mu) = \omega_+\mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0] + \omega_-\mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0]$. If, for example, ω_+ is large compared to ω_-, then choosing μ so as to make $\mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0]$ small is relatively more important than decreasing the value of $\mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0]$. Accordingly, μ_n^+ receives a higher weight in (3.5.3) in such a case.

3.6. EXAMPLES OF COVARIATE ADJUSTMENTS

In this section, we give primitive conditions for our high-level Assumptions 3.3 and 3.4, which concern the properties of the first-stage estimator. We consider in turn: linear, non-linear parametric, local linear, and generic machine learning adjustments. In Sections 3.6.1–3.6.3, where we consider methods suitable for settings with a low-dimensional covariate, we assume that the distribution of W_i does not change with n.

3.6.1. **Linear Adjustments.**

We define a linear estimator using observations close to the cutoff:

$$\widehat{\beta}_s = \arg\min_{\beta} \sum_{s \in I_s^c} K(X_i/h)(Y_i - \beta^\top(Z_i^\top T_i, Z_i^\top(1-T_i), X_i, T_iX_i, T_i, 1)^\top)^2. \tag{3.6.1}$$

Let $\widehat{\beta}_{s,Z}^+$ denote the first d components of $\widehat{\beta}_s$ and let $\widehat{\beta}_{s,Z}^-$ be the next d components of $\widehat{\beta}_s$. We define $\widehat{\mu}_{n,s}(z) = z^\top\widehat{\beta}_{s,Z}$, where $\widehat{\beta}_{s,Z} = \frac{1}{2}(\widehat{\beta}_{s,Z}^+ + \widehat{\beta}_{s,Z}^-)$.[11] Let $\bar{Z}_i = (1, Z_i^\top, X_i/h_1)^\top$. Assumptions 3.3 and 3.4 hold if we can ensure that the estimated slope coefficients concentrate around some deterministic sequence and the conditional expectation $\mathbb{E}[Z_i|X_i = x]$ is sufficiently smooth.

Assumption 3.6. *(i) Each component of* $\mathbb{E}[Z_i|X_i = x]$ *is twice differentiable on* $\mathcal{X} \setminus \{0\}$ *with L-Lipschitz continuous second derivative for some constant L; (ii) The limit as* $n \to \infty$ *of* $\mathbb{E}[K_{h_1}(X_i)\bar{Z}_i\bar{Z}_i^\top]$ *is non-singular; (iii)* $\mathbb{E}[Z_i^\top Z_i|X_i = x]$ *is bounded uniformly over* $x \in \mathcal{X}$*.*

[11] As discussed in Section 3.3.2, it suffices to estimate μ_n up to a constant.

Proposition 3.6.1. *Suppose that Assumption 3.6 holds and either (i) $h_1 \to 0$ and $nh_1 \to \infty$ or (ii) $h_1 \to c > 0$. Then Assumptions 3.3 and 3.4 are satisfied.*

This type of adjustments bears resemblance to the procedure of Calonico et al. (2019). Specifically, they obtain their estimator from a regression as in (3.6.1) but with two main differences. First, they using the whole sample. With these simple adjustments, cross-fitting is not necessary in our procedure, but it does not have any adverse effects. Second, they impose the restriction that $\widehat{\beta}_Z^+ = \widehat{\beta}_Z^-$. Doing so, implies by standard OLS algebra that $\widehat{\mu}_n(z)$ puts more weight to the side of the cutoff with the larger effective sample size. As can be seen in Section 3.5.3, this type of weighting is not optimal.[12]

3.6.2. **Non-linear Parametric Adjustments.** Suppose that the researcher uses some parametric specification $m_\beta(z) = \frac{1}{2}(m_\beta^-(z) + m_\beta^+(z))$ for the function μ_n, which can be based, e.g., on the logit or probit model. This specification might be correct or incorrect. The function m_β is known up to a finite-dimensional parameter $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$. We assume that there is an estimator $\widehat{\beta}$ converging to some nonrandom probability limit $\bar{\beta}$. Classic conditions for consistency in M-estimation problems are given, e.g., by Newey and McFadden (1994).

Assumption 3.7. *(i) For some $\bar{\beta}$ and $r_n \to 0$, $||\widehat{\beta} - \bar{\beta}||_\infty = O_p(r_n)$; (ii) For all $\beta_1, \beta_2 \in \mathcal{B}$, $z \in \mathcal{Z}$, and some constant G, $|m_{\beta_1}(z) - m_{\beta_2}(z)| \leq G||\beta_1 - \beta_2||_\infty$.*

Assumption 3.7 guarantees that the first-stages estimator converges in the supremum norm to some limiting function. With this mode of convergence, Assumption 3.3 follows trivially, and Assumption 3.4 also holds under regularity conditions on the conditional distribution of the additional covariates given the running variable. For concreteness, we assume that Z_i is continuously distributed given X_i, but analogous results can be derived if the additional covariates have a discrete distribution.

Proposition 3.6.2. *Suppose that Assumptions 3.1, 3.2, and 3.7 hold. Moreover, Z_i has bounded support and $\partial_x^j f_{Z|X}(z|x)$ is L-Lipschitz continuous in x for all z and $j \in \{0, 1\}$. Then Assumptions 3.3 and 3.4 are satisfied.*

3.6.3. **Nonparametric Adjustments.** We consider covariate adjustments based on classic nonparametric methods, which are suitable if the number of additional covariates is not too large. To fix ideas, we focus on local linear estimation (Fan and Gijbels, 1996), but similar results can be obtained for example for sieve estimation (Newey, 1997).

For $z \in \mathbb{R}^d$, we define the multivariate kernel as the product of univariate kernels, $\mathcal{K}_h(z) = \prod_{i=1}^d K_h(z_i)$, where $K_h(v) = \frac{1}{h}K(v/h)$.[13] We define estimators of $\mu_n^+(z)$ and $\mu_n^-(z)$

[12] A similar point is made by Negi and Wooldridge (2020) in the context of randomized experiments.

[13] The kernel chosen for the local linear first-stage estimator can be also different from the kernel used in the second stage.

using data in the complement of the sth fold as:

$$\begin{aligned}\widehat{\mu}^{+}_{n,s}(z) &= e_1^{\top} \arg\min_{\beta} \sum_{i \in I_s^c} T_i K_{h_X}(X_i) \mathcal{K}_{h_Z}(Z_i - z)(Y_i - \beta^{\top}(1, (Z_i - z)^{\top}, X_i))^2, \\ \widehat{\mu}^{-}_{n,s}(z) &= e_1^{\top} \arg\min_{\beta} \sum_{i \in I_s^c} (1 - T_i) K_{h_X}(X_i) \mathcal{K}_{h_Z}(Z_i - z)(Y_i - \beta^{\top}(1, (Z_i - z)^{\top}, X_i))^2.\end{aligned}$$

In Assumption 3.8 in Appendix 3.C.6, we impose standard assumptions on the data generating process for the local linear estimator.

Proposition 3.6.3. *Suppose that Assumptions 3.1, 3.2, and 3.8 hold. Further, assume that $h_X \to 0$, $h_Z \to 0$, $\log(n)/(nh_X h_Z^d) \to 0$, and $\partial_x^2 f_{Z|X}(z|x)$ is L-Lipschitz continuous in x for all z. Then Assumptions 3.3 and 3.4 are satisfied.*

Under Assumption 3.8 and the bandwidth conditions of Proposition 3.6.3, Masry (1996) shows that the local linear estimator is uniformly consistent. Using this result, Assumption 3.1 follows trivially. Assumption 3.4 also follows under the additional smoothness conditions; see the discussion in Appendix 3.A.2.

We emphasize that the bandwidth conditions are very mild, and they can be chosen, e.g., via cross-validation under further, standard regularity conditions. With a moderate number of covariates, it is optimal to choose a relatively large bandwidth, but this is allowed as long as they converge to zero. In general, with our method is advisable to oversmooth, rather than undersmooth when choosing the bandwidths in order to guarantee that the estimator is not too volatile. Oversmoothing comes at the cost of a possible increase in the variance of the final estimator, but it renders the normal approximation of the asymptotic distribution more reliable in finite samples.

3.6.4. **Adjustments Based on Machine Learning Methods.** We outline a general approach to ensuring that our high-level assumptions hold for many machine learning methods. Results about estimation of conditional expectations using machine learning methods typically concern convergence in mean square. We can make use of these results by estimating the functions μ_n^- and μ_n^+ based on narrow, fixed 'slices' of the data to the left and to the right of the cutoff, respectively.[14] Specifically, for any fixed h_1, we can readily obtain the result that the selected estimator belongs to some realization set $\mathcal{T}_n$ with probability approaching one, and

$$\sup_{\mu \in \mathcal{T}_n} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i \in \mathcal{X}_{h_1}\right] = o(1), \tag{3.6.2}$$

[14] Restricting the sample corresponds to weighting the observations based on a uniform kernel. Our reasoning applies also to any other kernel weighting scheme, e.g. using the triangular kernel.

where $\bar{\mu}_n(z) = \frac{1}{2}(\bar{\mu}_n^+(z) + \bar{\mu}_n^-(z))$ with $\bar{\mu}_n^+(z) = \mathbb{E}[Y_i|Z_i = z, X_i \in (0, h_1)]$ and $\bar{\mu}_n^-(z) = \mathbb{E}[Y_i|Z_i = z, X_i \in (-h_1, 0)]$. If the conditional distribution of the additional covariates given the running variable is sufficiently smooth on the interval $(-h_1, h_1)$, then the above property implies that Assumptions 3.3 and 3.4 hold; see Appendix 3.A for more details.

Primitive conditions for (3.6.2) are available for a variety of machine learning techniques, e.g. post-lasso (Belloni et al., 2012), random forests (Breiman, 2001; Wager and Athey, 2018), and deep neural networks (Farrell et al., 2021). Hence, we can flexibly choose a method that is best-suited for a given dataset under the assumptions imposed.

With fixed h_1, $\bar{\mu}_n$ might be different from μ_n. Our theory allows for that, but this procedure in general does not achieve the optimal variance $V(\mu_n)$. In the previous section, we show that for the local linear estimator, the optimal variance can be achieved by choosing h_1 that converges to zero. It would be interesting to formally study the setting with $h_1 \to 0$ for other methods. We leave this for future research.

3.7. SIMULATIONS

We compare the finite sample performance of our proposed estimator for different first-stage estimation methods in a Monte Carlo study.

3.7.1. **Setup.** We consider four models, which differ in the number of covariates entering the outcome equation, which we denote by $L \in \{0, 4, 10, 25\}$. The running variable X_i follows the uniform distribution over $[-1, 1]$. There are four independent, baseline covariates, denoted by Z_i^b, which are distributed uniformly over $[-1 + x^2, 1 + x^2]^4$ conditional on $X_i = x$. We generate further covariates based on the baseline covariates using Hermit polynomials. Let $b_l(Z_i^b)$ denote the l-th covariate. The outcome is generated according to the following model:

$$Y_i = D_i + \mu_L(X_i, Z_i) + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 0.25)$ and

$$\mu_L(X_i, Z_i) = \text{sign}(X_i) \cdot (X_i + X_i^2 - 2\,(X_i - 0.1)_+^2) + \bar{\iota}_L(\rho) \sum_{l=1}^{L} b_l(Z_i^b).$$

For positive L and ρ, we chose the coefficient $\bar{\iota}_L(\rho)$ so that $\mathbb{V}[\mu_L(0, Z_i)|X_i = 0] = \rho^2 \mathbb{V}[\varepsilon_i]$. In this definition, ρ represents the signal to noise ratio at the cutoff given the treatment status. It determines the scope for improvements from using covariates, but it does not affect the relative performance of different covariate adjustments. For concreteness, in the main text, we consider $\rho = 3$. We report simulation results for further values of ρ in Appendix 3.E. The results are based on $5,000$ simulation draws. The sample size is $2,000$

for the main results.

We consider in total seven implementations of the first-stage estimator: (i) the standard RD estimator with no covariate adjustments; (ii) the infeasible, optimal RD estimator with covariate adjustments based on the true conditional expectation function; (iii) the infeasible RD estimator with adjustments based on the best linear prediction on the population level of the true conditional expectation function given the four baseline covariates.[15] We consider four feasible adjustment functions based on:[16] (iv) a linear regression given the four baseline covariates; (v) a local linear regression given the four baseline covariates; (vi) a post-lasso regression given 200 covariates; and (vii) a random forest with the four baseline covariates.

To keep the exposition simple, in the main text, we consider only the bias-aware approach for the implementation of the second stage. Our procedure is based on the true bound on the second derivative of the conditional expectation of the outcome variable. The bandwidth is chosen to be optimal in terms of the estimated worst-case mean squared error. The main qualitative conclusions of our simulation study hold also for robust bias corrections and undersmoothing. We present these results in Appendix 3.E. There we also compare our estimators to the linear covariates adjustment method proposed by Calonico et al. (2019).[17]

3.7.2. **Simulations Results.** Table 3.1 reports estimation and inference results for different types of adjustments. The CIs for all estimators have simulated coverage rates close to their nominal ones.[18] First, we compare the standard RD estimator and the infeasible estimators. In Model 1, these estimators are numerically equal. In Models 2–4, where the covariates have some explanatory power for the outcome, the infeasible estimators have a substantially lower standard deviation than the standard estimator has. If the linear model is misspecified, the standard deviation of the optimal infeasible estimator is much smaller than that of the infeasible estimator with linear adjustments. We now turn to the feasible covariate-adjusted RD estimators. As predicted by Theorem 3.1, their mean standard deviations are close to those of their respective infeasible estimator, with only a slight increase due to the first-stage estimation.

[15] We obtain the population projection coefficients based on 10^7 draws with $X_i = 0$ and $\varepsilon_i = 0$. We fix this estimate through all simulations for each data generating process.

[16] In the first-stage, the observations are weighted using kernel weights with the bandwidth selected for the standard RD estimator.

[17] All computations are carried out with the statistical software `R`. The Hermit-polynomials are computed using the package `calculus`. To implement the first-stage estimators, we use the following packages: `np` for local polynomial regressions; `glmnet` for lasso regressions; `grf` for random forests, where predictions are based on 200 trees. In the second stage, a triangular kernel is used and EHW standard errors are computed. The bias-aware approach is based on the package `RDHonest`, and the other two approaches are implemented using the package `rdrobust`.

[18] In the considered models, the maximal bias is not achieved, so that the bias-aware CIs are conservative.

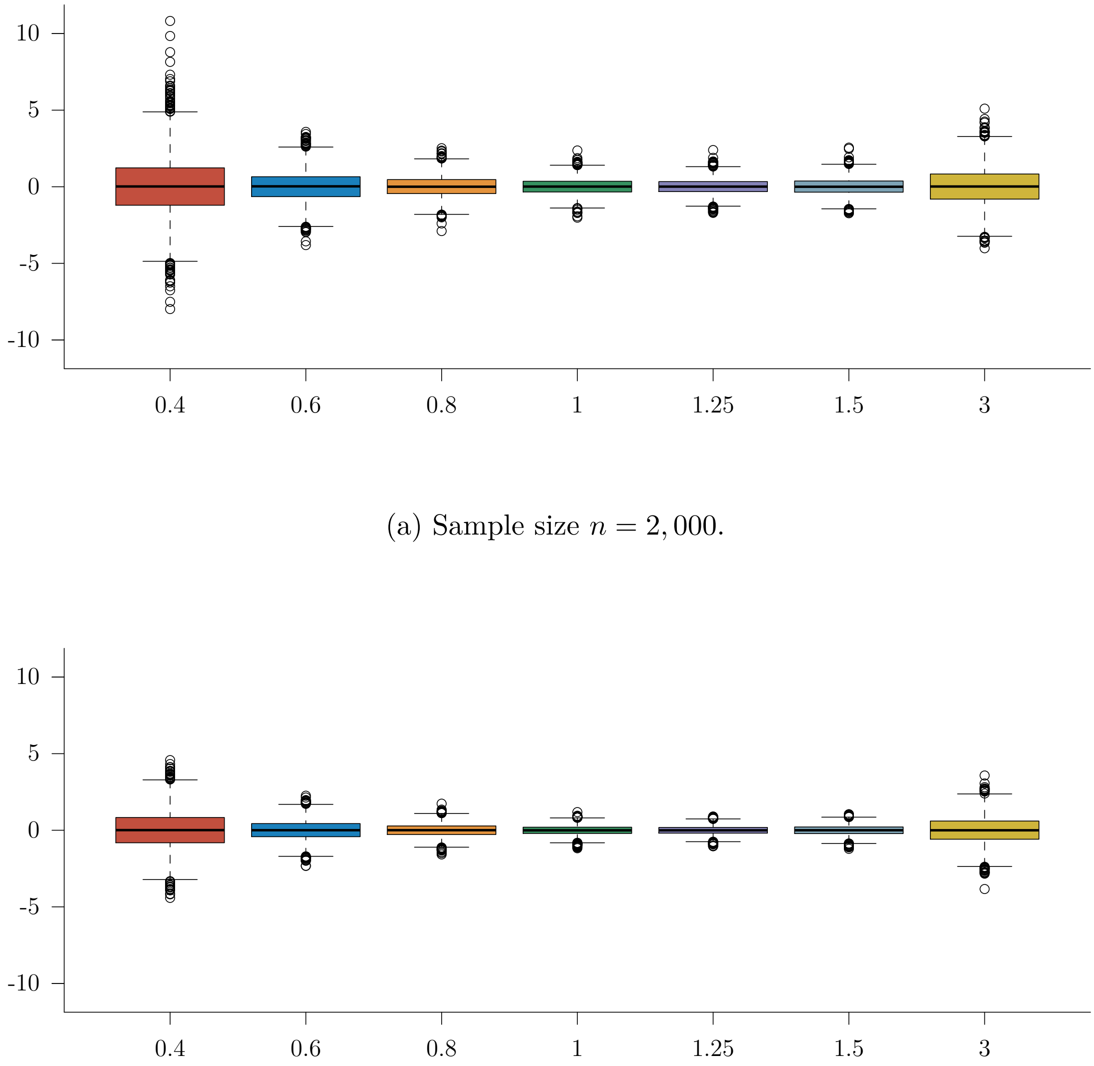
Table 3.1: Simulation results.

	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
	Model 1: L=0					**Model 2: L=4**				
Standard	97.0	-1.4	7.4	32.4	43.2	96.1	-7.1	18.6	81.8	68.8
Optimal Inf	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2
Linear Inf	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2
Linear	97.0	-1.4	7.4	32.7	43.3	96.7	-1.5	7.5	32.6	43.3
Local Linear	97.0	-1.4	7.4	32.7	43.3	96.8	-1.4	7.5	32.7	43.3
Lasso	96.7	-1.4	7.6	33.1	43.6	96.6	-2.1	8.8	38.3	46.6
Forest	96.8	-1.5	7.6	33.1	43.6	96.7	-2.1	8.7	37.9	46.5
	Model 3: L=10					**Model 4: L=25**				
Standard	96.4	-9.5	19.1	87.6	79.3	95.9	-6.3	18.5	81.0	68.5
Optimal Inf	96.5	-1.3	7.6	32.5	43.2	96.9	-1.3	7.4	32.4	43.2
Linear Inf	96.7	-4.8	12.7	56.2	61.8	96.8	-4.3	10.3	47.2	59.0
Linear	95.9	-4.0	13.7	59.1	59.7	96.5	-4.3	10.8	49.2	58.8
Local Linear	96.3	-1.6	8.3	35.6	45.2	96.8	-1.6	8.2	35.9	45.6
Lasso	96.2	-2.0	9.2	39.1	46.7	96.8	-1.4	7.7	34.0	44.3
Forest	96.6	-1.9	8.5	37.2	46.9	97.1	-2.2	9.3	41.3	49.0

Notes: Results based on 5,000 Monte Carlo draws for the bias-aware approach. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h).

In Figures 3.1 and 3.2, we compare the difference between the optimal infeasible RD estimator and two feasible ones: with adjustments based on local linear regression and post-lasso regression for several choices of the tuning parameters. In each simulation draw, we find the MSE-optimal tuning parameters via cross-validation, and then scale it down or up by different factors.[19] We consider two sample sizes, $n = 2,000$ and $n = 10,000$. We normalize the difference by the standard error of the optimal infeasible RD estimator.

[19] To facilitate comparisons of different covariate adjustments, in each simulation draw, we use the bandwidth selected for the standard RD estimator in the second stage across all different methods.

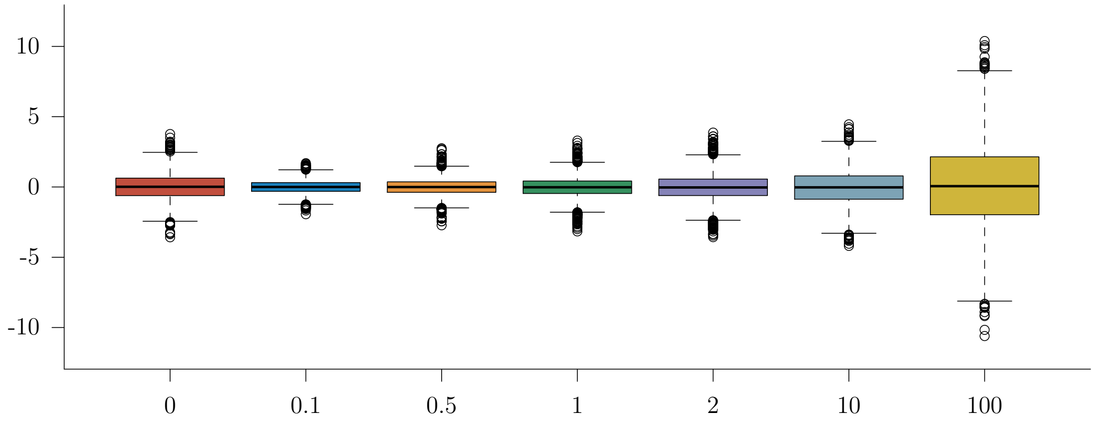


(a) Sample size $n = 2,000$.

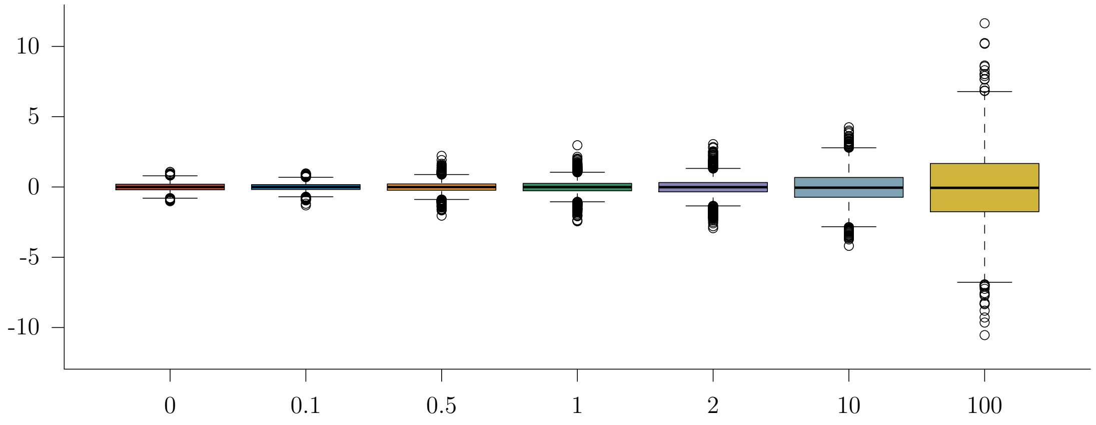
(b) Sample size $n = 10,000$.

Figure 3.1: Normalized difference of RD estimates with local linear adjustments.

Notes: Difference between optimal infeasible and feasible RD estimate normalized by standard deviation of infeasible estimator. We consider various scaling factors for the cross-validated MSE-optimal first-stage bandwidth. Simulations are based on Model 3. Panel (a) shows simulation results for $n = 2,000$, and Panel (b) for $n = 10,000$.



(a) Sample size $n = 2,000$.



(b) Sample size $n = 10,000$.

Figure 3.2: Normalized difference of RD estimates with post-lasso regression adjustments. *Notes:* Difference between optimal infeasible and feasible RD estimate normalized by standard deviation of infeasible estimator. We consider various scaling factors for the cross-validated MSE-optimal first-stage penalty parameter. Simulations are based on Model 3. Panel (a) shows simulation results for $n = 2,000$, and Panel (b) for $n = 10,000$.

In Figure 3.1, we observe that the normalized difference between the estimators is relatively small for a wide range of bandwidths around the optimal one. By comparing panels (a) and (b), we can see that these normalized differences become smaller as the sample size increases, which illustrates the asymptotic equivalence result in part (i) of Theorem 3.1. For a given sample size, the average absolute value of the normalized differences is U-shaped as a function of the bandwidth. If the bandwidth chosen in the first stage is too small, then the local linear estimator is very unstable. In this case, the property in Assumption 3.3 is not a good description of its finite-sample behavior, and the equivalence result in Theorem 3.1 fails. If the bandwidth is chosen to be too large, the local linear estimator has a relatively small variance, but it might be heavily biased, and it is effectively very similar to the linear estimator. In this case, the equivalence to an infeasible estimator holds with a different limiting sequence $(\bar{\mu}_n)_{n\in\mathbb{N}}$. We expect the estimator to be less efficient, but we emphasize that our inference procedure remain valid in this case.

Figure 3.2 shows a very similar pattern as Figure 3.1. If the penalty parameter in the lasso regression is chosen to be too small, effectively all covariates are classified as relevant, and the first-stage estimator has a high variance. In contrast, if the penalty parameter is chosen to be too large, very few covariates are classified as relevant. In this case, the covariate-adjusted RD estimator behaves similarly to the standard RD estimator.

3.8. CONCLUSIONS

Linear covariate adjustments are commonly used in RD designs to improve efficiency of the standard RD estimator. In this paper, we propose a class of RD estimators that allow for nonparametric covariate adjustments, which can reduce the variance of the RD estimator even further. We allow for a wide range of covariate adjustments under mild conditions. Despite using possibly highly-complex covariate adjustments, inference on the RD parameter can be conducted using standard methods available in the literature. We illustrate our results in a simulation study.

Appendix

3.A. FURTHER SUFFICIENT CONDITIONS FOR MAIN ASSUMPTIONS

In this section, we discuss sufficient conditions for our high-level Assumptions 3.3 and 3.4.

3.A.1. **Sufficient Conditions for Assumption 3.3.** We outline a generic way of ensuring that Assumption 3.3 holds, which can be employed for a wide range of estimators. For concreteness, we assume that the additional covariates are continuously distributed conditional on the running variable, but similar results can be derived for discrete distributions or intermediate cases.

Many results in the machine learning literature concern convergence in mean square, which means that we can obtain the following property:

$$\sup_{\mu\in\mathcal{T}_n}\mathbb{E}\left[(\mu(Z_i)-\bar{\mu}_n(Z_i))^2|X_i\in\mathcal{X}_h\right]=o(1). \tag{3.A.1}$$

We can infer our assumption from the above condition if the conditional distribution of the additional covariates does not change abruptly around the cutoff. Specifically, suppose that

$$\sup_{x\in\mathcal{X}_h}\sup_{z\in\mathcal{Z}_h}\frac{f_{Z|X}(z|x)}{f_{Z|X\in\mathcal{X}_h}(z)}<B, \tag{3.A.2}$$

for some constant B and h small enough. If the conditions in (3.A.1) and (3.A.2) hold, then Assumption 3.3 is satisfied because::

$$\begin{aligned}
&\sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[(\mu(Z_i)-\bar{\mu}_n(Z_i))^2|X_i=x\right]\\
&\qquad=\sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\int_{\mathcal{Z}_h}(\mu(Z_i)-\bar{\mu}_n(Z_i))^2 f_{Z|X\in\mathcal{X}_h}(z)\frac{f_{Z|X}(z|x)}{f_{Z|X\in\mathcal{X}_h}(z)}dz\\
&\qquad\leq B\sup_{\mu\in\mathcal{T}_n}\mathbb{E}\left[(\mu(Z_i)-\bar{\mu}_n(Z_i))^2|X_i\in\mathcal{X}_h\right]=o(1).
\end{aligned}$$

3.A.2. **Sufficient Conditions for Assumption 3.4.** We show that Assumption 3.4 can be inferred from the convergence imposed in Assumption 3.3 under mild additional smoothness conditions on the conditional distribution of the additional covariates given the running variable. This can be most intuitively seen when the support $\mathcal{Z}$ is discrete. In the continuous case some additional integrability conditions are needed.

Discrete Additional Covariates. Suppose that the support of the additional covariates, $\mathcal{Z}$, is finite. In this case, Assumption 3.3 implies that $\sup_{\mu\in\mathcal{T}_n}\sup_{z\in\mathcal{Z}_h}|\mu(z)-\bar{\mu}(z)|=o(1)$. Then for $j\in\{1,2\}$,

$$\partial_x^j\mathbb{E}[\mu(Z_i)-\bar{\mu}_n(Z_i)|X_i=x]=\sum_{z\in\mathcal{Z}}(\mu(z)-\bar{\mu}(z))\partial_x^j\mathbb{P}[Z_i=z|X_i=x].$$

Given Assumption 3.3, Assumption 3.4 holds if $\sup_{x\in\mathcal{X}_h\setminus\{0\}}\sup_{z\in\mathcal{Z}_h}\partial_x^1\mathbb{P}[Z_i=z|X_i=x]=O(1/h)$ and $\sup_{x\in\mathcal{X}_h\setminus\{0\}}\sup_{z\in\mathcal{Z}_h}\partial_x^2\mathbb{P}[Z_i=z|X_i=x]=O(1)$.

Continuous Additional Covariates. Suppose that the additional covariates are continuously distributed given the running variable, and that the conditional density $f_{Z|X}(z|x)$ is twice differentiable with respect to x on $\mathcal{X}\setminus\{0\}$ for all z. Further, assume that for $j\in\{0,1\}$, there exists a function $H_j(z)$ integrable over $\mathcal{Z}$ such that for all x_1,$x_2\in(0,h)$,

$$\left|\partial_x^j f_{Z|X}(z|x_1)-\partial_x^j f_{Z|X}(z|x_2)\right|+\left|\partial_x^j f_{Z|X}(z|-x_1)-\partial_x^j f_{Z|X}(z|-x_2)\right|\leq H_j(z)|x_1-x_2|.$$

In this setting, Assumption 3.4 holds if in addition to Assumption 3.3 for $j\in\{0,1\}$ either

(i) $\sup\limits_{\mu\in\mathcal{T}_n}\sup\limits_{z\in\mathcal{Z}_h}|\mu(z)-\bar{\mu}_n(z)|\to 0$, or

(ii) $\sup\limits_{x\in\mathcal{X}_h\setminus\{0\}}\mathbb{E}\left[\left(H_j(Z_i)/f_{Z|X}(Z_i|x)\right)^2\Big|X_i=x\right]<\infty$.

The first condition requires that the first-stage estimator converges in the supremum norm. This condition is satisfied for classic nonparametric estimators such as kernel and sieve estimators, see, e.g., Masry (1996); Newey (1997).

The second condition ensures that Assumption 3.4 holds in combination with L_2-convergence assumed in Assumption 3.3. The additional integrability condition holds for example if the conditional density $f_{Z|X}(z|x)$ is bounded away from zero and $\partial_x^j f_{Z|X}(z|x)$ is bounded for $j\in\{1,2\}$ uniformly in x and z.

3.B. RELATION TO THE LITERATURE

In this section, we compare our asymptotic results with those of Frölich and Huber (2019) and draw an analogy between our approach and double-robust estimation of the average treatment effect in randomized experiments. We also discuss the relation to estimation based on Neyman-orthogonal moments.

3.B.1. **Comparison with Frölich and Huber (2019).**

Our procedure with the local linear estimator in the first stage is related to that proposed by Frölich and Huber (2019).

Under our assumptions, for sharp designs with the same kernels of order $\lambda = 2$ used in both stages, their bias expression simplifies to:

$$\begin{aligned} bias^{FH} = & \frac{\bar{\nu}}{2} \int (\mu_n^+(z) - \mu_n^-(z) - \tau) \frac{\partial_x^2 f(x,z)}{f_X(0)} dz h^2 \\ & + \frac{\bar{\nu}}{2} \int (\partial_x^2 \mu_n(x,z)|_{x=0^+} - \partial_x^2 \mu_n(x,z)|_{x=0^-}) f_{Z|X}(z|0) dz h_x^2 \\ & + \frac{\nu_2}{2} \sum_{l=1}^{L} \int (\partial_{z_l}^2 \mu_n^+(z) - \partial_{z_l}^2 \mu_n^-(z)) f_{Z|X}(z|0) dz h_z^2, \end{aligned}$$

where $\mu_n(x, z) = \mathbb{E}[Y_i|X_i = x, Z_i = z]$, $\bar{\nu}$ is the "boundary bias kernel constant" defined before Theorem 3.1, and $\nu_2 = \int v^2 k(v) dv$. This expression has a more complicated than the bias in Theorem 3.1, and it does not simplify further under the additional smoothness assumption in Corollary 3.1.

The asymptotic variance equals the variance of our proposed estimator when the first-stage estimator is consistent, $\mathcal{V}^{FH} = V(\mu_n)$. The procedure of Frölich and Huber (2019), however, allows for at most three continuous additional covariates if a second-order kernel is used in the first-stage local linear regression.

3.B.2. **Analogy with ATE estimation.** RD designs are very similar in nature to randomized controlled trials. Conditional on the running variable being close to the cutoff, if the distribution of the covariates evolves continuously through the cutoff, the probability of observing a unit with any given value of the additional covariate is approximately the same to the left and to the right of the cutoff. Hence, the treatment is as if randomly assigned and the propensity score is constant.

In an experiment where the treatment probability is constant across covariates, the augmented inverse probability weighted estimator of the average treatment effect is given by:

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{m}_1(Z_i) - \widehat{m}_0(Z_i) + \frac{T_i(Y_i - \widehat{m}_1(Z_i))}{\widehat{p}} - \frac{(1 - T_i)(Y_i - \widehat{m}_0(Z_i))}{1 - \widehat{p}} \right), \tag{3.B.1}$$

where, $\widehat{m}_t(z)$ is an estimator of $\mathbb{E}[Y_i|Z_i = z, T_i = t]$ for $t \in \{0, 1\}$, and $\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} T_i$ is the proportion of treated units.

This estimator can be also represented as the difference in means in the treatment and control group of a modified outcome variable:

$$\widehat{\tau} = \frac{\sum_{i=1}^{n} T_i(Y_i - \widehat{m}(Z_i; \widehat{p}))}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1 - T_i)(Y_i - \widehat{m}(Z_i; \widehat{p}))}{\sum_{i=1}^{n} (1 - T_i)}, \tag{3.B.2}$$

where $\widehat{m}(z; \widehat{p}) = (1 - \widehat{p})\widehat{m}_1(z) + \widehat{p}\,\widehat{m}_0(z)$. Our proposed estimator is analogous to the expression in (3.B.2) in the sense that it is the difference between estimates from the treatment and control group, except that we replace the estimated propensity score $\widehat{p}$ with the known one, which equals one half.

3.B.3. **Insensitivity to the First Stage.** In two-stage estimation procedures, the first stage generally affects the properties of the final estimator. This complication, however, can be avoided using estimators based on so-called Neyman-orthogonal moments (Neyman, 1959, 1979), whose derivative with respect to the nuisance parameter estimated in the first stage is zero. This method has been recently used in the semiparametric literature in settings where a, possibly high-dimensional, nuisance parameter is estimated using machine learning methods; see, e.g., Belloni et al. (2017); Chernozhukov et al. (2018). In our context, Neyman-orthogonality means that

$$\partial^1_\mu \left(\mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-]\right)\Big|_{\mu=\mu_n} = 0, \tag{3.B.3}$$

where ∂^k_μ denotes the k-th functional derivative in all possible directions.

Our setting is related to estimation problems with Neyman-orthogonal moments but it differs in two main aspects. The property in Equation (3.4.1) is much stronger than (3.B.3) because functional derivatives of all orders evaluated at any function $\mu \in \mathcal{M}_n$ vanish. However, this property holds only conditional on the running variable been at the cutoff, whereas any estimation procedure has to rely on the data in some neighborhood of the cutoff.

3.C. PROOFS OF MAIN RESULTS

3.C.1. **Additional Notation.** We use the following notation throughout the proofs. For $s \in [S]$, $i \in I_{s(i)}$, and $j \in \{0, 1\}$, we define the local linear weights as

$$\begin{aligned}
w^{(j)}_{i,s}(h) &= w^{(j)}_{i,s,+}(h) - w^{(j)}_{i,s,-}(h),\\
w^{(j)}_{i,s,+}(h) &= e^\top_{j+1} Q^{-1}_{s,+} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_{s,+} = \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i, \widetilde{X}_i^\top \mathbf{1}\{X_i \geq 0\},\\
w^{(j)}_{i,s,-}(h) &= e^\top_{j+1} Q^{-1}_{s,-} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_{s,-} = \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i\widetilde{X}_i^\top \mathbf{1}\{X_i < 0\},
\end{aligned}$$

with $\widetilde{X}_i = (1, X_i)^\top$. We omit the index s if the sum is taken over the whole sample and we omit the superscript (j) if $j = 0$.

Further, for $\mu \in \mathcal{M}_n$, we let

$$\begin{aligned}
T_{s,+}(\mu) &= \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i \mu(Z_i) \mathbf{1}\{X_i \geq 0\}\\
T_{s,-}(\mu) &= \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i \mu(Z_i) \mathbf{1}\{X_i < 0\}.
\end{aligned}$$

Let $m(x;\mu) = \mathbb{E}[\bar{\mu}(Z_i) - \mu(Z_i)|X_i = x]$. We define $\beta_0(\mu) = m(0;\mu)$, $\beta_1^+(\mu) = \partial_x m(x;\mu)|_{x=0^+}$, and $\beta_1^-(\mu) = \partial_x m(x;\mu)|_{x=0^-}$, and further $\beta^+(\mu) = (\beta_0(\mu), \beta_1^+(\mu))$ and $\beta^-(\mu) = (\beta_0(\mu), \beta_1^-(\mu))$. Let $H = \text{diag}(1,h)$ and $\mathbb{I}_2 = \text{diag}(1{,}1)$.

3.C.2. **Proof of Theorem 3.1.** The proof of Theorem 3.1 is preceded by two lemmas.

Lemma 3.C.1. *Suppose that Assumption 3.2 holds. Then for all $s \in [S]$ it holds that:*

(i) For all $j \in \mathbb{N}$,

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^n K(X_i/h)(X_i/h)^j T_i = \bar{\nu}_j f_X(0^+) + o_P(1),\\
&\frac{1}{nh}\sum_{i=1}^n K(X_i/h)(X_i/h)^j (1-T_i) = \bar{\nu}_j f_X(0^-) + o_P(1),\\
&\frac{1}{nh}\sum_{i=1}^n K(X_i/h)(X_i/h)^j T_i = \frac{S}{nh}\sum_{i\in I_s} K(X_i/h)(X_i/h)^j T_i + O_P((nh)^{-1/2}),\\
&\frac{1}{nh}\sum_{i=1}^n K(X_i/h)(X_i/h)^j (1-T_i) = \frac{S}{nh}\sum_{i\in I_s} K(X_i/h)(X_i/h)^j (1-T_i) + O_P((nh)^{-1/2}).
\end{aligned}$$

(ii) For $j \in \{0,1\}$, $h^{2j}\sum_{i\in I_s} w_{i,s}^{(j)}(h)^2 = O_P((nh)^{-1})$ and $h^j \sum_{i\in I_s} |w_{i,s}^{(j)}(h) X_i^2| = O_P(h^2)$.

Proof. Standard kernel calculations. □

Lemma 3.C.2. *Suppose that Assumptions 3.1–3.4 hold. Then*

$$G_{s,\star}^{(j)} \equiv e_{j+1}^\top H(Q_{s,\star}^{-1} T_{s,\star}(\bar{\mu}_n - \widehat{\mu}_{n,s}) - \beta^\star(\bar{\mu}_n - \widehat{\mu}_{n,s})) = o_p(h^2 + (nh)^{-1/2})$$

for all $s \in [S]$, $\star \in \{+,-\}$, and $j \in \{0,1\}$.

Proof. We analyze the expectation and variance of $G_{s,\star}^{(j)}$ conditional on $\mathcal{X}_n$ and $(W_j)_{j\in I_s^c}$.

First, we consider the expectation. It holds with probability approaching one that

$$\begin{aligned}
|\mathbb{E}[G_{s,\star}^{(j)}|\mathcal{X}_n, (W_j)_{j\in I_s^c}]| &= \left|\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\mathbb{E}[\bar{\mu}_n(Z_i) - \widehat{\mu}_{n,s}(Z_i)|X_i, (W_j)_{j\in I_s^c}]\right|\\
&\le \sup_{\mu\in\mathcal{T}_n}\left|\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\mathbb{E}[\bar{\mu}_n(Z_i) - \mu(Z_i)|X_i]\right|
\end{aligned}$$

By Taylor's theorem with the mean-value form of the remainder, it holds that

$$m(X_i;\mu) = m(0;\mu) + \partial_x m(x;\mu)|_{x=0^\star} X_i + \frac{1}{2}\partial_x^2 m(\widetilde{x}_i;\mu) X_i^2,$$

for some $\widetilde{x}_i$ between 0 and X_i. Using standard local linear algebra and the triangle inequality, we obtain that

$$
\begin{aligned}
|\mathbb{E}[G_{s,\star}^{(j)}|\mathcal{X}_n,(W_j)_{j\in I_s^c}]| &\leq \sup_{\mu\in\mathcal{T}_n}\left|\frac{1}{2}\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\partial_x^2 m(\widetilde{x}_i;\mu)X_i^2\right| \\
&\leq \sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h\setminus\{0\}}\frac{1}{2}|\partial_x^2 m(x;\mu)|\sum_{i\in I_s}\left|w_{i,s,\star}^{(j)}(h)X_i^2\right| = o_p(h^2),
\end{aligned}
$$

where we use Lemma 3.C.1 and Assumption 3.4 in the last step.

Second, we consider the conditional variance. It holds with probability approaching one that

$$
\begin{aligned}
\mathbb{V}\left[G_{s,\star}^{(j)}|\mathcal{X}_n,(W_j)_{j\in I_s^c}\right] &= \sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2\mathbb{V}\left[\bar{\mu}_n(Z_i)-\widehat{\mu}_{n,s}(Z_i)|\mathcal{X}_n,(W_j)_{j\in I_s^c}\right] \\
&\leq \sup_{\mu\in\mathcal{T}_n}\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2\mathbb{E}[(\bar{\mu}_n(Z_i)-\mu(Z_i))^2|X_i] \\
&\leq \sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}[(\bar{\mu}_n(Z_i)-\mu(Z_i))^2|X_i=x]\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2 \\
&= o_p((nh)^{-1}).
\end{aligned}
$$

where we use Lemma 3.C.1 and Assumption 3.3 in the last step. The conditional convergence implies the unconditional one (see Chernozhukov et al., 2018, Lemma 6.1), which concludes the proof. □

Proof of Theorem 3.1. We prove the three parts separately.
Part (i) It holds that:

$$
\begin{aligned}
&\widehat{\tau}_{CF}(h;\widehat{\mu}_n)-\widehat{\tau}(h;\bar{\mu}_n) \\
&= e_1^\top\sum_{s=1}^S\left\{Q_+^{-1}T_{s,+}(\bar{\mu}_n-\widehat{\mu}_{n,s})-Q_-^{-1}T_{s,-}(\bar{\mu}_n-\widehat{\mu}_{n,s})\right\} \\
&= e_1^\top\sum_{s=1}^S Q_+^{-1}Q_{s,+}(Q_{s,+}^{-1}T_{s,+}(\bar{\mu}_n-\widehat{\mu}_{n,s})-\beta^+(\bar{\mu}_n-\widehat{\mu}_{n,s}))+e_1^\top\sum_{s=1}^S Q_+^{-1}Q_{s,+}\,\beta^+(\bar{\mu}_n-\widehat{\mu}_{n,s}) \\
&-e_1^\top\sum_{s=1}^S Q_-^{-1}Q_{s,-}(Q_{-,s}^{-1}T_{s,-}(\bar{\mu}_n-\widehat{\mu}_{n,s})-\beta^-(\bar{\mu}_n-\widehat{\mu}_{n,s}))-e_1^\top\sum_{s=1}^S Q_-^{-1}Q_{s,-}\,\beta^-(\bar{\mu}_n-\widehat{\mu}_{n,s}). \\
&\equiv A_1+A_2-A_3-A_4.
\end{aligned}
$$

In the following, we consider each of the four terms separately. First, note that

$$
A_1 = e_1^\top H^{-1}\sum_{s=1}^S HQ_+^{-1}HH^{-1}Q_{s,+}H^{-1}H(Q_{s,+}^{-1}T_{s,+}(\bar{\mu}_n-\widehat{\mu}_{n,s})-\beta^+(\bar{\mu}_n-\widehat{\mu}_{n,s}))
$$

By Lemma 3.C.1, for all $s \in [S]$, it holds that

$$HQ_+^{-1}HH^{-1}Q_{s,+}H^{-1} = \frac{1}{S}\mathbb{I}_2 + O_P((nh)^{-1/2}), \tag{3.C.1}$$

where throughout the proof we assume that the term $O_P((nh)^{-1/2})$ has conformable dimensions. Using Lemma 3.C.2 and noting that $e_1^\top H^{-1} = e_1^\top$, we obtain that $A_1 = o_p(h^2 + (nh)^{-1/2})$.

Second, it holds that

$$A_2 = e_1^\top H^{-1} \sum_{s=1}^{S} HQ_+^{-1}HH^{-1}Q_{s,+}H^{-1}H\,\beta^+(\bar{\mu}_n - \widehat{\mu}_{n,s}).$$

Using equation (3.C.1), we obtain that

$$\begin{aligned} A_2 &= \frac{1}{S}\sum_{s=1}^{S}(e_1^\top + O_p((nh)^{-1/2}))H\,\beta^+(\bar{\mu}_n - \widehat{\mu}_{n,s}) \\ &= \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s})(1 + O_p((nh)^{-1/2})) + h\beta_1^+(\bar{\mu}_n - \widehat{\mu}_{n,s})O_p((nh)^{-1/2}) \\ &= \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s}) + o_p((nh)^{-1/2}), \end{aligned}$$

where we use the fact $\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s}) = o_p(1)$ by Assumption 3.3 and $h\beta_1^+(\bar{\mu}_n - \widehat{\mu}_{n,s}) = o_p(1)$ by Assumption 3.4 for all $s \in [S]$.

Using analogous calculations, we can show that $A_3 = o_P(h^2 + (nh)^{-1/2})$ and $A_4 = \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s}) + o_P((nh)^{-1/2})$, which concludes the proof of part (i).

Part (ii). By the conditional version of Lyapunov CLT, we obtain that

$$\mathrm{se}(h;\bar{\mu}_n)^{-1}(\widehat{\tau}(h;\bar{\mu}_n) - \mathbb{E}[\widehat{\tau}(h;\bar{\mu}_n)|\mathcal{X}_n]) \to \mathcal{N}(0,1).$$

where $\mathrm{se}^2(h;\bar{\mu}_n) = \sum_{i=1}^{n} w_i(h)^2\mathbb{V}[M_i(\bar{\mu}_n)|X_i = X_i]$. Further, using L-Lipschitz continuity of $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$, we obtain that

$$\begin{aligned} &\mathrm{se}^2(h;\bar{\mu}_n) \\ &\quad = \sum_{i=1}^{n} w_{i,-}(h)^2\mathbb{V}[M_i(\bar{\mu}_n)|X_i = 0^-] + \sum_{i=1}^{n} w_{i,+}(h)^2\mathbb{V}[M_i(\bar{\mu}_n)|X_i = 0^+] + o_p((nh)^{-1/2}). \end{aligned}$$

It then follows from standard local linear arguments, that $nh\,\mathrm{se}^2(h;\bar{\mu}_n) - V(\bar{\mu}_n) = o_P(1)$ and $\mathbb{E}[\widehat{\tau}(h;\bar{\mu}_n)|\mathcal{X}_n] - \tau = B(\bar{\mu}_n)h^2 + o_p(h^2)$.

Part (iii). The proof is discussed in Section 3.3.2. It also follows from Proposition 3.D.1. □

3.C.3. **Proof of Corollary 3.1.** This result follows directly from linearity of the second derivative operator.

3.C.4. **Definition of Standard Error.** We first introduce the notation. Let $\mu \in \mathcal{M}_n$. We denote the standard error by $\widehat{s}^2(h;\mu) = \sum_{i=1}^n w_i^2(h)\widehat{\sigma}_i^2(\mu)$, where

$$\widehat{\sigma}_i^2(\mu) = \frac{1}{1+H_i}\left(M_i(\mu) - \sum_{j\in\mathcal{R}_i} v_{j,i}M_j(\mu)\right)^2,$$

$$v_{j,i} = \widetilde{X_i}\left(\sum_{j\in\mathcal{R}_i}\widetilde{X_j}^\top\widetilde{X_j}\right)^{-1}\widetilde{X_j}^\top, \quad H_i = \widetilde{X_i}\left(\sum_{j\in\mathcal{R}_i}\widetilde{X_j}^\top\widetilde{X_j}\right)^{-1}\widetilde{X_i}$$

Here $\widetilde{X_i} = (1, X_i)$ and $\mathcal{R}_i$ is the set of the R nearest neighbors of unit i based on the running variable and within the same fold and on the same side of the cutoff as unit i. We note that by basic OLS algebra, the weights $v_{j,i}$ satisfy: $\sum_{j\in\mathcal{R}_i} v_{j,i} = 1$, $\sum_{j\in\mathcal{R}_i} v_{j,i}(X_j - X_i) = 0$, and $\sum_{j\in\mathcal{R}_i} v_{j,i}^2 = H_i$.

We further let $\widehat{s}_s^2(h;\mu) = \sum_{i\in I_s} w_i^2(h)\widehat{\sigma}_i^2(\mu)$, so that $\widehat{s}^2(h;\mu) = \sum_{s=1}^S \widehat{s}_s^2(h;\mu)$. Similarly, we define $\mathrm{se}_s^2(h;\mu) = \sum_{i\in I_s} w_i^2(h)\sigma_i^2(\mu)$ and $\mathrm{se}^2(h;\mu) = \sum_{s=1}^S \mathrm{se}_s^2(h;\mu)$.

3.C.5. **Proof of Proposition 3.4.1.** Using the triangular inequality, we first note that

$$\begin{aligned}|nh\,\widehat{s}_{CF}^2(h;\widehat{\mu}_n) - V(\bar{\mu}_n)| &\le nh|\widehat{s}_{CF}^2(h;\widehat{\mu}_n) - \mathrm{se}^2(h;\bar{\mu}_n)| + |nh\,\mathrm{se}^2(h;\bar{\mu}_n) - V(\bar{\mu}_n)|\\ &\le S\max_{s\in[S]} nh|\widehat{s}_s^2(h;\widehat{\mu}_{n,s}) - \mathrm{se}_s^2(h;\bar{\mu}_n)| + o_p(1),\end{aligned}$$

where the second inequality follows from the proof of Theorem 3.1. The main step in this proof is to show that for any $s \in [S]$ and conditional on $\mathcal{X}_n$ and $(W_j)_{j\in I_s^c}$, it holds that

$$nh|\widehat{s}_s^2(h;\widehat{\mu}_{n,s}) - \mathrm{se}_s^2(h;\bar{\mu}_n)| = o_P(1). \tag{3.C.2}$$

We remark that the condition in (3.C.2) would essentially follow from the results of Noack and Rothe (2021) if $\mathbb{V}[M_i(\mu)|X_i = x]$ was L-Lipschitz continuous for all $\mu \in \mathcal{T}_n$. Our setting is different as we impose L-Lipschitz continuity only for the function $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$. Still, some steps of our proof follow from the proof of Theorem 4 of Noack and Rothe (2021). We note that

$$\begin{aligned}&\widehat{s}_s^2(h;\widehat{\mu}_{n,s}) - \mathrm{se}_s^2(h;\bar{\mu}_n)\\ &= (\mathbb{E}[\widehat{s}_s^2(h;\bar{\mu}_n)|\mathcal{X}_n] - \mathrm{se}_s^2(h;\bar{\mu}_n)) + (\widehat{s}_s^2(h;\widehat{\mu}_{n,s}) - \mathbb{E}[\widehat{s}_s^2(h;\widehat{\mu}_{n,s})|\mathcal{X}_n, (W_j)_{j\in I_s^c}])\\ &\quad + (\mathbb{E}[\widehat{s}_s^2(h;\widehat{\mu}_{n,s}) - \widehat{s}_s^2(h;\bar{\mu}_n)|\mathcal{X}_n, (W_j)_{j\in I_s^c}])\\ &\equiv G_1 + G_2 + G_3.\end{aligned}$$

In the following, we show that each of the three terms is of order $o_P((nh)^{-1})$. First, it follows from the proof of Theorem 4 of Noack and Rothe (2021) that $G_1 = o_P((nh)^{-1})$ as $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$ is L-Lipschitz continuous by Assumption 3.5.

Second, it is clear that $\mathbb{E}[G_2|\mathcal{X}_n, (W_j)_{j\in I_s^c}] = 0$. Further, it follows that with probability approaching one,

$$\mathbb{E}[G_2^2|\mathcal{X}_n, (W_j)_{j\in I_s^c}] \leq \sup_{\mu\in\mathcal{T}_n} \mathbb{E}\left[\left(\widehat{s}_s^2(h;\mu) - \mathbb{E}[\widehat{s}_s^2(h;\mu)|\mathcal{X}_n]\right)^2\right] = o_p((nh)^{-2}),$$

where the last equality follows from the proof of Theorem 4 of Noack and Rothe (2021) using boundedness of the fourth conditional moment assumed in the proposition.

We now consider G_3. We note that with probability approaching one

$$\begin{aligned}|G_3| &= |\sum_{i\in I_s} w_i^2(h)\mathbb{E}[\widehat{\sigma}_i^2(\widehat{\mu}_{n,s}) - \widehat{\sigma}_i^2(\bar{\mu}_n)|\mathcal{X}_n, (W_j)_{j\in I_s^c}]| \\ &\leq \sup_{j\in I_s\colon X_j\in\mathcal{X}_h} \sup_{\mu\in\mathcal{T}_n} \left|\mathbb{E}[\widehat{\sigma}_j^2(\mu) - \widehat{\sigma}_j^2(\bar{\mu}_n)|\mathcal{X}_n]\right| \sum_{i\in I_s} w_i(h)^2.\end{aligned}$$

Following Noack and Rothe (2021), we note that for any $\mu \in \mathcal{T}_n$ and any $i \in I_s$

$$\begin{aligned}\mathbb{E}[\widehat{\sigma}_i(\mu)|\mathcal{X}_n] = \sigma_i^2(\mu) &+ \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_{j,i}^2(\sigma_j^2(\mu) - \sigma_i^2(\mu))\right) \\ &+ \frac{1}{1+H_i}\left(\mathbb{E}[M_i(\mu)|X_i] - \sum_{j\in\mathcal{R}_i} v_{j,i}\mathbb{E}[M_j(\mu)|X_j]\right)^2.\end{aligned} \tag{3.C.3}$$

In the following, we denote by C a positive constant, which might be different from line to line. By a second-order Taylor-expansion and by a simple OLS-algebra, it holds for the last term in the above expression that

$$\begin{aligned}&\sup_{i\in I_s\colon X_i\in\mathcal{X}_h} \sup_{\mu\in\mathcal{T}_n} \frac{1}{1+H_i}\left(\mathbb{E}[M_i(\mu)|X_i] - \sum_{j\in\mathcal{R}_i} v_{j,i}\mathbb{E}[M_j(\mu)|X_j]\right)^2 \\ &\leq C \sup_{i\in I_s\colon X_i\in\mathcal{X}_h} \sup_{j\in\mathcal{R}_i} |X_i - X_j|^4 \sup_{x\in\mathcal{X}_h} \sup_{\mu\in\mathcal{T}_n} (\partial_x^2\mathbb{E}[M_i(\mu)|X_i = x])^2 = o_p(1),\end{aligned} \tag{3.C.4}$$

where we used that $\frac{1}{1+H_i}\sum_{j\in\mathcal{R}_i} v_{j,i}^2 \leq 1$ and $\sup_{x\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n} \partial_x^2\mathbb{E}[M_i(\mu)|X_i = x] = O(1)$ by Assumptions 3.4 and 3.5.

Using (3.C.3) and (3.C.4), we obtain that

$$\begin{aligned}
&\sup_{i\in I_s\colon X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}|\mathbb{E}[\widehat{\sigma}_i^2(\mu)-\widehat{\sigma}_i^2(\bar{\mu}_n)|\mathcal{X}_n]| \\
&\le \sup_{i\in I_s\colon X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}\left|\sigma_i^2(\mu)-\sigma_i^2(\bar{\mu}_n)+\frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i}v_{j,i}^2(\sigma_j^2(\mu)-\sigma_j^2(\bar{\mu}_n)+\sigma_i^2(\bar{\mu}_n)-\sigma_i^2(\mu))\right)\right| \\
&\quad + o_p(1) \\
&\le C\sup_{i\in I_s\colon X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}|\sigma_i^2(\mu)-\sigma_i^2(\bar{\mu}_n)|+o_p(1) \\
&\le C\sup_{x\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}|\mathbb{V}[M_i(\mu)|X_i=x]-\mathbb{V}[M_i(\bar{\mu}_n)|X_i=x]|+o_p(1)=o_p(1),
\end{aligned}$$

where we used that $\frac{1}{1+H_i}\sum_{j\in\mathcal{R}_i}v_{j,i}^2\le 1$ and Assumption 3.3.

Since $\sum_{i\in I_s}w_i(h)^2=O_p((nh)^{-1})$, we conclude that $G_3=o_P((nh)^{-1})$.

3.C.6. **Proofs for sufficient conditions in Section 3.6.**

Proof of Proposition 3.6.1. We start by showing that Assumption 3.3 holds. It follows from basic OLS algebra that there exists $\bar{\beta}_Z$ such that for all $s\in[S]$ it holds that $\|\widehat{\beta}_{s,Z}-\bar{\beta}_Z\|_\infty=O_P((nh_1)^{-1/2})$. This implies that $\widehat{\beta}_{s,Z}\in[\bar{\beta}_{s,Z}\pm(nh_1)^{-1/2}v_n]$ w.p.a. 1. Let $v_n\to\infty$ be a sequence s.t. $(nh_1)^{-1/2}v_n\to 0$. We define

$$\mathcal{T}_n=\{\mu:\mu(z)=\beta^\top z,\text{ where }\beta\in\mathcal{B}_n=[\bar{\beta}_Z\pm(nh_1)^{-1/2}v_n]\}.$$

By construction, $\bar{\mu}\in\mathcal{T}_n$ and $\mathbb{P}[\widehat{\mu}_{n,s}\in\mathcal{T}_n]=1+o(1)$ for all $s\in[S]$. Assumption 3.3 follows by noting that

$$\sup_{\beta\in\mathcal{B}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[(\beta^\top Z_i-\bar{\beta}_Z^\top Z_i)^2|X_i=x\right]\le d\sup_{\beta\in\mathcal{B}_n}\|\beta-\bar{\beta}_Z\|_\infty^2\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[Z_i^\top Z_i|X_i=x\right]=o(1).$$

We now consider Assumption 3.4. For $j\in\{1,2\}$, all $\beta\in\mathcal{B}_n$ and $x\in\mathcal{X}\setminus\{0\}$, we have that

$$\partial_x^j\mathbb{E}\left[\beta^\top Z_i-\bar{\beta}_Z^\top Z_i|X_i=x\right]=(\beta_Z-\bar{\beta}_Z)^\top\partial_x^j\mathbb{E}\left[Z_i|X_i=x\right],$$

which concludes this proof. □

Proof of Proposition 3.6.2. We start by showing that Assumption 3.3 holds. Let v_n be a sequence such that $v_n\to\infty$ and $r_nv_n\to 0$. We define

$$\mathcal{T}_n=\{\mu:\mu(z)=m_\beta(z),\text{ where }\beta\in\mathcal{B}_n=[\bar{\beta}\pm r_nv_n]\}.$$

By construction, $\bar{\mu}\in\mathcal{T}_n$ and $\mathbb{P}[\widehat{\mu}_{n,s}\in\mathcal{T}_n]=1+o(1)$ for all $s\in[S]$. Assumption 3.3 follows by noting that

$$\sup_{\beta\in\mathcal{B}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}[(m_\beta(Z_i)-m_{\bar\beta}(Z_i))^2|X_i=x]\leq\sup_{\beta\in\mathcal{B}_n}\|\beta-\bar\beta\|_\infty^2G^2=o(1).$$

We now consider Assumption 3.4. Under the assumptions made, for $j\in\{1,2\}$, all $\beta\in\mathcal{B}_n$, and $x\in\mathcal{X}\setminus\{0\}$, we have that

$$\partial_x^j\mathbb{E}[m_\beta(Z_i)-m_{\bar\beta}(Z_i)|X_i=x]=\int(m_\beta(z)-m_{\bar\beta}(z))\partial_x^jf_{Z|X}(z|x)dz.$$

It then follows that for $j\in\{1,2\}$

$$\begin{aligned}&\sup_{\beta\in\mathcal{B}_n}\sup_{x\in\mathcal{X}_h\setminus\{0\}}\left|\partial_x^j\mathbb{E}[m_\beta(Z_i)-m_{\bar\beta}(Z_i)|X_i=x]\right|\\&\qquad\leq G\sup_{\beta\in\mathcal{B}_n}\|\beta-\bar\beta\|_\infty\int_{\mathcal{Z}}|\partial_x^jf_{Z|X}(z|x)|dz=o_P(1),\end{aligned}$$

which concludes the proof. □

For completeness, we restate the classic assumptions for uniform convergence of the local linear estimator used by Masry (1996).

Assumption 3.8. *(i) (X_i,Z_i) are continuously distributed, and $\mathcal{X}$ and $\mathcal{Z}$ are compact and convex; (ii) The joint density $f(x,z)$ is bounded, has bounded first-order derivatives, and is bounded away from zero for all $(x,z)\in\mathcal{X}\times\mathcal{Z}$; (iii) $\mathbb{E}[Y_i|X_i=x,Z_i=z]$ is twice continuously differentiable w.r.t. x and z and the second derivatives are Lipschitz continuous; (iv) $\sup_{x,z}\mathbb{E}[|Y_i|^{2+\delta}|X_i=x,Z_i=z]<\infty$ for some constant $\delta>0$; (v) For $j\in\{0,...,3\}$, $H_j(u)\equiv u^jK(u)$ is Lipschitz continuous;*

Proof of Proposition 3.6.3. By Theorem 6 of Masry (1996), $\sup_{z\in\mathcal{Z}_h}\|\widehat\mu_n(z)-\mu_n(z)\|=O_P(r_n)$, where $r_n=o(1)$. Hence, the set $\mathcal{T}_n$ can be chosen s.t. $\sup_{\mu\in\mathcal{T}_n}\|\mu(Z_i)-\mu_n(Z_i)\|_\infty=o(1)$. Assumption 3.3 follows trivially. Assumption 3.4 is also satisfied, as discussed in Section 3.A.2. □

3.D. VARIANCE CALCULATIONS

In this section, we provide formal derivations for the optimality result discussed in Section 3.5.3 and for the discussion of variance reductions in comparison to the standard RD estimator discussed in Section 3.4. Recall that

$$\begin{aligned}\widetilde V(\mu)&=\omega_+\mathbb{V}[M_i(\mu)|X_i=0^+]+\omega_-\mathbb{V}[M_i(\mu)|X_i=0^-],\\\mu_n^*(z)&=\frac{\omega_-}{\omega_-+\omega_+}\mu_n^-(z)+\frac{\omega_+}{\omega_-+\omega_+}\mu_n^+(z).\end{aligned}$$

We obtain the variance $V(\mu)$ and the function μ_n as a special case when $\omega_+=\omega_-=1$.

Proposition 3.D.1. *Suppose that Assumptions 3.1–3.5 hold. Then for all $\mu \in \mathcal{M}_n$, it holds that:*

(i) $\widetilde{V}(\mu_n^*) \leq \widetilde{V}(\mu)$ *with* $\widetilde{V}(\mu_n^*) = \widetilde{V}(\mu)$ *if and only if* $\mathbb{V}[\mu(Z_i) - \mu_n^*(Z_i)|X_i = 0] = 0$*;*

(ii) $\widetilde{V}(\mu) < \widetilde{V}(0)$ *if and only if* $\mathbb{V}[\mu_n(Z_i) - \mu(Z_i)|X_i = 0] < \mathbb{V}[\mu_n(Z_i)|X_i = 0]$.

Proof. Fix μ, $\widetilde{\mu} \in \mathcal{M}_n$. By basic properties of the conditional expectation, we have that

$$\widetilde{V}(\mu) = \omega_+ \mathbb{V}[Y_i - \mu_n^+(Z_i)|X_i = 0^+] + \omega_- \mathbb{V}[Y_i - \mu_n^-(Z_i)|X_i = 0^-] + \widetilde{\mathcal{V}}(\mu),$$

where the first two terms on the right-hand side do not depend on μ, and

$$\widetilde{\mathcal{V}}(\mu) = \omega_+ \mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0] + \omega_- \mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0].$$

Further, it holds that

$$\begin{aligned}\widetilde{\mathcal{V}}(\mu) = \widetilde{\mathcal{V}}(\mu_n^* + \mu - \mu_n^*) &= \omega_+ \mathbb{V}\left[\frac{\omega_-}{\omega_+ + \omega_-}(\mu_n^+(Z_i) - \mu_n^-(Z_i)) - (\mu(Z_i) - \mu_n^*(Z_i))|X_i = 0\right] \\ &+ \omega_- \mathbb{V}\left[\frac{-\omega_+}{\omega_+ + \omega_-}(\mu_n^+(Z_i) - \mu_n^-(Z_i)) - (\mu(Z_i) - \mu_n^*(Z_i))|X_i = 0\right] \\ &= \widetilde{V}(\mu_n^*) + (\omega_+ + \omega_-)\mathbb{V}[\mu(Z_i) - \mu_n^*(Z_i)|X_i = 0].\end{aligned}$$

Hence, $\widetilde{V}(\mu) < \widetilde{V}(\widetilde{\mu})$ if and only if $\mathbb{V}[\mu(Z_i) - \mu_n^*(Z_i)|X_i = 0] < \mathbb{V}[\widetilde{\mu}(Z_i) - \mu_n^*(Z_i)|X_i = 0]$, and similarly with equalities instead of inequalities. Both parts of the lemma follow. □

3.E. ADDITIONAL SIMULATION RESULTS

In this section, we present further simulation results. Table 3.E.1 extends the results in Table 3.1. Apart from the bias-aware approach discussed in the main text, we consider bandwidth choices and confidence intervals based on robust bias corrections and undersmoothing.[20] The qualitative conclusions about the relative performance of different first-stage estimators in different models remain the same as discussed in the main text.

The simulated mean bandwidth of robust bias corrections is on average smaller than that of the bias-aware approach, and the confidence intervals are larger. This feature is known in the nonparametric literature. In the last two rows of Table 3.E.1 we report the results using the procedure of Calonico et al. (2019). In this simulation setting, they are essentially the same as the results for our procedure with a linear adjustment function.

In Table 3.E.2, we report simulation results for Model 3 for different values of the signal-to-noise ratio. This illustrates that the potential gains from covariate adjustments are large if the covariates explain a large portion of variation in the outcome variable.

[20] The bandwidth for undersmoothing is chosen as $n^{-1/20}$ times the MSE-optimal bandwidth estimated using the `rdrobust` package.

Table 3.E.1: Full simulation results for different numbers of relevant covariates.

		Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
		Model 1: L=0					**Model 2: L=4**					**Model 3: L=10**					**Model 4: L=25**				
Standard	BA	97.0	-1.4	7.4	32.4	43.2	96.1	-7.1	18.6	81.8	68.8	96.4	-9.5	19.1	87.6	79.3	95.9	-6.3	18.5	81.0	68.5
	RBC	94.8	1.5	11.0	41.5	29.9	94.7	0.0	35.1	130.9	30.5	94.6	1.1	37.1	140.0	26.9	94.2	1.6	39.3	145.7	24.5
	US	94.9	0.6	11.3	42.5	20.5	94.5	-1.1	36.0	133.4	20.9	94.7	0.1	38.0	142.4	18.4	94.3	0.8	40.5	148.4	16.7
Optimal Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2	96.5	-1.3	7.6	32.5	43.2	96.9	-1.3	7.4	32.4	43.2
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.6	1.5	11.3	41.5	29.9	94.2	1.5	11.0	41.4	30.0
	US	94.9	0.6	11.3	42.5	20.5	94.5	0.3	11.3	42.5	20.4	93.7	0.5	11.6	42.6	20.4	94.4	0.5	11.3	42.5	20.5
Linear Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2	96.7	-4.8	12.7	56.2	61.8	96.8	-4.3	10.3	47.2	59.0
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.7	1.3	23.4	85.9	26.3	94.6	0.7	19.9	75.4	19.7
	US	94.9	0.6	11.3	42.5	20.5	94.5	0.3	11.3	42.5	20.4	94.1	0.3	23.9	87.6	18.0	94.2	0.2	20.5	76.6	13.4
Linear	BA	97.0	-1.4	7.4	32.7	43.3	96.7	-1.5	7.5	32.6	43.3	95.9	-4.0	13.7	59.1	59.7	96.5	-4.3	10.8	49.2	58.8
	RBC	94.8	1.5	11.0	41.8	30.0	94.3	1.4	11.1	41.8	29.9	94.0	1.6	25.0	91.8	27.9	94.3	0.7	21.6	81.0	20.3
	US	95.1	0.6	11.3	42.9	20.5	94.6	0.3	11.4	42.8	20.5	94.2	0.6	25.6	93.7	19.1	94.4	0.2	22.2	82.4	13.9
Local Linear	BA	97.0	-1.4	7.4	32.7	43.3	96.8	-1.4	7.5	32.7	43.3	96.3	-1.6	8.3	35.6	45.2	96.8	-1.6	8.2	35.9	45.6
	RBC	94.5	1.5	11.1	41.9	30.0	94.5	1.4	11.1	41.9	29.9	94.3	1.4	12.7	47.1	29.3	94.2	1.5	13.0	49.0	27.8
	US	94.9	0.6	11.4	42.9	20.5	94.7	0.4	11.4	43.0	20.5	94.3	0.5	13.1	48.2	20.0	94.3	0.5	13.5	50.1	19.0
Lasso	BA	96.7	-1.4	7.6	33.1	43.6	96.6	-2.1	8.8	38.3	46.6	96.2	-2.0	9.2	39.1	46.7	96.8	-1.4	7.7	34.0	44.3
	RBC	94.4	1.5	11.6	43.5	28.8	95.0	1.2	13.8	52.1	29.1	93.9	1.3	14.6	53.0	29.5	94.3	1.1	13.2	49.0	24.3
	US	95.1	0.7	11.8	44.5	19.7	94.7	0.2	14.2	53.2	19.9	94.1	0.4	15.0	54.2	20.2	94.2	0.5	13.5	50.0	16.6
Forest	BA	96.8	-1.5	7.6	33.1	43.6	96.7	-2.1	8.7	37.9	46.5	96.6	-1.9	8.5	37.2	46.9	97.1	-2.2	9.3	41.3	49.0
	RBC	94.6	1.5	11.3	42.5	29.9	94.9	1.0	13.4	50.7	29.7	94.0	1.0	15.2	56.0	23.3	94.0	0.8	18.6	68.8	19.8
	US	94.6	0.6	11.6	43.6	20.5	94.8	0.0	13.8	51.8	20.3	94.3	0.3	15.5	57.0	15.9	94.3	0.4	19.1	70.1	13.6
CCFT	RBC	94.5	1.4	11.0	41.3	29.7	93.9	1.3	11.1	41.3	29.7	93.4	1.3	23.5	85.1	26.3	94.3	0.7	20.1	74.6	19.6
	US	94.4	0.6	11.4	42.2	20.3	94.0	0.3	11.4	42.2	20.3	93.4	0.3	24.1	86.3	18.0	93.5	0.2	20.8	75.2	13.4

Notes: Results based on 5,000 Monte Carlo draws based on Model 3 explained in the main text. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h). Bandwidth and confidence intervals are constructed based on the bias-aware approach (BA), robust bias correction (RBC), and undersmoothing (US).

Table 3.E.2: Simulation results for different signal-to-noise ratios.

		Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
		$\rho = 3$					$\rho = 1$					$\rho = 5$					$\rho = 10$				
Standard	BA	96.2	-8.9	19.6	87.5	79.3	96.4	-3.1	9.8	43.3	52.0	96.0	-14.7	29.6	134.7	95.5	95.5	-16.8	58.6	241.5	99.9
	RBC	94.7	1.1	37.2	139.6	26.9	94.1	0.8	16.4	61.2	27.9	94.6	0.9	59.8	226.5	26.7	94.7	-0.1	118.3	446.9	26.7
	US	94.3	0.4	38.2	142.1	18.4	93.8	-0.2	16.9	62.4	19.1	94.5	-0.2	61.4	230.3	18.3	94.8	-0.5	120.5	454.4	18.3
Optimal Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.4	32.5	43.2	96.5	-1.3	7.6	32.4	43.2	96.9	-1.3	7.3	32.5	43.2
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.5	1.5	11.3	41.5	29.8	94.2	1.5	11.0	41.4	30.0
	US	94.9	0.6	11.3	42.5	20.5	94.5	0.3	11.3	42.5	20.4	93.6	0.6	11.6	42.6	20.4	94.5	0.5	11.3	42.5	20.5
Linear Inf	BA	96.0	-4.8	12.9	56.2	61.9	96.2	-1.9	8.3	36.1	46.1	96.4	-8.5	18.1	81.5	76.6	95.9	-14.1	33.0	146.9	96.3
	RBC	94.2	1.2	23.1	85.8	26.4	94.0	1.2	13.1	48.9	28.3	93.8	1.2	35.8	131.6	25.8	94.8	1.2	66.2	252.9	25.6
	US	93.9	0.6	23.8	87.5	18.0	94.4	0.2	13.4	49.9	19.4	94.1	0.1	36.6	134.0	17.7	95.0	0.6	67.2	257.4	17.5
Linear	BA	96.1	-4.0	13.8	59.1	59.8	96.1	-1.9	8.4	36.5	45.9	95.6	-7.2	21.2	90.7	74.0	95.8	-14.1	37.0	161.5	95.4
	RBC	94.0	1.7	24.9	91.9	27.9	94.0	1.2	13.2	49.3	28.6	94.0	1.7	42.2	155.3	28.8	94.6	2.3	83.4	312.8	29.0
	US	93.9	1.0	25.6	93.8	19.1	94.3	0.3	13.5	50.4	19.5	94.2	0.7	43.3	158.4	19.7	94.8	1.0	85.5	318.8	19.8
Local Linear	BA	96.7	-1.7	8.2	35.6	45.1	96.5	-1.5	7.8	33.9	44.1	96.5	-1.7	8.7	37.3	46.3	97.0	-2.6	10.2	45.1	52.3
	RBC	94.4	1.5	12.5	47.1	29.3	94.2	1.3	11.7	43.9	29.7	94.0	1.5	13.7	50.5	28.8	94.6	1.6	17.4	65.1	27.6
	US	94.7	0.7	12.9	48.2	20.1	94.3	0.4	12.0	45.0	20.3	94.0	0.6	14.1	51.7	19.7	94.6	0.7	18.0	66.4	18.9
Lasso	BA	96.8	-2.0	9.1	39.3	46.9	96.8	-1.6	7.7	34.0	44.5	96.1	-2.7	11.5	48.4	51.0	96.2	-4.9	18.1	75.8	61.6
	RBC	93.8	1.5	14.4	53.3	29.6	94.3	1.0	12.4	46.9	26.5	93.9	1.5	18.5	67.7	31.1	94.1	1.9	32.6	117.7	32.2
	US	94.3	0.6	14.9	54.6	20.2	94.4	0.3	12.7	47.9	18.1	94.2	0.6	19.1	69.2	21.3	94.2	0.8	33.3	120.1	22.0
Forest	BA	96.6	-1.9	8.5	37.2	46.9	96.5	-1.6	7.7	33.7	44.3	96.7	-2.6	10.0	43.8	51.1	96.3	-5.8	14.7	64.5	64.5
	RBC	94.1	1.1	15.1	56.1	23.1	94.1	1.1	12.1	45.3	27.7	94.5	0.8	18.9	70.6	21.8	93.9	0.5	31.7	116.1	20.7
	US	94.3	0.6	15.5	57.2	15.8	94.5	0.2	12.4	46.2	19.0	95.0	0.2	19.3	71.8	14.9	94.2	0.1	32.5	118.1	14.2
CCFT	RBC	93.8	1.2	23.3	85.0	26.3	93.5	1.1	13.2	48.6	28.1	93.4	1.2	36.0	130.3	25.8	94.6	1.3	66.2	250.4	25.6
	US	93.3	0.6	24.0	86.3	18.0	93.9	0.2	13.5	49.4	19.2	93.3	0.2	36.9	132.0	17.7	94.7	0.5	67.4	253.4	17.5

Notes: Results based on 5,000 Monte Carlo draws based on Model 3 explained in the main text. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h). Bandwidth and confidence intervals are constructed based on the bias-aware approach (BA), robust bias correction (RBC), and undersmoothing (US).

References

Abadie, A., G. W. Imbens, and F. Zheng (2014): “Inference for misspecified models with fixed regressors,” *Journal of the American Statistical Association*, 109, 1601–1614.

Anderson, T. W., H. Rubin, et al. (1949): “Estimation of the parameters of a single equation in a complete system of stochastic equations,” *Annals of Mathematical statistics*, 20, 46–63.

Andrews, D. (1994): “Asymptotics for semiparametric econometric models via stochastic equicontinuity,” *Econometrica*, 62, 43–72.

Andrews, D. W. and G. Soares (2010): “Inference for parameters defined by moment inequalities using generalized moment selection,” *Econometrica*, 78, 119–157.

Armstrong, T. B. and M. Kolesár (2020): “Simple and honest confidence intervals in nonparametric regression,” *Quantitative Economics*, 11, 1–39.

Barendse, S. (2020): “Efficiently Weighted Estimation of Tail and Interquantile Expectations,” *Working Paper.*

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): “Sparse models and methods for optimal instruments with an application to eminent domain,” *Econometrica*, 80, 2369–2429.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): “Program evaluation and causal inference with high-dimensional data,” *Econometrica*, 85, 233–298.

Bickel, P. J. (1975): “One-Step Huber Estimates in the Linear Model,” *Journal of the American Statistical Association*, 70, 428–434.

Breiman, L. (2001): “Random forests,” *Machine learning*, 45, 5–32.

Cai, Z. and X. Wang (2008): “Nonparametric estimation of conditional VaR and expected shortfall,” *Journal of Econometrics*, 147, 120–130.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the effect of bias estimation on coverage accuracy in nonparametric inference," *Journal of the American Statistical Association*, 113, 767–779.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression Discontinuity Designs Using Covariates," *The Review of Economics and Statistics*, 101, 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A practical introduction to regression discontinuity designs: Foundations*, Cambridge University Press.

Cattaneo, M. D., M. Jansson, and X. Ma (2020): "Simple local polynomial density estimators," *Journal of the American Statistical Association*, 115, 1449–1455.

Chen, S. (2008): "Nonparametric Estimation of Expected Shortfall," *Journal of Financial Econometrics*, 6, 87–107.

Chen, X. and C. A. Flores (2015): "Bounds on treatment effects in the presence of sample selection and noncompliance: the wage effects of Job Corps," *Journal of Business & Economic Statistics*, 33, 523–540.

Cheng, M.-Y. (1997): "Boundary aware estimators of integrated density derivative products," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 191–203.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Dimitriadis, T., S. Bayer, et al. (2019): "A joint quantile and expected shortfall regression framework," *Electronic Journal of Statistics*, 13, 1823–1871.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66, CRC Press.

Fan, J., T.-C. Hu, and Y. K. Truong (1994): "Robust non-parametric function estimation," *Scandinavian journal of statistics*, 433–446.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2020): "Estimation of Conditional Average Treatment Effects With High-Dimensional Data," *Journal of Business & Economic Statistics*, 0, 1–15.

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep neural networks for estimation and inference," *Econometrica*, 89, 181–213.

Frölich, M. and M. Huber (2019): "Including Covariates in the Regression Discontinuity Design," *Journal of Business & Economic Statistics*, 37, 736–748.

Gerard, F., M. Rokkanen, and C. Rothe (2016): "Identification and inference in regression discontinuity designs with a manipulated running variable," .

——— (2020): "Bounds on treatment effects in regression discontinuity designs with a manipulated running variable," *Quantitative Economics*, 11, 839–870.

Guerre, E. and C. Sabbah (2012): "Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function," *Econometric Theory*, 87–129.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Hong, S.-Y. (2003): "Bahadur representation and its applications for local polynomial estimates in nonparametric M–regression," *Journal of Nonparametric Statistics*, 15, 237–251.

Horowitz, J. L. and C. F. Manski (1995): "Identification and robustness with contaminated and corrupted data," *Econometrica*, 63, 281–302.

Imbens, G. and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of economic studies*, 79, 933–959.

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

Imbens, G. W. and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

Jones, M. C. (1993): "Simple boundary correction for kernel density estimation," *Statistics and computing*, 3, 135–146.

Jurečková, J. (1984): "Regression quantiles and trimmed least squares estimator under a general design," *Kybernetika*, 20, 345–357.

Kato, K. (2012): "Weighted Nadaraya–Watson Estimation of Conditional Expected Shortfall," *Journal of Financial Econometrics*, 10, 265–291.

Kennedy, E. H. (2020): "Optimal doubly robust estimation of heterogeneous causal effects," *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 1229.

Koenker, R. and G. Bassett Jr (1978): "Regression quantiles," *Econometrica*, 33–50.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

Lejeune, M. and P. Sarda (1992): "Smooth Estimators of Distribution and Density Functions," *Computation Statistics & Data Analysis*, 14, 457–471.

Linton, O. and Z. Xiao (2013): "Estimation of and inference about the expected shortfall for time series with infinite variance," *Econometric Theory*, 29, 771–807.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571–599.

Masry, E. and J. Fan (1997): "Local polynomial estimation of regression functions for mixing processes," *Scandinavian Journal of Statistics*, 24, 165–179.

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

Negi, A. and J. M. Wooldridge (2020): "Revisiting regression adjustment in experiments with heterogeneous treatment effects," *Econometric Reviews*, 1–31.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Neyman, J. (1959): "Optimal asymptotic tests of composite hypotheses," *Probability and statsitics*, 213–234.

——— (1979): "C (α) tests and their use," *Sankhyā: The Indian Journal of Statistics, Series A*, 1–21.

Noack, C. and C. Rothe (2021): "Bias-aware inference in fuzzy regression discontinuity designs," *arXiv preprint arXiv:1906.04631.*

Pollard, D. (1991): "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 186–199.

Ruppert, D. and R. J. Carroll (1978): "Robust regression by trimmed least squares estimation," Tech. rep., Department of Statistics, University of North Carolina at Chapel Hill.

——— (1980): "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.

Scaillet, O. (2005): "Nonparametric estimation of conditional expected shortfall," *Insurance and Risk Management Journal*, 74, 639–660.

Semega, J., M. Kollar, E. A. Shrider, and J. Creamer (2020): *Income and Poverty in the United States: 2019*, United States Census Bureau.

Semenova, V. (2020): "Better Lee Bounds," *arXiv preprint arXiv:2008.12720.*

Shorack, G. R. et al. (1974): "Random means," *The Annals of Statistics*, 2, 661–675.

Stoye, J. (2009): "More on confidence intervals for partially identified parameters," *Econometrica*, 77, 1299–1315.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Wager, S. and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016): "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.

Zhang, J. L. and D. B. Rubin (2003): "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death"," *Journal of Educational and Behavioral Statistics*, 28, 353–368.

Curriculum Vitae

Tomasz Olma

Education:

2017–2021	University of Mannheim (Germany) *Ph.D. in Economics*
2016–2017	University of California, Berkeley (USA) *Visiting Student, Department of Economics*
2015–2017	University of Mannheim (Germany) *M.Sc. in Economics*
2012–2015	University of Warsaw (Poland) *B.Sc. in Mathematics*
2011–2014	Warsaw School of Economics (Poland) *B.Sc. in Quantitative Methods in Economics and Information Systems*

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehorde vorgelegen.

Mannheim, den 7. Juni 2021 Tomasz Olma