Robust Methods for Program Evaluation

by

Nestor Sebastian Calonico

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2014

Doctoral Committee:

Associate Professor Matias D. Cattaneo, Co-Chair
Professor Jeffrey A. Smith, Co-Chair
Professor Richard A. Hirth
Professor Lutz Kilian

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Robust Methods for Program Evaluation

by

Sebastian Calonico

My dissertation research focuses on different approaches to conduct robust estimation and inference in the context of program evaluation.

In Chapter 1, I look at the effects of teacher and peer characteristics on student achievement in the STAR Project conducted in Tennessee in the late 1980s. As in standard linear models, the proposed approach considers two types of unobservables: school-specific effects and idiosyncratic disturbances. It generalizes previous empirical research by allowing both effects to enter the structural function nonseparably. No functional form assumptions are needed for identification. Instead, it uses an exchangeability condition in the way that covariates affect the distribution of the school-specific effects. The model permits nonparametric distributional and counterfactual analysis of heterogeneous effects: it extends policy analysis beyond marginal or discrete changes to consider distributional effects originating from a counterfactual change in the distribution of characteristics of classrooms, peers and teachers. Also, these impacts can be analyzed on any feature of the distribution of student achievement, such as quantiles and inequality measures. The empirical analysis looks at the effects of class size, teacher experience and gender composition of the classroom

on test scores. Findings suggest that nonseparable heterogeneity is an important source of individual-level variation in academic performance. The impact of class size is considerably larger using my approach: students in smaller classes benefit about 0.3 standard deviations, compared to a 0.16 effect obtained using a standard linear model. Also, teacher experience has a stronger, nonlinear impact. Still, the distributional analysis suggests that these gains are hard to achieve when facing resource constraints.

In Chapter 2, a joint work with Matias Cattaneo and Rocio Titiunik, we study robust inference in the context of regression discontinuity (RD) design. In the RD approach, units are assigned to treatment based on whether their value of an observed covariate exceeds a known cutoff. Local polynomial estimators are routinely employed to construct confidence intervals for treatment effects. The performance of these confidence intervals in applications, however, may be seriously hampered by their sensitivity to the specific bandwidth employed. Available bandwidth selectors typically yield a "large" bandwidth, leading to data-driven confidence intervals that may be severely biased, with empirical coverage well below their nominal target. We propose new theory-based, more robust confidence interval estimators for average treatment effects at the cutoff in sharp RD, sharp kink RD, fuzzy RD and fuzzy kink RD designs. Our proposed confidence intervals are constructed using a bias-corrected RD estimator together with a novel standard error estimator. For practical implementation, we discuss mean-square error optimal bandwidths, which are by construction not valid for conventional confidence intervals but valid with our robust approach, and consistent standard error estimators based on our new variance formulas. Among other possibilities, our results give formal justification to simple inference procedures based on increasing the order of the local polynomial estimators employed. We find in a simulation study that our confidence intervals exhibit close-to-correct empirical coverage and good empirical interval length on average, remarkably improving upon

the alternatives available in the literature.

Finally, in Chapter 3, also written jointly with Matias Cattaneo and Rocio Titiunik, we present new results regarding RD plots. Exploratory data analysis plays a central role in applied statistics and econometrics. Specially in the RD approach, the use of graphical analysis has been advocated because it provides both easy presentation and transparent validation of the design [e.g., Imbens and Lemieux (2008, Section 3) and Lee and Lemieux (2010, Section 4.1)]. RD plots are nowadays widely used in applications, despite its formal properties being unknown: these plots are typically presented employing *ad hoc* choices of tuning parameters, which makes these procedures less automatic and more subjective. We formally study the most common RD plot based on an evenly-spaced binning of the data, and propose an optimal data-driven choice for the number of bins. This leads to an RD plot that is constructed objectively using the data available. In addition, we introduce an alternative RD plot based on quantile-spaced binning, study its formal properties, and propose the corresponding optimal data-driven choice for the number of bins. The main proposed data-driven selectors employ spacings-based estimators, which are simple and easy to implement in applications because they do not require additional choices of tuning parameters. Altogether, our results offer two alternative RD plots that are objective and automatic when implemented, thereby providing a reliable benchmark for empirical work using RD designs. We illustrate the performance of our automatic RD plots using two empirical applications and a Monte Carlo study.

CHAPTER I

Identifying Distributional Effects of Teachers and Peers in Nonseparable Models

1.1 Introduction

The effects of educational inputs such as class size, teaching quality and school resources on student achievement have long been studied in the economic literature. In a highly influential work, Hanushek (1986) concludes that the literature does not provide strong evidence of a consistent relationship between school resources and student performance. A positive effect of school inputs, particularly teaching quality, has instead been highlighted in more recent work. For example, Card and Krueger 1992; 1996 find a positive relationship between school resources and student achievement, showing that both low pupil-teacher ratios and high quality school systems lead to higher future earnings for students. Mixed conclusions have been reached on the effect of class size on student performance: while some studies conclude that small classes do not improve student achievement (e.g., (Hanushek, 2003), (Hoxby, 2000)), others find evidence of a positive impact (e.g., Krueger (1999), Krueger and Whitmore (2001), Angrist and Lavy (1999)).

These contrasting results have usually been attributed to econometric problems that make it difficult to recover the causal effect of educational inputs on student

performance, especially those related to omitted variable bias and reverse causality. Early studies have often relied on data in which the allocation of students to classes was not the result of an exogenous assignment. For example, schools might assign less able students to smaller classes, or better teachers to larger ones. In other cases, the allocation of students to classes is not exogenous due to parent decisions, for example parents more concerned about the education of their children may choose schools with a smaller class size or more experienced teachers.

With the aim to provide more reliable estimates, recent studies have relied on controlled randomized experiments or natural experiments. Most notably, a number of works have used data from the STAR Project, conducted in Tennessee from 1985-89. This was a large-scale, longitudinal experimental study of reduced class size, where students and teachers were randomly allocated to different class sizes. It motivated a large body of research on the effects of different classroom characteristics (not only class size, but also other factors such as teacher experience) on student performance both in the short and long-run. Most of the studies conclude that smaller classes increase student achievement, even after controlling for school fixed effects and teacher characteristics (e.g., Krueger (1999), Krueger and Whitmore (2001), Nye, Konstantopoulos, and Hedges (2004)).

Besides relying on an experimental setting, a common feature of all these studies is that they are based on linear models, where we can account for unobserved school heterogeneity by including school dummies, which can be handled with standard linear panel methods such as within-group transformations. I propose to extend this approach by considering a more general, nonseparable model that does not impose any functional form or parametric assumptions. In particular, no additivity or monotonicity assumptions are required for identification. Using the STAR Project data, I look at the influence of teacher experience, class size and gender composition of the classroom on student performance on standardized tests. The main motivation

behind this analysis is based on several important limitations of the standard model, which usually assumes a linear specification of the form:

$$Y_{ics} = \mu_s + P_{cs}\beta + Z'_{ics}\gamma + \varepsilon_{ics} \tag{1.1}$$

where Y_{ics} is a measure of achievement (e.g., kindergarten test scores) for student i assigned to classroom c in school s, P_{cs} is a classroom characteristic (e.g., class size or teacher experience). Finally, Z_{ics} accounts for other student and teacher characteristics that affect student performance. The main interest is on the parameter β, e.g, the impact of class size or teacher experience on test scores. The model also includes two types of unobserved heterogeneity: a school fixed effect μ_s and an individual specific component, ε_{ics}.

The model presents several limitations, mostly derived from the *linearity* assumption. This is crucial for identification, since the school fixed effect μ_s is usually differenced out. Besides being subject to model misspecification, this imposes important limitations for the analysis of *heterogeneity* in terms of the relationship between the impact of the covariates and μ_s. The marginal effect of P_{cs} on Y_{ics} (e.g., a marginal change in the gender composition of the classroom) is: $\partial Y_{ics}/\partial P_{cs} = \beta$, or $\beta\,(p_1 - p_0)$ for a discrete change (e.g., a reduction in class size). Given the additively separable assumption, the model fails to capture the heterogeneity that comes from μ_s, such as unobserved school characteristics or other attributes that affect all members of the school but cannot be observed in the data. Additionally, it rules out the possibility of *heterogeneous* treatment effects, which is often an important feature of the data (e.g., Heckman, Smith, and Clements (1997) and Djebbari and Smith (2008)). For example, it does not account for the possibility that the effect of the same reduction in class size could be larger for schools with better reputations.

Additionally, the linearity assumption limits other aspects of the analysis. First, regarding what features of the distribution of student achievement are considered.

Usually the analysis focuses on average outcomes; that is, the effects on the conditional expectation of Y_{ics}. Heterogeneity is most commonly accounted for by looking at subgroup impacts based on demographic characteristics. A small number of studies also look at quantile treatment effects (e.g., Jackson and Page (2013)), but the inclusion of fixed-effects is not straightforward as they can no longer be differenced out and additional assumptions are required. Second, it also limits the type of policy analysis that can be conducted. In a linear model, β measures the impact of a marginal or discrete change in P_{cs}. This might not be very informative in terms of policy implementation. For example, if we care about reallocations of individuals across groups as opposed to infeasible increases in the population. These reallocations can be characterized by obeying a particular feasibility constraint that should be accounted for. For example, one might be interested on the distributional effects of a policy that reduces gender segregation in the classroom, while keeping the total number of students of both genders fixed.

I propose a general method trying to account for these limitations. First, I use a nonseparable model of the form:

$$Y_{ics} = m\left(\mu_s, P_{cs}, Z_{ics}, \varepsilon_{ics}\right).$$

where the $m\left(\cdot\right)$ function is assumed unknown and left completely unspecified. Nonseparable models have been widely studied in the econometrics literature (e.g., Matzkin 2007; 2013). In the model I employ, no additivity or monotonicity assumptions are required for identification of certain parameters related to the effect of teacher and peer characteristics on student achievement. In addition, both μ_s and ε_{ics} can be of any dimension and interact with the covariates in general ways, in particular allowing for *heterogeneous treatment effects*. For example, for a marginal change:

$$\frac{\partial Y_{ics}}{\partial P_{cs}} = \frac{\partial m\left(p, z, \mu, \varepsilon\right)}{\partial p}.$$

Now the effect is allowed to be different even for students with the same observed characteristics. The same is true for a discrete change:

$$m\left(\mu_s, p_1, Z_{ics}, \varepsilon_{ics}\right) - m\left(\mu_s, p_0, Z_{ics}, \varepsilon_{ics}\right)$$

The method I propose goes beyond the effect of marginal and discrete changes of the covariates. In particular, I extend the analysis to consider counterfactual changes in the marginal distribution of P_{cs}, and their effect on the unconditional distribution of the outcome. For example, my method allows me to study the effects of a policy that modifies the distribution of teacher experience by reducing the number of less experienced teachers. Additionally, the impact of these counterfactual policies can be identified on any feature of the distribution of Y_{ics}. This includes, for example, the mean, quantiles and other functionals such as inequality measures.

The identification strategy is based on a control function approach that disentangles the direct effect of P_{cs} on Y_{ics} by keeping the distribution of unobservables fixed. The main concern is the possible correlation between the school fixed effects μ_s and the policy variable P_{cs}. This is handled via an exchangeability assumption (Altonji and Matzkin (2005)) on the conditional distribution of μ_s given observable class characteristics, which imposes that they cannot be ordered in a particular way in each school.

Using data from the STAR Project, I look at the influence of class size, teacher experience and gender composition of the class on test scores. My findings suggest that nonseparable heterogeneity is an important source of individual-level variation in the academic performance of kindergarten students. Using the nonseparable model, the impact of class size is considerably larger: students in smaller classes benefit

about 0.3 standard deviations, compared to a 0.16 effect obtained with a linear model. Also, teacher experience has a stronger, nonlinear impact: students assigned to more experienced teachers perform better in standardized test scores, and the gain increases with years of experience. Still, conducting a counterfactual distributional analysis I find that these gains in student performance are hard to achieve when facing resource constraints. For example, I find that a policy that reduces the size of some classes while keeping the number of students and teachers fixed generates a lower impact on test scores.

The remainder of the chapter is organized as follows. Section 2 describes the STAR Project and discusses some of the related literature. Identification, estimation and implementation of the proposed model are discussed in Section 3. Section 4 presents my empirical findings, comparing them with previous approaches. Finally, I discuss the main conclusions in Section 5. Proofs to the theorems are included in the appendix.

1.2 STAR Project

The Student/Teacher Achievement Ratio (STAR) Project was conducted in Tennessee during 1985-89. It was a large-scale, 4-year, longitudinal, experimental study of reduced class size, where students and teachers were randomly assigned to classes of different sizes. It included 79 schools from inner-city, rural, urban, and suburban locations, and over 6,000 students per grade level (for students in kindergarten and grades 1 to 3).

A large body of research has looked at the relationship between class size and student performance in nonexperimental settings, but the STAR Project was the first large-scale experiment to address this issue. In the absence of an experiment, the effect of a policy may be confounded by other observed or unobserved factors that may be correlated with the policy. In this case, the experiment only manipulated

class size and did not provide additional teacher training, new curriculum, or any other intervention.

In the original implementation of the experiment, students were to remain with the same randomly assigned class type from kindergarten through the end of the third grade. In practice, however, there were several deviations. Students who entered a participating school after the first year of the program were added to the experiment and randomly assigned to a class type. There was a substantial number of new entrants: 45 percent of eventual participants entered after kindergarten, due in part because, at the time, kindergarten was not required in Tennessee. A relatively large fraction of students exited the STAR Project schools (45 percent of overall participants) due to school moves, grade retention, or grade skipping. In addition, in response to parental concerns about fairness to students, all students in regular and regular-aide classes were randomized again in the first grade. Finally, a smaller number of students (about 10 percent of participants) were moved from one type of class to another in a nonrandom manner. Most of these moves reportedly were due to student misbehavior and not typically the result of parental requests to move their child to a small class. Still, if families felt that their child would be better served by attending smaller classes (or were upset that their child was randomly assigned to a regular class), this might yield a differential attrition rate or better attendance rate by class type. For these reasons, in this chapter I focus only on the sample of students who entered the project in kindergarten.

Ideally one would check randomization with a pretest to ensure that there are no measurable differences in the dependent variable by class type before the program began. Unfortunately, no baseline survey was collected. Still, several authors (e.g., Krueger (1999)) investigated this issue by comparing student characteristics that are related to student achievement but cannot be manipulated in response to treatment, such as student race, gender and age, finding no systematic differences in observable

characteristics across class type. Another drawback is that initial random assignment was not recorded, but instead initial enrollment was measured. This could be a concern if, for example, parents successfully lobbied for a class change in the days between class assignments and the beginning of school. Krueger (1999) presented evidence from a subset of the data suggesting that this was very unlikely. Finally, it is also important that teachers were randomly assigned. If the most effective teachers were disproportionately placed with small (or regular) classes, then the class-size effect would pick up this effect as well. Based on the data available, Krueger (1999) finds no observed within-school differences across observed characteristics of teachers, such as race, gender, experience level, or highest level of education.

In terms of external validity, there are a few aspects of the sample that may limit the validity of generalizing the STAR Project findings to other settings. In order to be eligible to participate in the program, schools were required to have a minimum-size cohort of fifty-seven students, enough to sustain both a regular and a regular-aide classroom of twenty-two students and one small class of fifteen students. As a result, the schools that participated were about 25 percent larger, on average, than other Tennessee schools. Because of requirements imposed by the legislature for geographic diversity, schools in inner cities were overrepresented, and the students included were more economically disadvantaged and more likely to be African-American than those in the state overall. Even though the percentage of non-white participants closely mirrors the percentage in the United States overall (33 versus 31 percent), there were very few Hispanic and Asian students in Tennessee at the time compared to the rest of the nation. Finally, average school spending in Tennessee was about three-fourths of the nationwide average, and teachers were less likely to have a master's degree. Krueger (1999) and Schanzenbach (2006) provide additional details on the implementation of the programs.

Numerous studies have used the STAR Project to show that class size, teacher

quality, and peer characteristics have significant (both in a statistical sense and in magnitude) causal impacts on test scores (e.g., Schanzenbach (2006)). In addition, there is a large literature about their long-term impacts. Krueger (1999) finds that, on average, performance on standardized tests increases by four percentile points the first year students attend small classes, and this advantage expands by about one percentile point per year in subsequent years. The effects are larger for minority students and those on free/reduced lunch programs. Other studies have shown that students assigned to small classes are more likely to complete high school Finn, Gerber, and Boyd-Zaharias (2005), take the SAT or ACT college entrance exams and less likely to be arrested (Krueger and Whitmore (2001)).

Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) analyze the long-term impacts of the STAR Project on college attendance, earnings, retirement savings, home ownership, and marriage by linking the original data to administrative data from tax returns.

More recently, Dynarski, Hyman, and Schanzenbach (2011) also find that students in small classes are more likely to enroll and complete college. However, very few studies look at distributional impacts beyond subgroup analysis in the STAR Project. Jackson and Page (2013) find heterogeneity across achievement quantiles, with the largest test score gains being at the top of the achievement distribution.

I contribute to this literature by providing a nonparametric, distributional evaluation of the impact of teachers, peers, and other class attributes on student performance in standardized tests. By looking at the effect of some classmate characteristics, the approach also relates to the peer effects literature, in particular to "contextual" effects models as described in Manski (1993).[1]

[1] See, e.g., Durlauf (2004) and Sacerdote (2011) for reviews, and Bramoulle, Djebbari, and Fortin (2009), Boucher, Bramoullé, Djebbari, and Fortin (2012) for recent empirical applications.

1.3 The Model

The performance on a standardized test for student i in classroom c from school s, Y_{ics}, is assumed to be generated through the nonseparable model:

$$Y_{ics} = m\left(\mu_s, P_{cs}, Z_{ics}, \varepsilon_{ics}\right) \tag{1.2}$$

for $i = 1, \cdots, I_{cs}$, $c = 1, \cdots, C_s$ and $s = 1, \cdots, S$, where (P_{cs}, Z_{ics}) is a d_x-dimensional vector of covariates, with P_{cs} a scalar that could be any classroom, teacher or peer characteristic whose effect on test scores we want to study. This flexible specification allows for general types of interaction between μ_s and P_{cs}, since no assumption is made on the functional form of $m(\cdot)$. This could be either a structural equation that describes the causal relationship between the variables, or a reduced form equation from a general structural system.

I consider several features of the relationship between test scores Y_{ics} and the class charateristic P_{cs}. First, I look at two parameters that have a straightforward interpretation and can be compared to the β coefficient from the linear model (1.1): a *Weighted Average Derivative Function* for continuous variables P_{cs}, and a *Discrete Changes Function* that evaluates P_{cs} at different points, useful for discrete random variables such as class size or teacher experience. Finally, I introduce a *Counterfactual Distribution Function* that measures the effect of general changes in the distribution of P_{cs} on the marginal distribution of Y_{ics}.

Definition I.1. When $m(\mu, p, z, \varepsilon)$ is differentiable in p and p is continuously distributed, the Local Average Response is:

$$\delta_{ics}\left(p, z\right) = \int \frac{\partial m\left(p, z, \mu, \varepsilon\right)}{\partial p} \mathbf{d}F_{\mu_s, \varepsilon_{ics}|P_{cs}, Z_{ics}}\left(\mu, \varepsilon | p, z\right) \tag{1.3}$$

where $F_{\mu_s, \varepsilon_{ics}|P_{cs}, Z_{ics}}\left(\mu, \varepsilon | p, z\right)$ is the distribution function of $(\mu_s, \varepsilon_{ics})$ conditional on

$P_{cs} = p$ and $Z_{ics} = z$.

That is, $\delta_{ics}(p, z)$ is the partial effect of P_{cs} on the expected value of Y_{ics}, evaluated at given values of P_{cs} and Z_{ics}, averaged over the distribution of unobservables. For example, this could measure the average effect on test scores of a marginal change in the gender composition of the class, when the proportion of females is 0.5. Note that, without assumptions on the dependence relationship among students, classrooms and schools, the average derivative function is indexed by (i, c, s). I will discuss this issue in more detail in Section 3.2.

One concern with the Local Average Response is that most common nonparametric estimators of (1.3) will exhibit low rates of convergence, especially when Z_{ics} is high-dimensional. Besides, in some contexts the objective of the analysis is not to predict the entire derivative curve of a conditional expectation function at each data point. Instead, we might be interested in an average version of (1.3) over all values of (P_{cs}, Z_{ics}). Then, I also consider Weighted Average Derivatives.

Definition I.2. The Weighted Average Derivative Function is

$$\delta^{\omega}_{ics} = \mathbb{E}\left[\frac{\partial m(p, z, \mu, \varepsilon)}{\partial p}\omega(p, z)\right] \tag{1.4}$$

where $\omega(p, z)$ is some specified weight function.

The weighted average derivative function is a well known parameter, and its identification and estimation have been extensively studied in the nonparametric and semiparametric literature, in part because it is possible to construct nonparametric estimators of (1.4) that attain parametric convergence rates. Certain regularity conditions are usually required on the regression functions, the data and the weights ω, such as compact support on (P_{cs}, Z_{ics}), bounded higher moments of Y_{ics} and derivatives of the $m(\cdot)$ function. See, e.g., Cattaneo, Crump, and Jansson (2010), Cattaneo, Crump, and Jansson (2013a), Cattaneo, Crump, and Jansson (2013b) and references

therein for a more detailed discussion. Also, see Newey and Stoker (1993) for efficiency results for average derivative estimators. I discuss implementations issues of (1.4) in section 3.2.

For discrete variables such as class size or years of teacher experience, we are instead interested in finite changes rather than infinitesimal ones. In this case, we can use:

Definition I.3. A Discrete Changes Function

$$\Delta_{ics}\left(p'', p'\right) = \int \left[m\left(p'', z, \mu, \varepsilon\right) - m\left(p', z, \mu, \varepsilon\right)\right] \mathbf{d}F_{Z_{ics},\mu_s,\varepsilon_{ics}|P_{cs}}\left(z, \mu, \varepsilon | p'\right) \tag{1.5}$$

is defined for a change between $P_{cs} = p''$ to $P_{cs} = p'$.

Finally, I discuss a Counterfactual Distributions Function that measures the effect of a counterfactual change in the distribution of P_{cs} on the marginal distribution of Y_{ics}. The parameter of interest in this case is:

Definition I.4. The Counterfactual Distribution Function

$$F_{Y^*_{ics}}\left(y\right) \equiv P[m(\mu_s, P^*_{cs}, Z_{ics}, \varepsilon_{ics}) \leq y] \tag{1.6}$$

is the marginal distribution of $Y^*_{ics} \equiv m(\mu_s, P^*_{cs}, Z_{ics}, \varepsilon_{ics})$ obtained by evaluating the function $m(\cdot)$ at values P^*_{cs}, where the distribution of P_{cs} changed from $F_{P_{cs}}$ to $F_{P^*_{cs}}$.

Now the research question is: how would the unconditional distribution of student performance Y_{ics} change if a policy maker could exogenously shift the values of P_{cs} to some P^*_{cs}, i.e., what is the difference between the distribution of Y_{ics} and that of the counterfactual random variable Y^*_{ics}. This new distribution can be obtained in different ways. For example, it could come from a transformation of the original random variable (such as a policy that consists of reducing the number of less experienced teachers), or from a different population (e.g. the distribution of teacher experience

from another state, different demographic groups or time periods, etc.). This type of counterfactual analysis have been extensively studied in other areas of economics (see, e.g., Fortin, Lemieux, and Firpo (2011)).

Note also that P_{cs}^{*} can be dependent or independent of P_{cs}. Both cases can be considered in the same framework, with only different implications in terms of implementation, as I discuss in the next section. Finally, the proposed approach also works for the case in which P_{cs} is different for each student, P_{ics}. For example, Dee (2004) looks at the effect of attending a class with teachers of similar characteristics as the students. Also, it does not have to consist only of classroom means (e.g. average age of peers). Glewwe (1997) points out the limitations associated with using the mean of peer characteristics without taking into account their overall distribution and how failure to do so can yield seriously misleading results. For example, $P_{ics} = \left(I^{-1}\sum_{i=1}^{I} Z_{ics}^{1-\zeta}\right)^{1-\zeta}$ accounts for other characteristics of the distribution according to the parameter ζ.

In all cases, the object of interest is the distribution $F_{Y_{ics}^{*}}$ and how it compares to $F_{Y_{ics}}$. The difference between them is called a distributional policy effect. In general, this approach can be used to conduct inference on $F_{Y_{ics}^{*}}$ as a whole, its moments and quantiles, or some functionals of it, such as inequality measures. The next section discusses the assumptions required for identification of all three parameters.

1.3.1 Identification

The main identification concern is the possible correlation between P_{cs} and μ_s. There are basically two ways in which P_{cs} affects Y_{ics}: a direct effect through the function $m(\cdot)$, and an indirect effect through the distribution of μ_s. In a linear approach, one could simply remove the effect of μ_s by differencing it out. This is no longer possible in a nonseparable model, so additional assumptions are required. In particular, I assume the existence of a vector V_s including information at the school

level, such that the school fixed effect is independent of P_{cs} once we condition on V_s. Then, we can isolate the direct effect of P_{cs} on Y_{ics}. For example, one approach would be to rely on a selection on observables type of assumption, and then construct the V_s vector with a rich set of school characteristics. Instead, I employ an exchangeability assumption that fits well in the context of the STAR Project. I develop this idea in more detail in the next section. For identification purposes, it is only required that the vector V_s satisfies:

Assumption I.5. $\mu_s \perp (Z_{ics}, P_{cs}) | V_s$

Note that V_s has the role of a control function, and there could be many choices of V_s satisfying this condition, each implying different restrictions on the model (see, e.g., Matzkin 2007; 2013 and references therein). I discuss a particular strategy to construct V_s in Section 3.1.1. The main idea is that, by controlling for V_s, I can isolated the direct effec of P_{cs} on Y_{ics} without the influence of μ_s.

The next assumption refers to the individual specific heterogeneity, ε_{ics}. In the context of the STAR Project, the random allocation of teachers and students to classroom ensures that $\varepsilon_{ics} \perp P_{cs}$. For example, let ε_{ics} represent family involvement in their children's education. Given random assignment of students and teachers into classrooms, it is expected that this student-specific characteristic is uncorrelated with the class size assigned to the student. More generally, I allow the independence of ε_{ics} and P_{cs} to be conditional:

Assumption I.6. $\varepsilon_{ics} \perp P_{cs} | (\mu_s, Z_{ics}, V_s)$

The first two assumptions are sufficient for identification of the average derivative and discrete changes functions, and have been previously proposed in a similar context by Altonji and Matzkin (2005). The result is given in Theorems 1 and 2.

Theorem I.7 (Identification of Weighted Average Derivatives)**.** *Under (A.1)-(A.2):*

$$\delta_{ics}^{\omega} = \mathbb{E}\left[\frac{\partial \mathbb{E}\left(Y_{ics}|P_{cs}=p, Z_{ics}=z, V_s=v\right)}{\partial P_{cs}}\omega(p,z)\right] \tag{1.7}$$

which also requires $\mathbb{E}[|\partial \mathbb{E}\left(Y_{ics}|P_{cs}=p, Z_{ics}=z, V_s=v\right)/\partial P_{cs}|] < \infty$.

The idea behind the identification of δ_{ics}^{ω} is straightforward. First, we calculate the partial effect of P_{cs} on Y_{ics} holding V_s constant. This holds the distribution of unobservables constant. Second, we compute the conditional distribution of $V_s|P_{cs}, Z_{ics}$ and recover $\delta_{ics}\left(p,z\right)$ by integrating out V_s. From this result, identification of the density weighted average derivatives follows directly by integrating over the joint distribution of (P_{cs}, Z_{is}). A similar identification strategy can be used for the discrete changes function:

Theorem I.8 (Identification of Discrete Changes)**.** *Under (A.1)-(A-2):*

$$\begin{aligned}\Delta_{ics}\left(p'',p'\right) &= \int \mathbb{E}\left(Y_{ics}|P_{cs}=p'', Z_{ics}=z, V_s=v\right)\mathbf{d}F_{Z_{ics},V_s|P_{cs}}\left(z,v|p'\right) \\ &- \mathbb{E}\left(Y_{ics}|P_{cs}=p'\right)\end{aligned} \tag{1.8}$$

The next two assumptions are specific to the counterfactual distribution analysis, and concern the type of distributions that can be considered for P_{cs}^{*}. A general assumption regarding the relationship between the counterfactual random variable and the unobservables is:

Assumption I.9. $\mu_s \perp \left(Z_{ics}\,, P_{cs}\,, P_{cs}^{*}\right)|V_s$ *and* $\varepsilon_{ics} \perp \left(P_{cs}, P_{cs}^{*}\right)|\left(\mu_s\,, Z_{ics}\,, V_s\right)$

This would be satisfied, for example, if P_{cs}^{*} is originated from a transformation of P_{cs}, $P_{cs}^{*} = \Gamma(P_{cs})$. Finally, I also impose a common support condition:

Assumption I.10. $sup(P_{cs}^{*}) \subseteq sup(P_{cs})$

This is required to achieve nonparametric identification due to the inability to extrapolate beyond the range observed in the data, restricting the policy experiments that can be considered to ones for which there is already some experience in the data. It could still be possible to give meaningful bounds on the counterfactual distribution when P_{cs}^{*} is allowed to take values outside of the support of P_{cs} with moderate probability. Identification follows:

Theorem I.11 (Identification of Counterfactual Distributions)**.** *Under* $(A.1)-(A.4)$*:*

$$F_{Y_{ics}^{*}}(y) = \mathbb{E}\left[F_{Y_{ics}|P_{cs},Z_{ics},V_s}(y, P_{cs}^{*}, Z_{ics}, V_s)\right] \tag{1.9}$$

That is, we can identify the unobserved marginal distribution of Y_{ics}^{*} by first computing the conditional CDF of Y_{ics} given P_{cs}, Z_{ics} and V_s. As in the previous cases, holding V_s holds the distribution of the fixed effects constant. Finally, the unconditional distribution can be obtained by integrating over the distribution of $(P_{cs}^{*}, Z_{ics}, V_s)$. Also note that, from these results, functionals such as quantiles and inequality measures are also identified.

Remark I.12. In all cases, an implicit assumption is the nonparametric identification of the regression function $\mathbb{E}(Y_{ics}|P_{cs}, Z_{ics}, V_s)$ for values of (P_{cs}, Z_{ics}, V_s) for which the conditional density of (V_s, Z_{ics}) given P_{cs} is positive. I discuss this issue in more detail in Section 3.1.1, after introducing the choice of V_s.

The methodological contribution of this chapter is to extend some previous results from the literature on distributional counterfactual effects and on nonseparable models, especially some recent contributions in a panel data context. Rothe 2010; 2012 proposes a nonparametric procedure to analyze counterfactual distributions using nonseparable models, but without accounting for group-invariant fixed effects. Recently, Chernozhukov, Fernández-Val, and Melly (2013) consider policy interventions that correspond to either changes in the distribution of covariates, or changes

in the conditional distribution of the outcome given covariates, or both. This idea also contributes to research on nonseparable models, especially to some recent work for panel data. For a review of earlier contributions in a cross-sectional context, see Matzkin 2007; 2013. One important difference is that these models usually focus on the identification of local average structural derivatives (LASD), for which additional assumptions are required. For example, monotonicity on the unobservables (e.g., Altonji and Matzkin (2005), Evdokimov (2010)). Su, Hoderlein, and White (2010) discuss several limitations of this assumption. Alternatively, other papers restrict the analysis to a subpopulation for which the covariates do not change over time (e.g., Hoderlein and White (2012), Chernozhukov, Fernández-Val, Hahn, and Newey (2013)). None of these assumptions are employed in the identification results in Theorems 1 to 3.

1.3.1.1 Exchangeability

To find a vector V_s satisfying $(A.1)$ I use the notion of exchangeability, first introduced in the nonseparable models literature by Altonji and Matzkin (2005). Graham, Imbens, and Ridder (2010) also use an exchangeability assumption but at the student level, and in a different model, to study segregation by gender in kindergarten. Without loss of generality, I assume that there are two classrooms for each school, $C = 2$. In the present context, exchangeability is defined as:

Definition I.13. The conditional distribution of μ_s given $(X_{1s}\,, X_{2s})$ is exchangeable in $(X_{1s}\,, X_{2s})$ if $F_{\mu_s|X_{1s}\,,X_{2s}}\left(\mu|x_1\,, x_2\right) = F_{\mu_s|X_{1s}\,,X_{2s}}\left(\mu|x_2\,, x_1\right)$, where $X_{cs} = (P_{cs}\,, Z_{cs})$ is a vector of classroom characteristics.

This means that the conditional distribution $F_{\mu_s|X_{1s}\,,X_{2s}}\left(\mu|x_1\,, x_2\right)$ is invariant to permutations of its arguments. That is, the subscript c is uninformative, and the information that (X_{1s}, X_{2s}) provides is independent of the order in which the elements are collected. This does not imply that there are no classroom effects, but that

classrooms cannot be ordered in a particular way for all schools. The order could be natural in other contexts, such as in panel data (if, for example, we were following classroom over time). In that context, it rules out any type of dynamic behavior. Here I assume that there are no a priori reasons for the first classroom to have a different effect than the second one on the distribution of μ_s. An important restriction is the implication that the same equality holds for any subset of the data.[2] For $C = 2$ this implies $F_{\mu_s|X_{1s}}(\mu|x_1) = F_{\mu_s|X_{2s}}(\mu|x_2)$ which means that observable characteristics of each classroom provide the same information regarding the distribution of the school fixed effects. As opposed to a conditional independence assumption where we need to find a rich enough set of variables to include in V_s such that Assumption 1 is satisfied, the exchangeability assumption can hold for any of the elements in X_{cs}.

Example I.14. Let $\mu_s \in \{H, L\}$, so schools can be either low or high quality type. Also, $X_{cs} \in \{1, 2\}$ represents years of teacher experience. One possible scenario where the exchangeability assumption would not hold is when high quality schools always assign more experienced teachers to classroom 1. Then, it might be that $P(\mu_s = H|X_{1s} = 1, X_{2s} = 2) = 0$ while $P(\mu_s = H|X_{1s} = 2, X_{2s} = 1) > 0$.

Example I.15. Suppose classrooms are numbered by the extend of external distraction (e.g., nice views out the window, external noise, broken chairs, etc). Then, teacher assignment should be invariant to these choices.

Example I.16. We can also gain some intuition by looking at types of distributional assumptions that would lead to exchangeability. Let $P_{cs} = P_s + \widetilde{P}_{cs}$ and $\mu_s = \theta P_s + \widetilde{\mu}_s$, where $P_s \backsim N(0, 1)$, $\widetilde{P}_{cs} \backsim N(0, 1)$, and $\widetilde{\mu}_s \backsim N(0, 1)$, all i.i.d. Then, $F_{\mu_s|P_{1s},P_{2s}}(\mu|p_1, p_2) = F_{\mu_s|V_s}(\mu|p_1 + p_2)$ by properties of the normal distribution.

To sum up, exchangeability is a reasonable assumption in the context of the STAR Project, where teachers and students were randomly assigned to each classroom,

[2] Note that *i.i.d.* $\Rightarrow$ *exchangeability* $\Rightarrow$ *stationarity* $\Rightarrow$ *identically distributed.*

which, even though classrooms were of different size, this is observed and can be accounted for by including class size as one of the elements of X_{cs}. This assumption can be used to construct a vector V_s satisfying $(A.1)$. First, the Fundamental Theorem of Symmetric *Polynomial* Functions states that any symmetric polynomial can be written in terms of elementary symmetric functions. Together with the Weierstrass approximation theorem, this implies that if $F_{\mu_s|X_{1s},X_{2s}}(\mu|x_1, x_2)$ is exchangeable in (X_{1s}, X_{2s}), it can be approximated arbitrarily close by a function of the form:

$$F_{\mu_s|X_{1s},X_{2s}}(\mu|x_1, x_2) = F_{\mu_s|V_s}(\mu|v) \tag{1.10}$$

where $V_s \equiv (V_s^1, V_s^2)$ are elementary symmetric polynomials of (X_{1s}, X_{2s}). For example, when X_{cs} is a scalar, $V_s = (X_{1s} + X_{2s}, X_{1s}X_{2s})$. Finally, note that (1.10) implies $(A.1)$: $\mu_s \perp X_{cs}|V_s$.

As mentioned before, an implicit assumption for the results in Theorems 1-3 is the nonparametric identification of $\mathbb{E}[Y_{ics}|P_{cs}, Z_{ics}, V_s]$ for values of (P_{cs}, Z_{ics}, V_s) for which the conditional density of (V_s, Z_{ics}) given P_{cs} is positive. This requires enough variability on $P_{cs}|Z_{ics}, V_s$. Altonji and Matzkin (2005) discuss several alternatives to guarantee this condition, but most of them require imposing additional restrictions to the model. Instead, I propose exploiting the additional variability arising from the inclusion of more elements in the vector of classroom characteristics X_{cs} (which could also be at the student level). I use the results for elementary symmetric functions for vectors developed in Weyl (1939). For example, let $X_s = (X_{1s}, X_{2s})$, with $X_{cs} = (P_{cs}, Z_{cs})$. Then, $V_s = (P_{1s}Z_{1s} + P_{2s}Z_{2s}, P_{1s}Z_{1s}P_{2s}Z_{2s})$.

1.3.2 Implementation

The estimation of all parameters of interest can be based on the identification results in Theorems 1 to 3. First, I impose assumptions on the dependence across

students, classrooms and schools.

Assumption I.17. *(a) The sequence* $\{Y_s, X_s\}_{s=1}^{S}$ *is i.i.d., where* (Y_s, X_s) *is a vector including information for all classrooms and students in school* s*:* $Y_s \equiv (Y_{1s}, \cdots, Y_{Cs})$ *and* $X_s \equiv ((P_{1s}, Z_{1s}), \cdots, (P_{Cs}, Z_{Cs}))$*, where* $Y_{cs} \equiv (Y_{1cs}, \cdots, Y_{Ics})$ *and* $Z_{cs} \equiv (Z_{1cs}, \cdots, Z_{Ics})$*. (b) Additionally, I assume that observations are identically distributed across* $i = 1, \cdots, I_{cs}$ *and* $c = 1, \cdots, C_s$*.*

Assumption 5 (b) arises naturally in a context of exchangeability of classrooms across schools, as stated in Definition 5. Then, we can omit the indexes (i, c, s) from the left hand side of (1.4), (1.5) and (1.6). To estimate the Counterfactual Distribution Function Estimator I use:

$$\widehat{F}_{Y^*}(y) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{C_s}\sum_{c=1}^{C_s}\left(\frac{1}{I_{cs}}\sum_{i=1}^{I_{cs}}\widehat{F}_{Y_{ics}|P_{cs},Z_{ics},V_s}\left(y\,,P_{cs}^{*}\,,Z_{ics}\,,V_s\right)\right)\right] \tag{1.11}$$

where $\widehat{F}_{Y_{ics}|P_{cs},Z_{ics},V_s}(y|p,z,v)$ is an estimator of the conditional distribution of Y_{ics} given $(P_{cs} = p, Z_{ics} = z, V_s = v)$. This conditional distribution function can be estimated by either a semi-parametric approach (e.g., inverting a conditional quantile model), or by fully nonparametric methods (e.g., a kernel CDF estimator). I choose a semi-parametric approach with a prominent role in empirical work: a *Distribution Regression Model*. This approach was first developed in Foresi and Paracchi (1992) and recently extended by Chernozhukov, Fernández-Val, and Melly (2013). The estimator of the conditional CDF is:

$$\widehat{F}_{Y_{ics}|P_{cs},Z_{ics},V_s}(y|p,z,v) = \Lambda\left(\rho(p,z,v)'\widehat{\theta}(y)\right) \tag{1.12}$$

where $\Lambda(\cdot)$ is a link function (such as probit or logit), and $\widehat{\theta}$ is obtained by fitting a

binary choice model of the event $1\{Y_i \leq y\}$ on $\rho(P_{cs}, Z_{ics}, V_s)$

$$\widehat{\theta}(y) = \arg\max_{b \in \mathbb{R}^{d_x+d_v}} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{i=1}^{I_{cs}} \left(1\{Y_{ics} \leq y\} \ln\left[\Lambda\left(\rho(P_{cs}, Z_{ics}, V_s)' b\right)\right]\right. \tag{1.13}$$

$$\left. +1\{Y_{ics} > y\} \ln\left[1 - \Lambda\left(\rho(P_{cs}, Z_{ics}, V_s)' b\right)\right]\right)$$

with $(P_{cs}, Z_{ics}) \in \mathbb{R}^{d_x}$, $V_s \in \mathbb{R}^{d_v}$ and $\rho(\cdot)$ is a vector of transformations (polynomials or b-splines). The distribution regression model is flexible in the sense that, for any given link function Λ, we can approximate the conditional distribution function arbitrarily well by using a rich enough $\rho(\cdot)$. It generalizes location regression by allowing the slope coefficients $\beta(y)$ to depend on the threshold index y. As opposed to other semiparametric alternatives (such as a quantile regression model), it does not require smoothness of the conditional density, since the approximation is done pointwise in the threshold y, and thus handles continuous, discrete, or mixed Y without any special adjustment (see Chernozhukov, Fernández-Val, and Melly (2013) for further details). In summary, the counterfactual distributions are estimated using the following algorithm:

Algorithm 1. *(Estimation of Counterfactual Distributions) (i) Apply the distribution regression model (1.12) to obtain estimates $\widehat{F}_{Y_{ics}|P_{cs}, Z_{ics}, V_s}$ using data on $(Y_{ics}, P_{cs}, Z_{ics}, V_s)$ for $i = 1, \cdots, I_{cs}$, $c = 1, \cdots, C_s$ and $s = 1, \cdots, S$. (ii) Compute the unconditional distribution $\widehat{F}_{Y^*}(y)$ in (1.11) by evaluating the estimator in (i) on $(y, P^*_{cs}, Z_{ics}, V_s)$ and taking the average over students, classrooms and schools.*

Next, to estimate average derivatives I employ a simple unweighted version

$$\delta = \mathbb{E}\left[\frac{\partial m(p, z, \mu, \varepsilon)}{\partial p}\right] \tag{1.14}$$

that can be compared with β in (1.1). An estimator of Average Derivatives is:

$$\widehat{\delta} = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{C_s}\sum_{c=1}^{C_s}\left(\frac{1}{I_{cs}}\sum_{i=1}^{I_{cs}}\frac{\partial\widehat{\mathbb{E}}\left(Y_{ics}|P_{cs},Z_{ics},V_s\right)}{\partial P_{cs}}\right)\right] \tag{1.15}$$

where $\partial\widehat{\mathbb{E}}\left(Y_{ics}|P_{cs},Z_{ics},V_s\right)/\partial P_{cs}$ is a nonparametric series estimators of the first derivative of the regression function $\mathbb{E}\left(Y_{ics}|P_{cs},Z_{ics},V_s\right)$ with respect to P_{cs}. Let $\tilde{X}_{ics} = (P_{cs},Z_{ics},V_s)$ and $g_0(x) = \mathbb{E}[Y_{ics}|\tilde{X}_{ics} = x]$ denote the true conditional expectation. Series methods approximate the unknown $g_0(x)$ with a flexible parametric function $g_K(x.\vartheta)$ where ϑ is an unknown coefficient vector. The integer K is the dimension of ϑ and indexes the complexity of the approximation. Let $\pi^K(x) = (\pi_{1K}(x),\cdots,\pi_{KK}(x))'$ be a vector of approximating (basis) functions having the property that a linear combination can approximate $g_0(x)$, then a *Linear Series Estimator* of $g_0(x)$ takes the form:

$$\widehat{g}(x) = \pi^K(x)'\widehat{\vartheta} \tag{1.16}$$

with $\hat{\vartheta} = (\Pi'\Pi)^{-}\Pi'Y$, where Y is the vector containing all values of Y_{ics} and Π is a vector including $\pi^K(x)$ for all values of $\tilde{X}_{ics}$. From (1.16), we can construct a series estimator of the derivative of the regression function as:

$$\frac{\widehat{\partial g(x)}}{\partial x} = \frac{\partial\pi^K(x)}{\partial x}'\widehat{\vartheta} \tag{1.17}$$

Two popular choices for series estimators are power series and splines. Let r be the dimension of x, and $\lambda = (\lambda_1,\cdots,\lambda_r)'$ a vector of nonnegative integers, i.e. a multi-index, with norm $|\lambda| = \sum_{j=1}^{r}\lambda_j$, and let $z^\lambda \equiv \prod_{j=1}^{r}(z_j)^{\lambda_j}$. For a sequence $(\lambda(k))_{l=1}^{\infty}$ of distinct such vectors, a power series approximation has $\pi_{kK}(x) = x^{\lambda(k)}$. Regression splines are linear combinations of functions that are smooth piecewise polynomials of a given order with fixed knots (joint points). For additional references on series

estimators, see, e.g., Newey (1997a), Chen (2007a), and Cattaneo and Farrell (2013a). To sum up, the Average Derivative Estimator can be implemented using the following procedure:

Algorithm 2. *(Estimation of Density Weighted Average Derivatives) (i) Estimate the derivative of the regression function* $\mathbb{E}\left(Y_{ics}|P_{cs}, Z_{ics}, V_s\right)$ *using the series estimator (1.17) with data on* $(Y_{ics}, P_{cs}, Z_{ics}, V_s)$ *for* $i = 1, \cdots, I_{cs}$, $c = 1, \cdots, C_s$ *and* $s = 1, \cdots, S$. *(ii) Compute (1.15) averaging over students, classrooms and schools.*

Finally, for the Discrete Changes Estimator:

$$\begin{aligned} \widehat{\Delta}\left(p'', p'\right) &= \int \widehat{\mathbb{E}}\left(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v\right) \widehat{f}_{Z_{ics},V_s|P_{cs}}\left(z, v|p'\right) dzdv \\ &- \widehat{\mathbb{E}}\left(Y_{ics}|P_{cs} = p'\right) \end{aligned} \tag{1.18}$$

where $\widehat{\mathbb{E}}\left(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v\right)$ and $\widehat{\mathbb{E}}\left(Y_{ics}|P_{cs} = p'\right)$ are nonparametric series estimators of the regression function, and $\widehat{f}_{Z_{ics},V_s|P_{cs}}\left(z, v|p\right)$ is a nonparametric kernel estimator for the joint density of (V_s, Z_{ics}) conditional on $P_{cs} = p$, given by:

$$\widehat{f}_{Z_{ics},V_s|P_{cs}}\left(z, v|p\right) = \frac{S^{-1}\sum_{s=1}^{S} C_s^{-1}\sum_{c=1}^{C_s} I_{cs}^{-1}\sum_{i=1}^{I_{cs}} K_{h_0}(P_{cs} - p)K_{h_1}(Z_{ics} - z)K_{h_2}(V_s - v)}{S^{-1}\sum_{s=1}^{S} C_s^{-1}\sum_{c=1}^{C_s} I_{cs}^{-1}\sum_{i=1}^{I_{cs}} K_{h_0}(P_{cs} - p)} \tag{1.19}$$

with $K_h(u) = h^{-1}K(u/h)$ and (h_0, h_1, h_2) the bandwidths associated with (P_{cs}, Z_{ics}, V_s). The bandwidths can be obtained via cross-validation methods proposed in Fan and Yim (2004) and Hall, Racine, and Li (2004). The procedure can be summarized by:

Algorithm 3. *(Estimation of Discrete Changes) (i) Use the series estimator (1.16) to estimate the regression function* $\mathbb{E}\left(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v\right)$ *using data on* $(Y_{ics}, P_{cs}, Z_{ics}, V_s)$ *for* $i = 1, \cdots, I_{cs}$, $c = 1, \cdots, C_s$ *and* $s = 1, \cdots, S$. *(ii) Estimate the conditional density of* (V_s, Z_{ics}) *given* P_{cs} *using (1.19). (iii) Integrate the conditional expectation in (i) with respect to the density in (ii) to obtain the first term in (1.8). This can be done, for example, using Monte Carlo integration. (iv) Use the series*

estimator (1.16) to estimate the regression function $\mathbb{E}(Y_{ics}|P_{cs} = p)$ *and substract it from (iii) to obtain the final estimator.*

In all cases, I construct uniform confidence bands via nonparametric bootstrap with clusters at the school level.

1.4 Empirical Results

The primary Project STAR data consist of 11,601 students who participated for at least one year. It includes students demographic information, school and class identifiers, school and teacher information, experimental condition (class type) and achievement test scores. Achievement data continued to be collected through high school. This includes achievement test scores in grades 4 to 8, teachers' ratings of student behavior in grades 4 and 8, students' self-report of school engagement and peer effects in grade 8, mathematics, science, and foreign language courses taken in high school, SAT/ACT participation and scores and graduation/dropout information. The study also collected data on 1780 students in grades 1 to 3 in 21 comparison schools, matched with STAR schools but not participating in the experiment.

Table 1 presents the summary statistics of the final sample used for the empirical analysis. It consists of 5,781 students who started the project in kindergarten and have valid information on demographic characteristics and test scores. Females constitute 48 percent of the sample, average age at the beginning of 1985 is 4.7 years, 32 percent of the students are black, and 47 percent are eligible for the free/reduced lunch program. Mean years of teacher experience is 9.2, and classes have on average 19 students.

For all the analyses conducted below, the outcome Y_{ics} consists of standardized (to have mean zero and standard deviation one) SAT scores averaged across subjects (math, reading, listening and word study skills), as is common in the literature. The

policy variables P_{cs} are class size, teacher experience and proportion of females in the classroom. In all the models, I include additional control variables Z_{ics} accounting for student gender, race, age and free/reduced lunch status. I start by comparing the results obtained using a standard, linear fixed effects panel data model with the nonparametric estimates of density weighted average derivatives (1.4) and the discrete changes estimators (1.5). Tables 3 to 6 present these results for each policy variable, and for different implementations of the nonparametric series estimators.

The empirical analysis also includes a counterfactual study of the effect of different policies related to class size, teacher experience and proportion of females in the classroom, using the Counterfactual Distribution function (1.6). Results are presented in Figures 1 to 8. For each figure, the left panel (Panel (a)) displays the original distribution of the policy variable and the resulting counterfactual change. The right one (Panel (b)) reports *Quantile Treatment Effects* (QTE) for that policy, together with uniformly valid confidence intervals. The QTE estimator measures the impact of the counterfactual policy for quantile q as the difference in outcomes between the $q-$th student in the countefactual (treatment) distribution and the $q-$th student in the original (control) one. For instance, we can compare the median test score for the students in the original distribution and subtract from it the median test score for the students under the counterfactual policy to estimate the effect at the median of the achievement distribution. Note that this estimator will not identify the impact of the policy on a particular student who would have been at the $q-$th percentile in the absence of the policy. This interpretation is only appropriate if the policy causes no re-ordering of achievement ranks within the distribution. As discussed in Heckman, Smith, and Clements (1997) and more recently in Djebbari and Smith (2008), quantile treatment effects are simply differences between the treatment and control distributions, and recovering quantiles of the treatment effect distribution requires specific assumptions about the joint distribution of outcomes in the treatment

and control states (such as perfect positive or perfect negative dependence). Nevertheless, the QTE estimator provides substantial information about treatment effects heterogeneity.

1.4.1 Class Size

Empirical analyses in the STAR Project usually conclude that smaller classes increase student achievement, even after controlling for school fixed effects and teacher characteristics. Table 3 presents estimates of the effect of moving from a class size of 22 to 15 students (the median class sizes for regular and small classrooms in the STAR Project, respectively). In the linear model, this effect is simply $\widehat{\beta}(22-15)$, where β is the coefficient associated with class size in the linear model (1.1). Instead, the estimated effect using the discrete changes estimator (1.5) is:

$$\begin{aligned}\widehat{\Delta}(15,22) &= \int \widehat{\mathbb{E}}(Y_{ics}|P_{cs}=15,Z_{ics}=z,V_s=v)\,\widehat{f}_{Z_{ics},V_s|P_{cs}}(z,v|22)\,dzdv \\ &- \widehat{\mathbb{E}}(Y_{ics}|P_{cs}=22)\end{aligned}$$

Using a fixed-effects linear panel data model (Column 1), I find that students benefit about 0.16 standard deviations from assignment to a small class. This is in line with previous findings. However, the nonparametric estimates are actually larger and statistically significant for all the specifications. For example, the effect of assignment to a small class is between 0.3 and 0.43 standard deviations using power series estimators of the regression function. This suggests that unobserved heterogeneity at the school level plays an important role on the impact of class size on student performance. In turn, it could help explain previous findings of different impact estimates for demographic groups, as in Schanzenbach (2006). Still, now the effect is more general since unobservable factor are also accounted for. For instance, it is possible that the positive effect of a smaller class size is larger in a school with a better management.

Next, I extend the analysis by looking at *distributional effects* of class size policies using the counterfactual distribution estimator (1.6). The goal here is to see what policies regarding class size would be able to generate the gains in students' performance obtained in the previous analysis. I start with a policy that simply reduces class size in the largest classroom (*Policy 1*):

$$P_{cs}^{*} = \begin{cases} P_{cs} & \text{if } P_{cs} \leq 21 \\ P_{cs} - 5 & \text{if } P_{cs} > 21 \end{cases}$$

The QTE results are presented in Figure 1.1. The effect is positive throughout the achievement distribution, but heterogeneous, with the biggest impacts among children with scores near the top of the distribution. For example, the test score of a student at the 90^{th} percentile in the counterfactual distribution is almost a third of a standard deviation higher than the test score of a 90^{th} percentile student in the original distribution, whereas the difference at the 10^{th} percentile of the distribution are less than a tenth of a standard deviation. These estimates are in line with, although lower in magnitude, the estimates comparing small versus large class sizes in Jackson and Page (2013). High achievers could benefit more from smaller classes if, for instance, teachers in small classes are better able to identify high achievers and use instructional approaches that work well for them.

One potential concern with the previous policy is that it does not take into account feasibility or resource constraints. For example, in order to reduce class size according to *Policy 1*, the school would need to hire additional teachers or to enroll some students in additional classrooms. For this reason, I also look at *Policy 2* which keeps the number of students fixed by constructing the counterfactual variable as:

$$P_{cs}^{*} = \begin{cases} P_{cs} + 5 & \text{if } P_{cs} \leq 21 \\ P_{cs} - 5 & \text{if } P_{cs} > 21 \end{cases}$$

From the results in Figure 1.2 we can see that the QTE estimates are close to zero and statistically insignificant over the achievement distribution. This can be explained by the improvements in performance by the students in smaller classes being compensated by the worsening in performance by those in larger classes.

Overall, I conclude that the estimated impacts of class size are larger when using a nonseparable model, highlighting the relevance of accounting for heterogeneous treatment effects. The distributional analysis of counterfactual changes also suggests that the impacts are much smaller once we take into account feasibility constrains.

1.4.2 Teacher Experience

Teacher experience has traditionally been an important component of teacher policies in the U.S. school systems. Although recent debates have focused on the development and use of more direct measures of teacher performance (e.g., value-added models, standards-based evaluation), teacher experience continues to play a dominant role in most human resource policies. The underlying assumption is that experience promotes effectiveness and that experience gained over time enhances the knowledge, skills, and productivity of teachers.

Experience is among the most commonly studied teacher characteristic. Several studies find that the impact of experience is strongest during the first few years of teaching: on average, brand-new teachers are less effective than those with some experience (e.g., Rockoff (2004), Rivkin, Hanushek, and Kain (2005), Clotfelter, Ladd, and Vigdor 2007; 2010, Kane, Rockoff, and Staiger (2008), Harris and Sass (2011)), but the greatest productivity gains occur during their first few years on the job, after which their performance tends to level off. Empirical evidence suggests that, on average, students with teachers in their fifth year of teaching score between 5 and 15 percent of a standard deviation higher than students with teachers in their first year on the job (Atteberry, Loeb, and Wyckoff (2013)). There is also evidence that this

effect is stronger than the effects of teacher licensure tests scores, and even class size (e.g., Clotfelter, Ladd, and Vigdor (2007), Rivkin, Hanushek, and Kain (2005)).

In the STAR Project, Krueger (1999) finds small but positive effects of teacher experience, with a peak at about twenty years: students in classes where the teacher has twenty years of experience tend to score about three percentile points higher than those in classes where the teacher has zero experience, all else being equal. As a whole, however, he concludes that measured teacher characteristics explain relatively little of student achievement on standardized tests. More recently, Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) finds that students randomly assigned to more experienced kindergarten teachers have higher test scores, with the effect being roughly linear. Schanzenbach (2006) analyze the indirect effect of teacher experience by comparing the performance of students in small versus regular class size with teachers of different experience, finding considerable heterogeneity of impacts: students with more experienced teachers show large, stat other observable teacher characteristics such as advanced degrees,istically significant gains from reduced class size. In contrast, students who have a teacher with fewer than five years of experience show smaller and often not statistically significant gains from small classes. Recently, Mueller (2013) finds that this pattern exists at all deciles of the achievement distribution, but is less pronounced at lower deciles.

In Tables 4 and 5, I compare the estimates from a linear panel data model (with a quadratic term for the experience variable) to the discrete changes estimator. The goal is to study nonlinear effects of teacher experience on student performance by comparing students with teachers of different years of experience. First, I look at a change from 5 to 10 years of experience (Table 4). Then, I consider a change from 10

to 15 years in Table 5. Using the nonseparable model (1.2), the effects are:

$$\begin{aligned}
\widehat{\Delta}(10,05) &= \int \widehat{\mathbb{E}}(Y_{ics}|P_{cs}=10, Z_{ics}=z, V_s=v)\, \widehat{f}_{Z_{ics},V_s|P_{cs}}(z,v|05)\, dzdv \\
&- \widehat{\mathbb{E}}(Y_{ics}|P_{cs}=05) \\
\widehat{\Delta}(15,10) &= \int \widehat{\mathbb{E}}(Y_{ics}|P_{cs}=15, Z_{ics}=z, V_s=v)\, \widehat{f}_{Z_{ics},V_s|P_{cs}}(z,v|10)\, dzdv \\
&- \widehat{\mathbb{E}}(Y_{ics}|P_{cs}=10)
\end{aligned}$$

From Table 4, we can see that the estimate for $\Delta(10,05)$ is around 0.13. That is, students with a teacher with 10 years of experience perform 0.13 standard deviations higher than those with a teacher with only 5 years of experience. We can also see that the point estimates are precisely estimated. This effect is considerably larger than the one obtained using a quadratic model (0.027). Also, in Table 5, changing teacher experience from 10 to 15 years yields a larger impact (between 0.23 and 0.55), which is also larger than the one obtained from a quadratic model (0.063). That is, using the nonseparable model we find evidence of a strong and nonlinear effect of teacher experience. Again, this points to the importance of accounting for unobserved factors in the impact of teacher experience on student performance.

Finally, I look at *distributional effects*. *Policy 1* consists of a general increase of five years in teacher experience,

$$P_{cs}^{*} = P_{cs} + 5$$

From Figure 1.3, we can see that the effect is positive for all percentiles, but is slightly larger for those at the top quantiles. More importantly, the magnitude of the effect is considerably lower than what is obtained with the discrete changes estimator. To examine whether this could be due to a differential effect coming from teachers of different experience, the next two policies look at the differential effect of teacher

experience for relatively new versus more experienced teachers. *Policy 2* only affects classrooms with less experienced teachers:

$$P_{cs}^{*} = \begin{cases} P_{cs} & \text{if } P_{cs} > 5 \\ P_{cs} + 5 & \text{if } P_{cs} \leq 5 \end{cases}$$

Results are presented in Figure 1.4. The effect of Policy 2 is roughly constant over the achievement distribution. Overall, as with class size, the impacts of these policies are smaller than those obtained with the discrete changes estimators. This could be due to, for example, the impact of teacher experience coming from their interaction with other classrooms characteristics, such as class size (Mueller (2013)).

1.4.3 Proportion of Females

The idea that peers can affect student achievement is based on the assumption that students do not only learn from teachers but also from classmates. For example, students might teach one another by working in groups or having casual discussions, generating knowledge spillovers (see, e.g., Sacerdote (2011) for a review of this literature). One aspect of particular relevance in this context is the gender composition of the classroom. For example, the study of gender peer effects can shed light on the debate single-sex versus coeducational schools (Whitmore (2005)). Gender composition of the classroom could affect student performance in many ways. For example, a higher proportion of girls could improve classroom behavior, reduce classroom disruption and affect the level of violence, creating a better atmosphere for learning (Lavy and Schlosser (2011)). The presence of boys could intimidate girls from speaking up and influence student self-concepts or affect engagement with certain subjects. Finally, classroom composition could also affect the attitude and expectations of teachers towards the class, influencing the pace of teaching or their instructional methods (Cunningham and Andrews (1988)).

Several studies have examined the empirical role of gender composition of the classroom. Hoxby (2000) exploits gender variation in cohort composition in Texas elementary schools and finds that a higher share of girls raises student achievement in math and reading, both for boys and girls. Lavy and Schlosser (2011) find that, in Israeli middle-schools, a 10 percent point increase in the proportion of female students increases girls' math test scores by 3.7 percent of a standard deviation and boys' scores by 2.2 percent.

Let P_{cs} be the proportion of females in classroom c in school s. Table 6 compares the fixed effect estimator of β from model (1.1) to the density weighted average derivative estimator (1.3), for different choices of the series estimator of the derivative of the regression function. We can see that the impacts are considerably larger and statistically significant when we use the nonseparable model.

Next, I look at *distributional impacts*. First, *Policy 1* implies a 10 percent increase in the proportion of females for all classrooms,

$$P^*_{cs} = (1 + 0.1) \times P_{cs}$$

From the results in Figure 1.5, we can see that the effect of this policy is positive for all quantiles. There is some heterogeneity (with larger point estimates at the top of the distribution), but with wide confidence intervals. Overall, the impacts are smaller than those obtained in Table 6.

The next policy try to disentangle the mechanism behind the positive effect of the proportion of females on student performance. For example, the effect could be coming from either having more girls in the classroom or more students of the same gender. Then, *Policy 2* increases the proportion of females in a classroom with majority of girls, and decreases the proportion in the classrooms with majority of

boys:

$$P_{cs}^{*} = \begin{cases} (1+0.1) \times P_{cs} & \text{if } P_{cs} > 0.5 \\ (1-0.1) \times P_{cs} & \text{if } P_{cs} \leq 0.5 \end{cases}$$

From Figure 1.6 we can see that the impacts are now close to zero for all quantiles, suggesting that the effect is actually coming from a larger proportion of females, in line with previous findings in the literature. As with class size, imposing feasibility constraints affects the magnitude of the impacts, suggesting that the implementation of policies regarding gender composition of the classrooms should take into account additional interactions and explore additional channels through which gender peer effects influence student performance.

1.4.4 Tables and graphs

Table 1.1: Summary Statistics - Students

Variable	Mean	Std. Dev.	N
Age	4.71	0.34	5,847
Race (Black)	0.32	0.46	5,847
Female	0.48	0.51	5,847
Free Lunch Eligible	0.48	0.52	5,847
Rural School	0.46	0.49	5,847
Total Math Score SAT	485	47.7	5,844
Total Reading Score SAT	436	31.7	5,763

Table 1.2: Summary Statistics - Classrooms

Variable	Mean	Std. Dev.	N
Teacher Race (Black)	0.16	0.36	302
Teacher has Master Degree	0.36	0.48	302
Teacher Experience (years)	9.32	5.75	302
Class Size	19.4	4.14	302

Note: Original Sample Size: 6325, Sample with non-missing score information: 5907, Sample with non-missing values of class size, teacher experience and gender: 5886. The final sample excludes those with missing values in any of the covariates.

Table 1.3: Class Size

	OLS	**Power Series**			**Regression Splines**		
	Fixed Effects	(1)	(2)	(3)	(1)	(2)	(3)
Coefficient	0.167	0.296	0.399	0.432	0.226	0.224	0.321
Std. Error	(0.028)	(0.043)	(0.063)	(0.065)	(0.064)	(0.064)	(0.043)
Parameter	-	$K = 2$	$K = 4$	$K = 6$	TP 1	TP 2	ThP

Table 1.4: Teacher Experience - 5 to 10 years

	OLS	**Power Series**			**Regression Splines**		
	Fixed Effects	(1)	(2)	(3)	(1)	(2)	(3)
Coefficient	0.027	0.125	0.160	0.127	0.137	0.138	0.135
Std. Error	(0.012)	(0.022)	(0.035)	(0.045)	(0.040)	(0.040)	(0.022)
Parameter	-	$K = 2$	$K = 4$	$K = 6$	TP 1	TP 2	ThP

Table 1.5: Teacher Experience - 10 to 15 years

	OLS	**Power Series**			**Regression Splines**		
	Fixed Effects	(1)	(2)	(3)	(1)	(2)	(3)
Coefficient	0.063	0.228	0.390	0.554	0.339	0.335	0.247
Std. Error	(0.013)	(0.025)	(0.041)	(0.062)	(0.054)	(0.053)	(0.031)
Parameter	-	$K = 2$	$K = 4$	$K = 6$	TP 1	TP 2	ThP

Table 1.6: Proportion of Females

	OLS	**Power Series**		
	Fixed Effects	(1)	(2)	(3)
Coefficient	0.376	0.580	0.555	0.592
Std. Error	(0.013)	(0.122)	(0.126)	(0.129)
Parameter	-	$K = 2$	$K = 4$	$K = 6$

Notes: Regression Splines obtained using *mgcv* R-package

1.5 Conclusion

In this chapter, I look at the effects of teacher and peer characteristics on student achievement in the STAR Project conducted in Tennessee in the late 80s. As in standard linear models, I consider two types of unobservables: school-specific effects and idiosyncratic disturbances. The model generalizes previous empirical research

by allowing both effects to enter the structural function nonseparably. In particular, no functional form assumptions are needed for identification. Thus, the model permits nonparametric distributional and counterfactual analysis of heterogeneous effects. The main identification result uses an exchangeability assumption on the way that covariates affect the distribution of the school fixed effects. The model also extends policy analysis beyond marginal or discrete changes, to consider distributional effects originating from a counterfactual change in the distribution of characteristics of classrooms, peers and teachers. These impacts can also be analyzed on any feature of the distribution of student achievement, such as quantiles and inequality measures. In the empirical analysis, I look at the effects of class size, teacher experience and gender composition of the classroom on student test scores. Findings suggest that nonseparable heterogeneity is an important source of individual-level variation in the academic performance of kindergarten students in the STAR Project. Compared to previous results, the impact of class size is larger in magnitude and teacher experience has a stronger nonlinear impact. Still, conducting a counterfactual distributional analysis I find that these gains in student performance are hard to achieve when facing resource constraints.

Figure 1.1: Class Size - Policy I

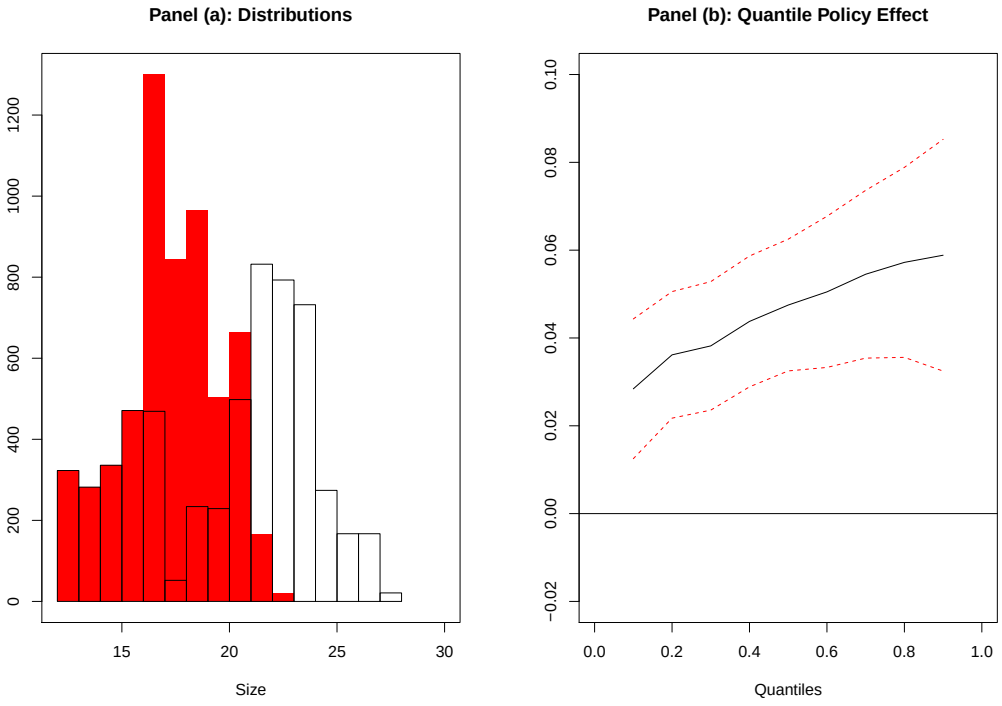


Figure 1.2: Class Size - Policy II

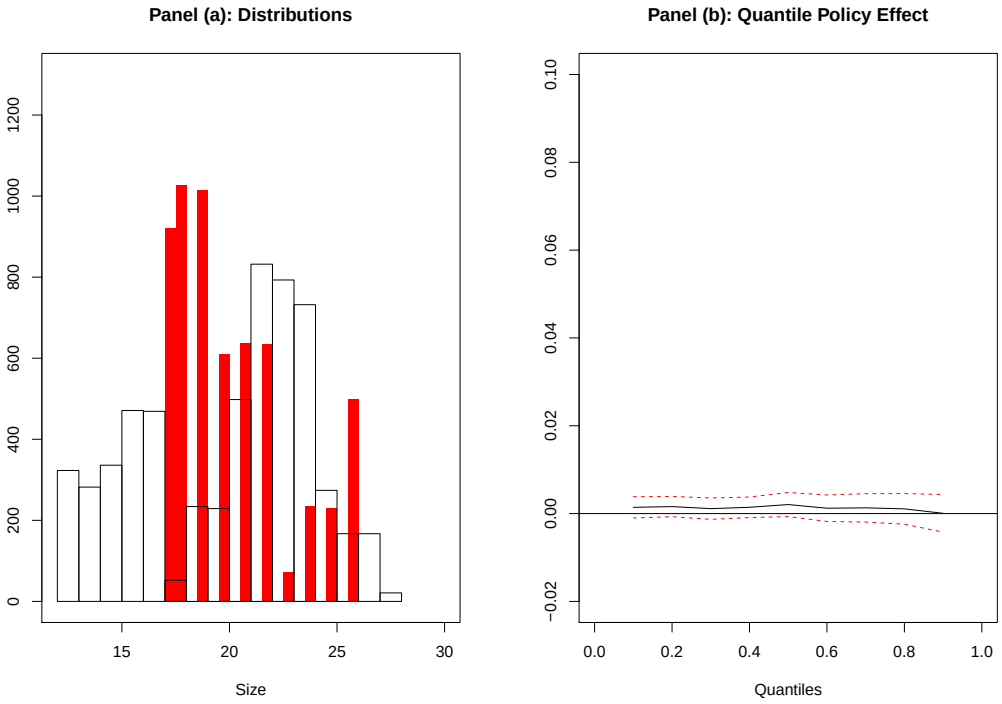


Figure 1.3: Teacher Experience - Policy I

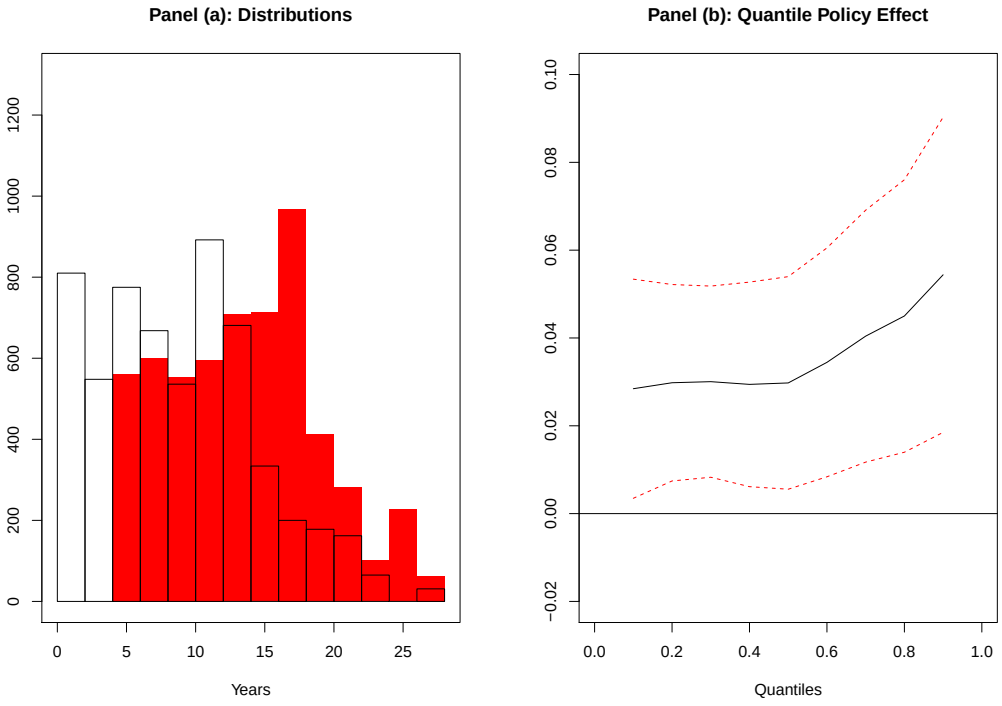


Figure 1.4: Teacher Experience - Policy II

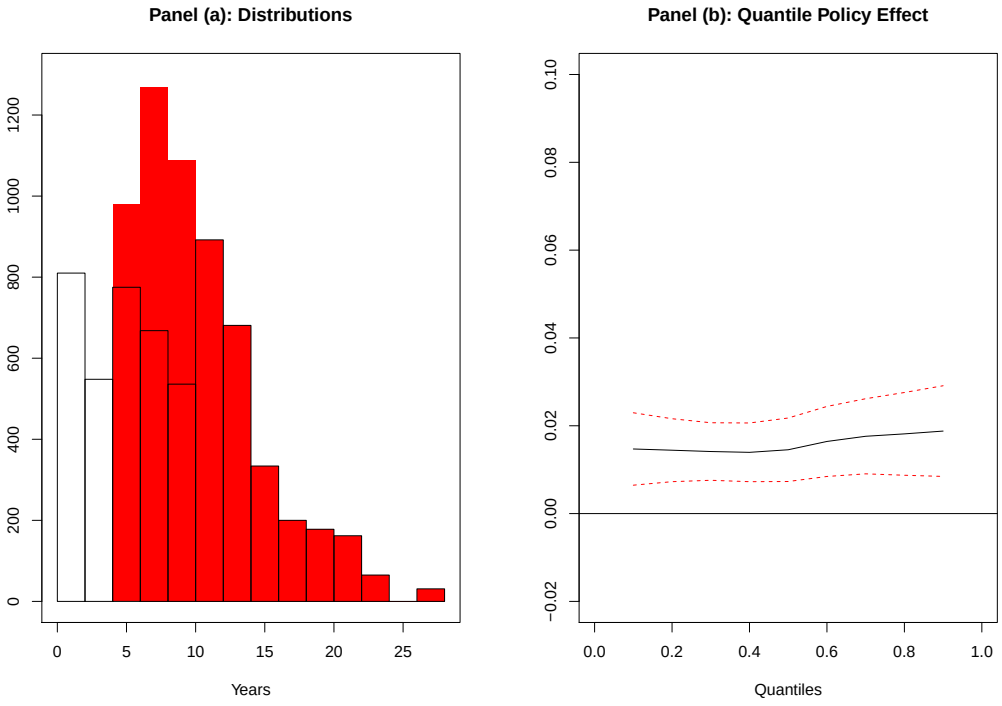


Figure 1.5: Proportion of Females - Policy I

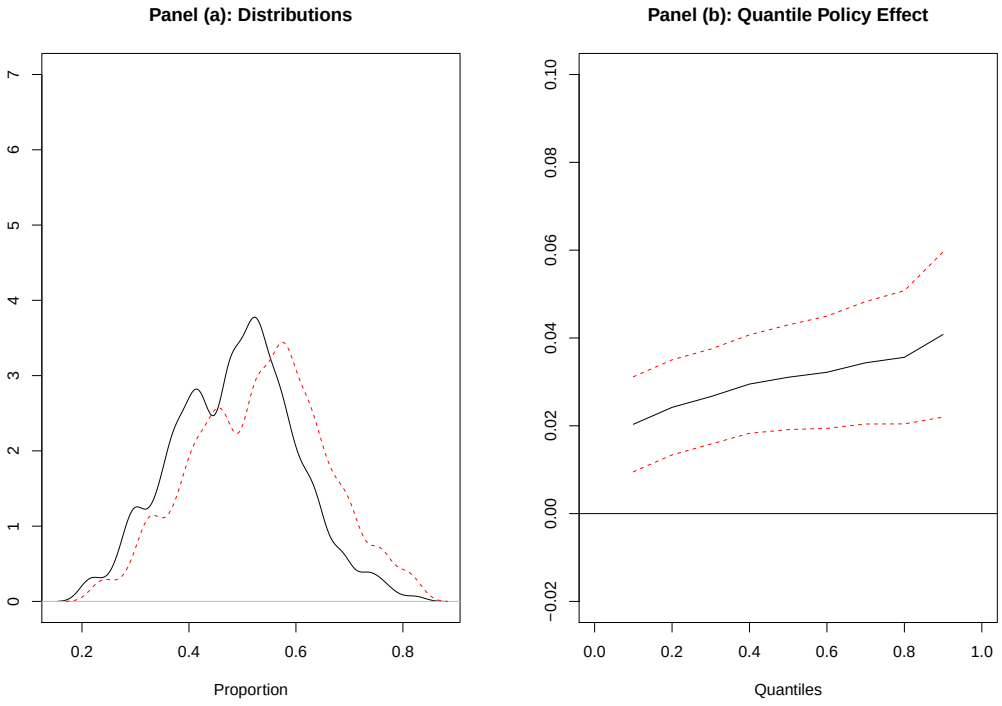
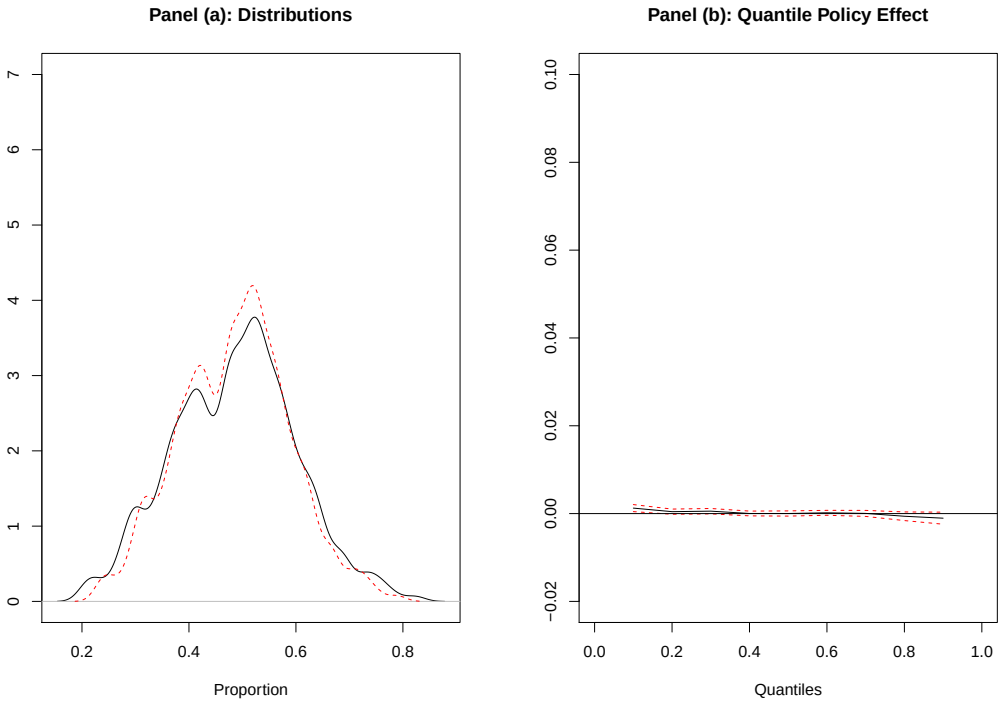


Figure 1.6: Proportion of Females - Policy II



CHAPTER II

Robust Nonparametric Confidence Intervals for Regression- Discontinuity Designs

2.1 Introduction

The regression discontinuity (RD) design has become one of the leading quasi-experimental empirical strategies in economics, political science, education and many other social and behavioral sciences (see van der Klaauw (2008), Imbens and Lemieux (2008), Lee and Lemieux (2010) and Dinardo and Lee (2011) for reviews). In this design, units are assigned to treatment based on their value of an observed covariate (also known as score or running variable), with the probability of treatment assignment jumping discontinuously at a known cutoff. For example, in its original application, Thistlethwaite and Campbell (1960) used this design to study the effects of receiving an award on future academic achievement, where the award was given to students whose test scores were above a cutoff. The idea of the RD design is to study the effects of the treatment using only observations near the cutoff to control for smoothly varying unobserved confounders. In the simplest case, flexible estimation of RD treatment effects approximates the regression function of the outcome given the score near the cutoff for control and treated groups separately, and computes the estimated effect as the difference of the values of the regression functions at the cutoff

for each group.

Nonparametric local polynomial estimators have received great attention in the recent RD literature, becoming the standard choice for estimation of RD treatment effects. This estimation strategy involves approximating the regression functions above and below the cutoff by means of weighted polynomial regressions, typically of order one or two, with weights computed by applying a kernel function on the distance of each observation's score to the cutoff. These kernel-based estimators require a choice of bandwidth for implementation, and several bandwidth selectors are now available in the literature. These bandwidth selectors are obtained by balancing squared-bias and variance of the RD estimator, a procedure that typically leads to bandwidth choices that are too "large" to ensure the validity of the distributional approximations usually invoked; that is, these bandwidth selectors lead to a non-negligible bias in the distributional approximation of the estimator. As a consequence, the resulting confidence intervals for RD treatment effects may be biased, having empirical coverage well below their nominal target. This implies that conventional confidence intervals may substantially over-reject the null hypothesis of no treatment effect.

To address this drawback in conventional RD inference, we propose new confidence intervals for RD treatment effects that offer robustness to "large" bandwidths such as those usually obtained from cross-validation or asymptotic mean-square-error minimization. [1] Our proposed confidence intervals are constructed as follows. We first bias-correct the RD estimator to account for the effect of a "large" bandwidth choice; that is, we recenter the usual t-statistic with an estimate of the leading bias. As it is well-known, however, conventional bias-correction alone delivers very poor finite-sample performance because it relies on a low-quality distributional approximation. Thus, in order to improve the quality of the distributional approximation

[1] For example, for the local-linear RD estimator, "small" and "large" bandwidths refer, respectively, to $nh_n^5 \to 0$ and $nh_n^5 \not\to 0$ (e.g., $nh_n^5 \to c \in \mathbb{R}_{++}$), where h_n is the bandwidth and n is the sample size. Section 2.2 discusses this case in detail, while the general case is given in the appendix.

of the bias-corrected t-statistic, we rescale it with a novel standard error formula that accounts for the additional variability introduced by the estimated bias. The new standardization is theoretically justified by a non-standard large-sample distributional approximation of the bias-corrected estimator, which explicitly accounts for the potential contribution that bias-correction may add to the finite-sample variability of the usual t-statistic.

Altogether, our proposed confidence intervals are demonstrably more robust to the bandwidth choice ("small" or "large"), as they are not only valid when the usual bandwidth conditions are satisfied (being asymptotically equivalent to the conventional confidence intervals in this case), but also continue to offer correct coverage rates in large samples even when the conventional confidence intervals do not (see Remarks II.5 and II.6 below). These properties are illustrated with an empirically motivated simulation study, which shows that our proposed data-driven confidence intervals exhibit close-to-correct empirical coverage and good empirical interval length on average.

Our discussion focuses on the construction of robust confidence intervals for the RD average treatment effect at the cutoff in four settings: sharp RD, sharp kink RD, fuzzy RD and fuzzy kink RD designs. These are special cases of our main theorems given in the appendix. In all cases, the bias-correction technique follows the standard approach in the nonparametrics literature (e.g., (Fan and Gijbels, 1996, Section 4.4, p. 116)), but our standard error formulas are different because they incorporate additional terms not present in the conventional formulas currently used in practice. The resulting confidence intervals allow for mean-square optimal bandwidth selectors and, more generally, enjoy demonstrable improvements in terms of allowed bandwidth sequences, coverage error rates and, in some cases, interval length (see Remarks II.5, II.7 and II.8 below). As a particular case, our results also justify confidence intervals estimators based on a local polynomial estimator of an order higher than the order

of the polynomial used for point estimation, a procedure that is easy to implement in applications (see Remark II.10 below). The new confidence intervals may be used both for inference on treatment effects (when the outcome of interest is used as an outcome in the estimation) as well as for falsification tests that look for null effects (when pretreatment or "placebo" covariates are used as outcomes in the estimation).

This chapter contributes to the emerging methodological literature on RD designs. See Hahn, Todd, and van der Klaauw (2001) and Lee (2008) for identification results, Porter (2003) for optimality results of local polynomial estimators, McCrary (2008) for specification testing, Lee and Card (2008) for inference with discrete running variables, Imbens and Kalyanaraman (2012) for bandwidth selection procedures for local-linear estimators, Otsu and Xu (2013) for empirical likelihood methods, Frandsen, Frölich, and Melly (2012) for quantile treatment effects, Card, Lee, Pei, and Weber (2012), Dong (2012) and Dong and Lewbel (2012) for kink RD designs, Marmer, Feir, and Lemieux (2012) for weak-IV robust inference in fuzzy RD designs, Cattaneo, Frandsen, and Titiunik (2014) for randomization-inference methods, and Calonico, Cattaneo, and Titiunik (2014a) for RD plots. More broadly, our results also contribute to the literature on asymptotic approximations for nonparametric local polynomial estimators (Fan and Gijbels (1996)), which are useful in econometrics (Ichimura and Todd (2007)) – see Remark II.11 and Calonico, Cattaneo, and Farrell (2014) for further discussion.

The rest of the chapter is organized as follows. Section 2.2 describes the sharp RD design, reviews conventional results and outlines our proposed robust confidence intervals. Section 2.3 discusses extensions to kink RD, fuzzy RD and fuzzy kink RD designs. Mean-square-error optimal bandwidths and their validity are examined in Section 2.4, while valid standard-errors are discussed in Section 2.5. Section 2.6 presents our simulation study, and Section 3.6 concludes. In the appendix, we summarize our general theoretical results, including extensions to arbitrary polynomial

orders, higher-order derivatives and related results for bandwidth selection, while in the online supplemental appendix (Calonico, Cattaneo, and Titiunik (2014d)) we collect the main mathematical proofs, other methodological and technical results, additional simulation evidence, and an empirical illustration employing household data from Progresa/Oportunidades. Companion `R` and `STATA` software packages are described in Calonico, Cattaneo and Titiunik 2014e; 2014b.

2.2 Sharp RD Design

In the canonical sharp RD design, $(Y_i(0), Y_i(1), X_i)'$, $i = 1, 2, \ldots, n$, is a random sample with $f(x)$ the Lebesgue density of X_i. Given a known threshold $\bar{x}$, set to $\bar{x} = 0$ without loss of generality, the observed score or forcing variable X_i determines whether unit i is assigned treatment ($X_i \geq 0$) or not ($X_i < 0$), while the random variables $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes with and without treatment, respectively. The observed random sample is $(Y_i, X_i)'$, $i = 1, 2, \ldots, n$, where $Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i$ with $T_i = \mathbf{1}(X_i \geq 0)$ and $\mathbf{1}(\cdot)$ is the indicator function.

The parameter of interest is $\tau_{\texttt{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = \bar{x}]$, the average treatment effect at the threshold. Under a mild continuity condition, Hahn, Todd, and van der Klaauw (2001) show that this parameter is nonparametrically identifiable as the difference of two conditional expectations evaluated at the (induced) boundary point $\bar{x} = 0$:

$$\tau_{\texttt{SRD}} = \mu_+ - \mu_-, \qquad \mu_+ = \lim_{x \to 0^+} \mu(x), \qquad \mu_- = \lim_{x \to 0^-} \mu(x), \qquad \mu(x) = \mathbb{E}[Y_i|X_i = x].$$

Throughout the chapter, we drop the evaluation point of functions whenever possible to simplify notation. Estimation in RD designs naturally focuses on flexible approximation, near the cutoff $\bar{x} = 0$, of the regression functions $\mu_-(x) = \mathbb{E}[Y_i(0)|X_i = x]$ (from the left) and $\mu_+(x) = \mathbb{E}[Y_i(1)|X_i = x]$ (from the right). We employ the following

assumption on the basic sharp RD model.

Assumption II.1. *For $\kappa_0 > 0$, the following holds in the neighborhood $(-\kappa_0, \kappa_0)$ around the cutoff $\bar{x} = 0$:*

(a) $\mathbb{E}[Y_i^4|X_i = x]$ is bounded, and $f(x)$ is continuous and bounded away from zero.

(b) $\mu_-(x) = \mathbb{E}[Y_i(0)|X_i = x]$ and $\mu_+(x) = \mathbb{E}[Y_i(1)|X_i = x]$ are S-times continuously differentiable.

(c) $\sigma_-^2(x) = \mathbb{V}[Y_i(0)|X_i = x]$ and $\sigma_+^2(x) = \mathbb{V}[Y_i(1)|X_i = x]$ are continuous and bounded away from zero.

Part (a) in Assumption III.1 imposes existence of moments, requires that the running variable X_i be continuously distributed near the cutoff, and ensures the presence of observations arbitrarily close to the cutoff in large samples. Part (b) imposes standard smoothness conditions on the underlying regression functions, which is the key ingredient used to control the leading biases of the RD estimators considered in this chapter. Part (c) puts standard restrictions on the conditional variance of the observed outcome, which may be different at either side of the threshold. We set $\sigma_+^2 = \lim_{x\to 0^+} \sigma^2(x)$ and $\sigma_-^2 = \lim_{x\to 0^-} \sigma^2(x)$, where $\sigma^2(x) = \mathbb{V}[Y_i|X_i = x]$. Higher-order derivatives of the unknown regression functions are denoted by $\mu_+^{(\nu)}(x) = \mathrm{d}^\nu \mu_+(x)/\mathrm{d}x^\nu$ and $\mu_+^{(\nu)}(x) = \mathrm{d}^\nu \mu_+(x)/\mathrm{d}x^\nu$, for $\nu < S$ (with S in Assumption III.1(b)). We also set $\mu_+^{(\nu)} = \lim_{x\to 0^+} \mu_+^{(\nu)}(x)$ and $\mu_-^{(\nu)} = \lim_{x\to 0^-} \mu_-^{(\nu)}(x)$; by definition, $\mu_+ = \mu_+^{(0)}$ and $\mu_- = \mu_-^{(0)}$.

Remark II.2 (Discrete running variable). Assumption III.1(a) rules out discrete-valued running variables. In applications where X_i exhibits many mass points near the cutoff, this assumption may still give a good approximation and our results might be used in practice. However, when X_i exhibits few mass points, our results do not apply directly without further assumptions and modifications, and other assumptions and inference approaches may be more appropriate; e.g., Cattaneo, Frandsen, and Titiunik (2014).

Throughout the chapter, we employ local polynomial regression estimators of various orders to approximate unknown regression functions (Fan and Gijbels (1996)). These estimators are particularly well-suited for inference in the RD design because of their excellent boundary properties (Cheng, Fan, and Marron (1997)). Section B.1.1 in the appendix describes these estimators in full generality and introduces detailed notation not employed in the main text to ease the exposition. We impose the following assumption on the kernel function employed to construct these estimators.

Assumption II.3. *For some $\kappa > 0$, the kernel function $k(\cdot) : [0, \kappa] \mapsto \mathbb{R}$ is bounded and nonnegative, zero outside its support, and positive and continuous on $(0, \kappa)$.*

Assumption II.3 permits all kernels commonly used in empirical work, including the triangular kernel $k(u) = (1 - u)\mathbf{1}(0 \leq u \leq 1)$ and the uniform kernel $k(u) = \mathbf{1}(0 \leq u \leq 1)$. Our results apply when different kernels are used on either side of the threshold, but we set $K(u) = k(-u) \cdot \mathbf{1}(u < 0) + k(u) \cdot \mathbf{1}(u \geq 0)$ for concreteness. This implies that, for $\kappa > 0$ in Assumption II.3, $K(\cdot)$ is symmetric, bounded and nonnegative on $[-\kappa, \kappa]$, zero otherwise, and positive and continuous on $(-\kappa, \kappa)$. For simplicity, we employ the same kernel function $k(\cdot)$ to form all estimators in the chapter.

2.2.1 Robust Local-Linear Confidence Intervals

Following Hahn, Todd, and van der Klaauw (2001) and Porter (2003), we consider confidence intervals based on the popular local-linear estimator of $\tau_{\texttt{SRD}}$, which is the difference in intercepts of two first-order local polynomial estimators, one from each side of the threshold. Formally, for a positive bandwidth h_n,

$$\hat{\tau}_{\texttt{SRD}}(h_n) = \hat{\mu}_{+,1}(h_n) - \hat{\mu}_{-,1}(h_n),$$

$$(\hat{\mu}_{+,1}(h_n), \hat{\mu}^{(1)}_{+,1}(h_n))' = \arg\min_{b_0, b_1 \in \mathbb{R}} \sum_{i=1}^{n} \mathbf{1}(X_i \geq 0)(Y_i - b_0 - X_i b_1)^2 \; K(X_i/h_n),$$

$$(\hat{\mu}_{-,1}(h_n), \hat{\mu}^{(1)}_{-,1}(h_n))' = \arg\min_{b_0,b_1\in\mathbb{R}} \sum_{i=1}^{n} \mathbf{1}(X_i < 0)(Y_i - b_0 - X_i b_1)^2 \; K(X_i/h_n).$$

Conventional approaches to constructing confidence intervals for $\tau_{\texttt{SRD}}$ using the local-linear estimator rely on the following large-sample approximation for the standardized t-statistic (see Lemma B.1(D) in the appendix for the general result): if $nh_n^5 \to 0$ and $nh_n \to \infty$, then

$$T_{\texttt{SRD}}(h_n) = \frac{\hat{\tau}_{\texttt{SRD}}(h_n) - \tau_{\texttt{SRD}}}{\sqrt{\mathsf{V}_{\texttt{SRD}}(h_n)}} \to_d \mathcal{N}(0,1)$$

$$\mathsf{V}_{\texttt{SRD}}(h_n) = \mathbb{V}[\hat{\tau}_{\texttt{SRD}}(h_n)|\mathcal{X}_n], \qquad \mathcal{X}_n = [X_1, \cdots, X_n]'.$$

This justifies the conventional (infeasible) $100(1-\alpha)$-percent confidence interval for $\tau_{\texttt{SRD}}$ given by

$$I_{\texttt{SRD}}(h_n) = \left[\hat{\tau}_{\texttt{SRD}}(h_n) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}_{\texttt{SRD}}(h_n)}\right],$$

with Φ^{-1}_{α} the upper α-quantile of the standard normal distribution (e.g., $\Phi^{-1}_{0.95} \approx 1.96$). In practice, a standard error estimator is needed to construct feasible confidence intervals because the variance $\mathsf{V}_{\texttt{SRD}}(h_n)$ involves unknown quantities, but for now we assume $\mathsf{V}_{\texttt{SRD}}(h_n)$ is known and postpone the issue of standard error estimation until Section 2.5. Even in this simplified known-variance case, the choice of the bandwidth h_n is crucial. The condition $nh_n^5 \to 0$ is explicitly imposed to eliminate the contribution of the leading bias to the distributional approximation, which depends on the unknown second derivatives $\mu^{(2)}_+$ and $\mu^{(2)}_-$ as described in Lemma B.1(B) in the appendix. This means that, in general, the confidence intervals $I_{\texttt{SRD}}(h_n)$ will have correct asymptotic coverage only if the bandwidth h_n is "small" enough to satisfy the bias-condition $nh_n^5 \to 0$.

Several approaches are available in the literature to select h_n, including plug-in rules and cross-validation procedures; see Imbens and Kalyanaraman (2012) for a recent account of the state-of-the-art in bandwidth selection for RD designs. Unfor-

tunately, these approaches lead to bandwidths that are too "large" because they do not satisfy the bias-condition $nh_n^5 \to 0$: minimizing the asymptotic mean squared error (MSE) of $\hat{\tau}_{\texttt{SRD}}(h_n)$ gives the optimal plug-in bandwidth choice $h_{\texttt{MSE}} = C_{\texttt{MSE}}\, n^{-1/5}$ with $C_{\texttt{MSE}}$ a constant, which by construction implies that $n(h_{\texttt{MSE}})^5 \to c \in (0, \infty)$ and hence leads to a first-order bias in the distributional approximation. This is a well-known problem in the nonparametric curve estimation literature. Moreover, implementing this MSE-optimal bandwidth choice in practice is likely to introduce additional variability in the chosen bandwidth that may lead to "large" bandwidths as well. Similarly, cross-validation bandwidth selectors tend to have low convergence rates, and thus also typically lead to "large" bandwidth choices; see, e.g., Ichimura and Todd (2007) and references therein. These observations suggest that commonly used local-linear RD confidence intervals may not exhibit correct coverage in applications due to the presence of a potentially first-order bias in their construction, as illustrated by the simulation evidence we present in Section 2.6. Since applied researchers often estimate RD treatment effects using local-linear regressions with MSE-optimal bandwidths and implicitly ignore the asymptotic bias of the estimator, the poor coverage of conventional confidence intervals we highlight potentially affects many RD empirical applications.

We propose a novel approach to inference based on bias correction to address this problem. Conventional bias correction seeks to remove the leading bias term of the statistic by subtracting off a consistent bias estimate, thus removing the impact of the potentially first-order bias. While systematic and easy to justify theoretically, this approach usually delivers poor performance in finite samples. We propose an alternative large-sample distributional approximation that takes bias correction as a starting point, but improves its performance in finite samples by accounting for the added variability introduced by the bias estimate.

To describe our approach formally, consider first the conventional bias correction

approach. The leading asymptotic bias of the local-linear estimator is

$$\mathbb{E}[\hat{\tau}_{\mathtt{SRD}}(h_n)|\mathcal{X}_n] - \tau_{\mathtt{SRD}} = h_n^2 \mathsf{B}_{\mathtt{SRD}}(h_n) \ \{1 + o_p(1)\}$$

$$\mathsf{B}_{\mathtt{SRD}}(h_n) = \frac{\mu_+^{(2)}}{2!}\mathcal{B}_{+,\mathtt{SRD}}(h_n) - \frac{\mu_-^{(2)}}{2!}\mathcal{B}_{-,\mathtt{SRD}}(h_n),$$

where $\mathcal{B}_{+,\mathtt{SRD}}(h_n)$ and $\mathcal{B}_{-,\mathtt{SRD}}(h_n)$ are asymptotically bounded, observed quantities (function of $\mathcal{X}_n$, $k(\cdot)$ and h_n) explicitly given in Lemma B.1(B) in the appendix. Therefore, a plug-in bias-corrected estimator is

$$\hat{\tau}_{\mathtt{SRD}}^{\mathtt{bc}}(h_n, b_n) = \hat{\tau}_{\mathtt{SRD}}(h_n) - h_n^2 \hat{\mathsf{B}}_{\mathtt{SRD}}(h_n, b_n)$$

$$\hat{\mathsf{B}}_{\mathtt{SRD}}(h_n, b_n) = \frac{\hat{\mu}_{+,2}^{(2)}(b_n)}{2!}\mathcal{B}_{+,\mathtt{SRD}}(h_n) - \frac{\hat{\mu}_{-,2}^{(2)}(b_n)}{2!}\mathcal{B}_{-,\mathtt{SRD}}(h_n),$$

with $\hat{\mu}_{+,2}^{(2)}(b_n)$ and $\hat{\mu}_{+,2}^{(2)}(b_n)$ denoting conventional local-quadratic estimators of $\mu_+^{(2)}$ and $\mu_-^{(2)}$, as described in Section B.1.1 in the appendix. Here, b_n is the so-called pilot bandwidth sequence, usually larger than h_n. As shown in the appendix for the general case, if $nh_n^7 \to 0$ and $h_n/b_n \to 0$, and other regularity conditions hold, then the bias-corrected (infeasible) t-statistic satisfies

$$T_{\mathtt{SRD}}^{\mathtt{bc}}(h_n, b_n) = \frac{\hat{\tau}_{\mathtt{SRD}}^{\mathtt{bc}}(h_n, b_n) - \tau_{\mathtt{SRD}}}{\sqrt{\mathsf{V}_{\mathtt{SRD}}(h_n)}} \to_d \mathcal{N}(0,1),$$

which justifies confidence intervals for $\tau_{\mathtt{SRD}}$ of the form:

$$I_{\mathtt{SRD}}^{\mathtt{bc}}(h_n, b_n) = \left[\left(\hat{\tau}_{\mathtt{SRD}}(h_n) - h_n^2 \hat{\mathsf{B}}_{\mathtt{SRD}}(h_n, b_n)\right) \pm \Phi_{1-\alpha/2}^{-1}\sqrt{\mathsf{V}_{\mathtt{SRD}}(h_n)}\right].$$

That is, in the conventional bias-correction approach, the confidence intervals are re-centered to account for the presence of the bias. This approach allows for potentially "larger" bandwidths h_n, such as the MSE-optimal choice, because the leading asymptotic bias is manually removed from the distributional approximation. In practice, b_n

may also be selected using an MSE-optimal choice, denoted $b_{\texttt{MSE}}$, which can be implemented by a plug-in estimate, denoted $\hat{b}_{\texttt{MSE}}$; see Section 2.4 for details. While bias correction is an appealing theoretical idea, a natural concern with the conventional large-sample approximation for the bias-corrected local-linear RD estimator is that it does not account for the additional variability introduced by the bias estimates $\hat{\mu}^{(2)}_{+,2}(b_n)$ and $\hat{\mu}^{(2)}_{-,2}(b_n)$ and thus the distributional approximation given above tends to provide a poor characterization of the finite-sample variability of the statistic. This large-sample approximation relies on the carefully tailored condition $h_n/b_n \to 0$, which makes the variability of the bias-correction estimate disappear asymptotically. However, h_n/b_n is never zero in finite samples.

Our alternative asymptotic approximation for bias-corrected local polynomial estimators removes the restriction $h_n/b_n \to 0$, leading to alternative confidence intervals for RD treatment effects capturing the (possibly first-order) effect of the bias correction to the distributional approximation. The alternative large-sample approximation we propose for the (properly centered and scaled) estimator $\hat{\tau}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)$ allows for the more general condition $\rho_n = h_n/b_n \to \rho \in [0, \infty]$, which in particular permits a pilot bandwidth b_n of the same order of (and possibly equal to) the main bandwidth h_n. This approach implies that the bias-correction term may not be asymptotically negligible (after appropriate centering and scaling) in general, in which case it will converge in distribution to a centered at zero normal random variable, provided the asymptotic bias is small. Thus, the resulting distributional approximation includes the contribution of both the point estimate $\hat{\tau}_{\texttt{SRD}}(h_n)$ and the bias estimate, leading to a different asymptotic variance in general. This idea is formalized in the following theorem.

Theorem II.4. *Let Assumptions III.1–II.3 hold with $S \geq 3$. If $n \min\{h_n^5, b_n^5\} \max\{h_n^2, b_n^2\} \to$*

0 *and* $n\min\{h_n, b_n\} \to \infty$, *then*

$$T^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n) = \frac{\hat{\tau}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n) - \tau_{\mathtt{SRD}}}{\sqrt{\mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)}} \to_d \mathcal{N}(0,1), \qquad \mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n) = \mathsf{V}_{\mathtt{SRD}}(h_n) + \mathsf{C}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n),$$

provided $\kappa\max\{h_n, b_n\} < \kappa_0$. *The exact form of* $\mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ *is given in Theorem B.2(V) in the appendix.*

Theorem II.4 shows that by standardizing the bias-corrected estimator by its (conditional) variance, the asymptotic distribution of the resulting bias-corrected statistic $T^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ is Gaussian even when the condition $h_n/b_n \to 0$ is violated. The standardization formula $\mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ depends explicitly on the behavior of $\rho_n = h_n/b_n$, and $\mathsf{C}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ may be interpreted as a correction to account for the variability of the estimated bias-correction term. The key practical implication of Theorem II.4 is that it justifies the more robust, theory-based $100(1-\alpha)$-percent confidence intervals:

$$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n) = \left[\left(\hat{\tau}_{\mathtt{SRD}}(h_n) - h_n^2\hat{\mathsf{B}}_{\mathtt{SRD}}(h_n, b_n)\right) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}_{\mathtt{SRD}}(h_n) + \mathsf{C}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)}\right].$$

We summarize important features of our main result in the remarks below.

Remark II.5 (Robustness). The distributional approximation in Theorem II.4 permits one bandwidth (but not both) to be fixed, provided this bandwidth is not too "large"; i.e., both must satisfy $\kappa\max\{h_n, b_n\} < \kappa_0$ for all n large enough, but only one needs to vanish. This theorem allows for all conventional bandwidth sequences and, in addition, permits other bandwidth sequences that would make $I_{\mathtt{SRD}}(h_n)$ and $I^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ invalid (i.e., $\mathbb{P}[\tau_{\mathtt{SRD}} \in I_{\mathtt{SRD}}(h_n)] \not\to 1-\alpha$ and $\mathbb{P}[\tau_{\mathtt{SRD}} \in I^{\mathtt{bc}}_{\mathtt{SRD}}(h_n)] \not\to 1-\alpha$).

Remark II.6 (Asymptotic variance). Three limiting cases are obtained depending on $\rho_n \to \rho \in [0, \infty]$.

Case 1: $\rho = 0$. In this case $h_n = o(b_n)$ and $\mathsf{C}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n) = o_p(\mathsf{V}_{\mathtt{SRD}}(h_n))$, thus making our approach asymptotically equivalent to the standard approach to bias-correction:

$\mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)/\mathsf{V}_{\mathtt{SRD}}(h_n) \to_p 1$.

Case 2: $\rho \in (0, \infty)$. In this case $h_n = \rho b_n$, a knife-edge case, where both $\hat{\tau}_{\mathtt{SRD}}(h_n)$ and $\hat{\mathsf{B}}_{\mathtt{SRD}}(h_n, b_n)$ contribute to the asymptotic variance.

Case 3: $\rho = \infty$. In this case $b_n = o(h_n)$ and $\mathsf{V}_{\mathtt{SRD}}(h_n) = o_p(\mathsf{C}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n))$, implying that the bias-estimate is first-order while the actual estimator $\hat{\tau}_{\mathtt{SRD}}(h_n)$ is of smaller order:

$$\mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)/\mathbb{V}[h_n^2 \hat{\mathsf{B}}_{\mathtt{SRD}}(h_n, b_n)|\mathcal{X}_n] \to_p 1$$

Remark II.7 (Higher-order implications). If h_n and b_n are chosen so that the confidence intervals have correct asymptotic coverage, then $I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ will have faster coverage error rates than $I_{\mathtt{SRD}}(h_n)$ (given the smoothness assumptions imposed). See Calonico, Cattaneo, and Farrell (2014) for further details.

Remark II.8 (Interval length). If $\rho_n = h_n/b_n \to \rho \in [0, \infty)$, then $I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ and $I_{\mathtt{SRD}}(h_n)$ have interval length proportional to $1/\sqrt{nh_n}$. If, in addition, h_n and b_n are chosen so that the confidence intervals have correct asymptotic coverage, then $I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ will have shorter interval length than $I_{\mathtt{SRD}}(h_n)$ for n large enough. However, because the proportionality constant is larger for $I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ than for $I_{\mathtt{SRD}}(h_n)$, the interval $I_{\mathtt{SRD}}(h_n)$ may be shorter than $I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ in small samples. See Section 2.6 for simulation evidence, and Calonico, Cattaneo, and Farrell (2014) for further details.

Remark II.9 (Bootstrap). Bootstrapping $\hat{\tau}_{\mathtt{SRD}}(h_n)$ or $T_{\mathtt{SRD}}(h_n)$ will not improve the performance of the conventional confidence intervals because the bootstrap distribution is centered at $\mathbb{E}[\hat{\tau}_{\mathtt{SRD}}(h_n)|\mathcal{X}_n]$. Bootstrapping $\hat{\tau}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ or $T^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ is possible but not advisable because these quantities are not asymptotically pivotal in general. Bootstrapping the asymptotically pivotal statistic $T^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n)$ is possible, as an alternative to the Gaussian approximation. See Horowitz (2001) for further details.

Remark II.10 (Special case $h_n = b_n$). If $h_n = b_n$ (and the same kernel function

$k(\cdot)$ is used), then $\hat{\tau}^{\mathsf{bc}}_{\mathtt{SRD}}(h_n, h_n)$ is numerically equivalent to the (not bias-corrected) local-quadratic estimator of $\tau_{\mathtt{SRD}}$, and $\mathsf{V}^{\mathsf{bc}}_{\mathtt{SRD}}(h_n, h_n)$ coincides with the variance of the latter estimator. This is true for any polynomial order used (see appendix and online supplemental appendix), which gives a simple connection between local polynomial estimators of order p and $p+1$ and manual bias-correction. Thus, this result provides a formal justification for an inference approach based on increasing the order of the RD estimator: choose h_n to be the MSE-optimal bandwidth for the local-linear estimator, but construct confidence intervals using a t-test based on the local-quadratic estimator instead. This approach corresponds to the case $h_n = b_n$ in Theorem II.4.

Remark II.11 (Nonparametrics and undersmoothing). Our results apply more broadly to nonparametric kernel-based curve estimation problems, and also offer a new theoretical perspective on the trade-off and connection between undersmoothing (i.e., choosing an ad-hoc "smaller" bandwidth) and explicit bias-correction. See Calonico, Cattaneo, and Farrell (2014) for further details.

Remark II.12 (Different bandwidths). All our results may be extended to allow for different bandwidths entering the estimators for control and treatment units. In this case, the different bandwidth sequences should satisfy the conditions imposed in the theorems.

2.3 Other RD Designs

We discuss three extensions of our approach to other empirically relevant settings: sharp kink RD, fuzzy RD and fuzzy kink RD designs. The result presented are special cases of Theorems B.2 and B.4 in the appendix. In all cases, the construction follows the same logic: (*i*) the conventional large-sample distribution is characterized, (*ii*) the leading bias is presented and a plug-in bias-correction is proposed, and (*iii*) the alternative large-sample distribution is derived to obtain the robust confidence

intervals.

2.3.1 Sharp Kink RD

In the sharp kink RD design, interest lies on the difference of the first derivative of the regression functions at the cutoff, as opposed to the differences in the levels of those functions (see, e.g., Card, Lee, Pei, and Weber (2012), Dong (2012), Dong and Lewbel (2012) and references therein). The estimand is $\tau_{\texttt{SKRD}} = \mu_+^{(1)} - \mu_-^{(1)}$.

Although a local-linear estimator could still be used in this context, it is more appropriate to employ a local-quadratic estimator due to boundary-bias considerations. Thus, we focus on the local-quadratic RD estimator $\hat{\tau}_{\texttt{SKRD}}(h_n) = \hat{\mu}_{+,2}^{(1)}(h_n) - \hat{\mu}_{-,2}^{(1)}(h_n)$, where $\hat{\mu}_{+,2}^{(1)}(h_n)$ and $\hat{\mu}_{-,2}^{(1)}(h_n)$ denote local-quadratic estimators of $\mu_+^{(1)}$ and $\mu_-^{(1)}$, respectively; see Section B.1.1 in the appendix. Lemma B.1(D) in the appendix gives $T_{\texttt{SKRD}}(h_n) = (\hat{\tau}_{\texttt{SKRD}}(h_n) - \tau_{\texttt{SKRD}})/\sqrt{\mathsf{V}_{\texttt{SKRD}}(h_n)} \to_d \mathcal{N}(0,1)$ with $\mathsf{V}_{\texttt{SKRD}}(h_n) = \mathbb{V}[\hat{\tau}_{\texttt{SKRD}}(h_n)|\mathcal{X}_n]$, which corresponds to the conventional distributional approximation. The MSE-optimal bandwidth choice for $\hat{\tau}_{\texttt{SKRD}}(h_n)$ is derived in Lemma II.17 in Section 2.4. This choice, among others, will again lead to a non-negligible first-order bias. Proceeding as before, we have $\mathbb{E}[\hat{\tau}_{\texttt{SKRD}}(h_n)|\mathcal{X}_n] - \tau_{\texttt{SKRD}} = h_n^2 \mathsf{B}_{\texttt{SKRD}}(h_n)\ \{1 + o_p(1)\}$ with $\mathsf{B}_{\texttt{SKRD}}(h_n) = \mu_+^{(3)}\mathcal{B}_{+,\texttt{SKRD}}(h_n)/3! - \mu_-^{(3)}\mathcal{B}_{-,\texttt{SKRD}}(h_n)/3!$, where $\mathcal{B}_{+,\texttt{SKRD}}(h_n)$ and $\mathcal{B}_{-,\texttt{SKRD}}(h_n)$ are asymptotically bounded observed quantities (function of $\mathcal{X}_n$, $k(\cdot)$ and h_n), also given in Lemma B.1(B).

A bias-corrected local-quadratic estimator of $\tau_{\texttt{SKRD}}$ is $\hat{\tau}_{\texttt{SKRD}}^{\texttt{bc}}(h_n, b_n) = \hat{\tau}_{\texttt{SKRD}}(h_n) - h_n^2 \hat{\mathsf{B}}_{\texttt{SKRD}}(h_n, b_n)$ with $\hat{\mathsf{B}}_{\texttt{SKRD}}(h_n, b_n) = \hat{\mu}_{+,3}^{(3)}(b_n)\mathcal{B}_{+,\texttt{SKRD}}(h_n)/3! - \hat{\mu}_{-,3}^{(3)}(b_n)\mathcal{B}_{-,\texttt{SKRD}}(h_n)/3!$, where $\hat{\mu}_{+,3}^{(3)}(b_n)$ and $\hat{\mu}_{-,3}^{(3)}(b_n)$ are the local-cubic estimators of $\mu_+^{(3)}$ and $\mu_-^{(3)}$, respectively; see Section B.1.1 in the appendix for details.

Theorem II.13. *Let Assumptions III.1–II.3 hold with $S \geq 4$. If $n \min\{h_n^7, b_n^7\} \max\{h_n^2, b_n^2\} \to$*

0 *and* $n\min\{h_n, b_n\} \to \infty$, *then*

$$T^{\texttt{rbc}}_{\texttt{SKRD}}(h_n, b_n) = \frac{\hat{\tau}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n) - \tau_{\texttt{SKRD}}}{\sqrt{\mathsf{V}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n)}} \to_d \mathcal{N}(0,1),$$

provided $\kappa \max\{h_n, b_n\} < \kappa_0$. *The exact form of* $\mathsf{V}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n)$ *is given in Theorem B.2(V) in the appendix.*

This theorem is analogous to Theorem II.4 for the sharp kink RD design, and derives the new variance formula $\mathsf{V}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n)$ capturing the additional contribution of the bias-correction to the sampling variability. The new variance also takes the form $\mathsf{V}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n) = \mathsf{V}_{\texttt{SKRD}}(h_n) + \mathsf{C}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n)$, where $\mathsf{C}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n)$ is the correction term. This result theoretically justifies the following more robust $100(1-\alpha)$-percent confidence interval for $\tau_{\texttt{SKRD}}$: $I^{\texttt{rbc}}_{\texttt{SKRD}}(h_n, b_n) = \left[\hat{\tau}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}^{\texttt{bc}}_{\texttt{SKRD}}(h_n, b_n)}\right]$.

2.3.2 Fuzzy RD

In the fuzzy RD design, actual treatment status may differ from treatment assignment and is thus only partially determined by the running variable. We introduce the following notation: $(Y_i(0), Y_i(1), T_i(0), T_i(1), X_i)'$, $i = 1, 2, \ldots, n$, is a random sample where in this case treatment status for each unit is $T_i = T_i(0) \cdot \mathbf{1}(X_i < 0) + T_i(1) \cdot \mathbf{1}(X_i \geq 0)$, with $T_i(0), T_i(1) \in \{0, 1\}$. The observed random sample now is $\{(Y_i, T_i, X_i)' : i = 1, 2, \ldots, n\}$. The estimand of interest is $\tau_{\texttt{FRD}} = (\mathbb{E}[Y_i(1)|X = 0] - \mathbb{E}[Y_i(0)|X = 0])/(\mathbb{E}[T_i(1)|X = 0] - \mathbb{E}[T_i(0)|X = 0])$, provided that $\mathbb{E}[T_i(1)|X = 0] - \mathbb{E}[T_i(0)|X = 0] \neq 0$. Under appropriate conditions, this estimand is nonparametrically identifiable as

$$\tau_{\texttt{FRD}} = \frac{\tau_{Y,\texttt{SRD}}}{\tau_{T,\texttt{SRD}}} = \frac{\mu_{Y+} - \mu_{Y-}}{\mu_{T+} - \mu_{T-}}$$

where here, and elsewhere as needed, we make explicit the outcome variable underlying the population parameter. That is, $\tau_{Y,\texttt{SRD}} = \mu_{Y+} - \mu_{Y-}$ with $\mu_{Y+} = \lim_{x \to 0^+} \mu_Y(x)$

and $\mu_{Y-} = \lim_{x\to 0^-} \mu_Y(x)$, $\mu_Y(x) = \mathbb{E}[Y_i|X_i = x]$, and $\tau_{T,\mathtt{SRD}} = \mu_{T+} - \mu_{T-}$ with $\mu_{T+} = \lim_{x\to 0^+} \mu_T(x)$ and $\mu_{T-} = \lim_{x\to 0^-} \mu_T(x)$, $\mu_T(x) = \mathbb{E}[T_i|X_i = x]$. We employ the following additional assumption.

Assumption II.14. *For $\kappa_0 > 0$, the following holds in the neighborhood $(-\kappa_0, \kappa_0)$ around the cutoff $\bar{x} = 0$:*
(a) $\mu_{T-}(x) = \mathbb{E}[T_i(0)|X_i = x]$ and $\mu_{T+}(x) = \mathbb{E}[T_i(1)|X_i = x]$ are S-times continuously differentiable.
(b) $\sigma^2_{T-}(x) = \mathbb{V}[T_i(0)|X_i = x]$ and $\sigma^2_{T+}(x) = \mathbb{V}[T_i(1)|X_i = x]$ are continuous and bounded away from zero.

A popular estimator in this setting is the ratio of two reduced form, sharp local-linear RD estimators:

$$\hat{\tau}_{\mathtt{FRD}}(h_n) = \frac{\hat{\tau}_{Y,\mathtt{SRD}}(h_n)}{\hat{\tau}_{T,\mathtt{SRD}}(h_n)} = \frac{\hat{\mu}_{Y+,1}(h_n) - \hat{\mu}_{Y-,1}(h_n)}{\hat{\mu}_{T+,1}(h_n) - \hat{\mu}_{T-,1}(h_n)},$$

again now making explicit the outcome variable being used in each expression. That is, for a random variable U (equal to either Y or T) we set $\hat{\mu}_{U+,1}(h_n)$ and $\hat{\mu}_{U-,1}(h_n)$ to be the local-linear estimators employing U_i as outcome variable; see Section B.1.1 in the appendix for details.

Under Assumptions III.1–II.14, and appropriate bandwidth conditions, the conventional large-sample properties of $\hat{\tau}_{\mathtt{FRD}}$ are characterized by noting that $\hat{\tau}_{\mathtt{FRD}}(h_n) - \tau_{\mathtt{FRD}} = \tilde{\tau}_{\mathtt{FRD}}(h_n) + R_n$ with $\tilde{\tau}_{\mathtt{FRD}}(h_n) = (\hat{\tau}_{Y,\mathtt{SRD}}(h_n) - \tau_{Y,\mathtt{SRD}})/\tau_{T,\mathtt{SRD}} - \tau_{Y,\mathtt{SRD}}(\hat{\tau}_{T,\mathtt{SRD}}(h_n) - \tau_{T,\mathtt{SRD}})/\tau^2_{T,\mathtt{SRD}}$ and R_n a higher-order reminder term. This shows that, to first-order, the fuzzy RD estimator behaves like a linear combination of two sharp RD estimators. Thus, as Lemma B.3(D) in the appendix shows,

$$T_{\mathtt{FRD}}(h_n) = \frac{\hat{\tau}_{\mathtt{FRD}}(h_n) - \tau_{\mathtt{FRD}}}{\sqrt{\mathsf{V}_{\mathtt{FRD}}(h_n)}} \to_d \mathcal{N}(0,1), \qquad \mathsf{V}_{\mathtt{FRD}}(h_n) = \mathbb{V}[\tilde{\tau}_{\mathtt{FRD}}(h_n)|\mathcal{X}_n].$$

The bias of the local-linear fuzzy RD estimator $\hat{\tau}_{\texttt{FRD}}(h_n)$ is $\mathbb{E}[\tilde{\tau}_{\texttt{FRD}}(h_n)|\mathcal{X}_n] = h_n^2 \mathsf{B}_{\texttt{FRD}}(h_n)\ \{1+ o_p(1)\}$ with

$$\mathsf{B}_{\texttt{FRD}}(h_n) = \left(\frac{1}{\tau_{T,\texttt{SRD}}}\frac{\mu^{(2)}_{Y+}}{2!} - \frac{\tau_{Y,\texttt{SRD}}}{\tau^2_{T,\texttt{SRD}}}\frac{\mu^{(2)}_{T+}}{2!}\right)\mathcal{B}_{+,\texttt{FRD}}(h_n) - \left(\frac{1}{\tau_{T,\texttt{SRD}}}\frac{\mu^{(2)}_{Y-}}{2!} - \frac{\tau_{Y,\texttt{SRD}}}{\tau^2_{T,\texttt{SRD}}}\frac{\mu^{(2)}_{T-}}{2!}\right)\mathcal{B}_{-,\texttt{FRD}}(h_n),$$

where $\mathcal{B}_{+,\texttt{FRD}}(h_n)$ and $\mathcal{B}_{-,\texttt{FRD}}(h_n)$ are also asymptotically bounded observed quantities (function of $\mathcal{X}_n$, $k(\cdot)$ and h_n) and given in Lemma B.3(B). A bias-corrected estimator of $\tau_{\texttt{SRD}}$ employing a local-quadratic estimate of the leading biases is $\hat{\tau}^{\texttt{bc}}_{\texttt{FRD}}(h_n, b_n) = \hat{\tau}_{\texttt{FRD}}(h_n) - h_n^2 \hat{\mathsf{B}}_{\texttt{FRD}}(h_n, b_n)$ with

$$\begin{aligned}\hat{\mathsf{B}}_{\texttt{FRD}}(h_n, b_n) \;=\; &\left(\frac{1}{\hat{\tau}_{T,\texttt{SRD}}(h_n)}\frac{\hat{\mu}^{(2)}_{Y+,2}(b_n)}{2!} - \frac{\hat{\tau}_{Y,\texttt{SRD}}(h_n)}{\hat{\tau}^2_{T,\texttt{SRD}}(h_n)}\frac{\hat{\mu}^{(2)}_{T+,2}(b_n)}{2!}\right)\mathcal{B}_{+,\texttt{FRD}}(h_n)\\ &- \left(\frac{1}{\hat{\tau}_{T,\texttt{SRD}}(h_n)}\frac{\hat{\mu}^{(2)}_{Y-,2}(b_n)}{2!} - \frac{\hat{\tau}_{Y,\texttt{SRD}}(h_n)}{\hat{\tau}^2_{T,\texttt{SRD}}(h_n)}\frac{\hat{\mu}^{(2)}_{T-,2}(b_n)}{2!}\right)\mathcal{B}_{-,\texttt{FRD}}(h_n).\end{aligned}$$

We propose to bias-correct the fuzzy RD estimator using its first-order linear approximation, as opposed to directly bias-correct $\hat{\tau}_{Y,\texttt{SRD}}(h_n)$ and $\hat{\tau}_{T,\texttt{SRD}}(h_n)$ separately in the numerator and denominator of $\hat{\tau}_{\texttt{FRD}}(h_n)$. The former approach seems more intuitive as it captures the leading bias of the actual estimator of interest.

Theorem II.15. *Let Assumptions III.1–II.14 hold with $S \geq 3$, and $\tau_{T,\texttt{SRD}} \neq 0$. If $n\min\{h_n^5, b_n^5\}\max\{h_n^2, b_n^2\} \to 0$ and $n\min\{h_n, b_n\} \to \infty$, then*

$$T^{\texttt{rbc}}_{\texttt{FRD}}(h_n, b_n) = \frac{\hat{\tau}^{\texttt{bc}}_{\texttt{FRD}}(h_n, b_n) - \tau_{\texttt{FRD}}}{\sqrt{\mathsf{V}^{\texttt{bc}}_{\texttt{FRD}}(h_n, b_n)}} \to_d \mathcal{N}(0,1),$$

provided that $h_n \to 0$ and $\kappa b_n < \kappa_0$. The exact form of $\mathsf{V}^{\texttt{bc}}_{\texttt{FRD}}(h_n, b_n)$ is given in Theorem B.4(V).

2.3.3 Fuzzy Kink RD

We retain the notation and assumptions introduced above for the fuzzy RD design. In the fuzzy Kink RD, the parameter of interest and plug-in estimators are, respectively,

$$\tau_{\texttt{FKRD}} = \frac{\tau_{Y,\texttt{SKRD}}}{\tau_{T,\texttt{SKRD}}} = \frac{\mu^{(1)}_{Y+} - \mu^{(1)}_{Y-}}{\mu^{(1)}_{T+} - \mu^{(1)}_{T-}}$$

$$\text{and} \qquad \hat{\tau}_{\texttt{FKRD}}(h_n) = \frac{\hat{\tau}_{Y,\texttt{SKRD}}(h_n)}{\hat{\tau}_{T,\texttt{SKRD}}(h_n)} = \frac{\hat{\mu}^{(1)}_{Y+,2}(h_n) - \hat{\mu}^{(1)}_{Y-,2}(h_n)}{\hat{\mu}^{(1)}_{T+,2}(h_n) - \hat{\mu}^{(1)}_{T-,2}(h_n)},$$

where $\hat{\tau}_{\texttt{FKRD}}(h_n)$ is based on two local-quadratic (reduced form) estimates; see Section B.1.1 in the appendix.

The linearization argument given for the fuzzy RD estimator applies here as well. Employing Lemma B.3(D) in the appendix once more, we verify that

$$T_{\texttt{FKRD}}(h_n) = (\hat{\tau}_{\texttt{FKRD}}(h_n) - \tau_{\texttt{FKRD}})/\sqrt{\mathsf{V}_{\texttt{FKRD}}(h_n)} \to_d \mathcal{N}(0,1)$$

with $\mathsf{V}_{\texttt{FKRD}}(h_n) = \mathbb{V}[\tilde{\tau}_{\texttt{FKRD}}(h_n)|\mathcal{X}_n]$, and $\mathbb{E}[\tilde{\tau}_{\texttt{FKRD}}(h_n)|\mathcal{X}_n] = h_n^2 \mathsf{B}_{\texttt{FKRD}}(h_n)\ \{1 + o_p(1)\}$ with

$$\begin{aligned}
\mathsf{B}_{\texttt{FKRD}}(h_n) &= (\mu^{(3)}_{Y+}/\tau_{T,\texttt{SKRD}} - \tau_{Y,\texttt{SKRD}}\mu^{(3)}_{T+}/\tau^2_{T,\texttt{SKRD}})\mathcal{B}_{+,\texttt{FKRD}}(h_n)/3! \\
&- (\mu^{(3)}_{Y-}/\tau_{T,\texttt{SKRD}} - \tau_{Y,\texttt{SKRD}}\mu^{(3)}_{T-}/\tau^2_{T,\texttt{SKRD}})\mathcal{B}_{-,\texttt{FKRD}}(h_n)/3!
\end{aligned}$$

where $\mathcal{B}_{-,\texttt{FKRD}}(h_n)$ and $\mathcal{B}_{-,\texttt{FKRD}}(h_n)$ are also given in Lemma B.3. A plug-in bias-corrected estimator of $\tau_{\texttt{FKRD}}$ employing local-cubic estimates of the leading biases is $\hat{\tau}^{\texttt{bc}}_{\texttt{FKRD}}(h_n, b_n) = \hat{\tau}_{\texttt{FKRD}}(h_n) - h_n^2 \hat{\mathsf{B}}_{\texttt{FKRD}}(h_n, b_n)$, where

$$\begin{aligned}
\hat{\mathsf{B}}_{\texttt{FKRD}}(h_n, b_n) &= (\hat{\mu}^{(3)}_{Y+,3}(b_n)/\hat{\tau}_{T,\texttt{SKRD}}(h_n) - \hat{\tau}_{Y,\texttt{SKRD}}(h_n)\hat{\mu}^{(3)}_{T+,3}(b_n)/\hat{\tau}^2_{T,\texttt{SKRD}}(h_n))\mathcal{B}_{+,\texttt{FKRD}}(h_n)/3! \\
&- (\hat{\mu}^{(3)}_{Y-,3}(b_n)/\hat{\tau}_{T,\texttt{SKRD}}(h_n) - \hat{\tau}_{Y,\texttt{SKRD}}(h_n)\hat{\mu}^{(3)}_{T-,3}(b_n)/\hat{\tau}^2_{T,\texttt{SKRD}}(h_n))\mathcal{B}_{-,\texttt{FKRD}}(h_n)/3!
\end{aligned}$$

Theorem II.16. *Let Assumptions III.1–II.14 hold with $S \geq 4$, and $\tau_{T,\texttt{SKRD}} \neq 0$. If*

$n \min\{h_n^7, b_n^7\} \max\{h_n^2, b_n^2\} \to 0$ *and* $n \min\{h_n^3, b_n\} \to \infty$, *then*

$$T^{\mathtt{rbc}}_{\mathtt{FKRD}}(h_n, b_n) = \frac{\hat{\tau}^{\mathtt{bc}}_{\mathtt{FKRD}}(h_n, b_n) - \tau_{\mathtt{FKRD}}}{\sqrt{\mathsf{V}^{\mathtt{bc}}_{\mathtt{FKRD}}(h_n, b_n)}} \to_d \mathcal{N}(0,1),$$

provided that $h_n \to 0$ *and* $\kappa b_n < \kappa_0$. *The exact form of* $\mathsf{V}^{\mathtt{bc}}_{\mathtt{FKRD}}(h_n, b_n)$ *is given in Theorem B.4(V).*

2.4 Validity of MSE-Optimal Bandwidth Selectors

Following Imbens and Kalyanaraman (2012), we derive MSE-optimal bandwidth choices for h_n and b_n that apply to all the RD settings discussed previously. These bandwidth choices are not valid when conventional distributional approximations are used, but they are fully compatible with our distributional approach. Let

$$\begin{aligned}
\Gamma_p &= \int_0^\infty K(u) r_p(u) r_p(u)' du \\
\vartheta_{p,q} &= \int_0^\infty K(u) u^q r_p(u) du \\
\Psi_p &= \int_0^\infty K(u)^2 r_p(u) r_p(u)' du
\end{aligned}$$

where $r_p(x) = (1, x, \cdots, x^p)'$ and e_ν is the conformable $(\nu+1)$-th unit vector (e.g., $e_1 = (0,1,0)'$ if $p = 2$). See Section B.1.1 in the appendix for more details.

2.4.1 Sharp Designs

To handle the sharp RD and shark kink RD designs together, as well as the choice of pilot bandwidths, we introduce more general notation. The estimands in the sharp RD designs may be written as $\tau_\nu = \mu_+^{(\nu)} - \mu_-^{(\nu)}$ with, in particular, $\tau_{\mathtt{SRD}} = \tau_0$ and $\tau_{\mathtt{SKRD}} = \tau_1$. The p-th order local-polynomial RD estimators are $\hat{\tau}_{\nu,p}(h_n) = \hat{\mu}_{+,p}^{(\nu)}(h_n) - \hat{\mu}_{-,p}^{(\nu)}(h_n)$

with $\nu \leq p$ with, in particular, $\hat{\tau}_{\texttt{SRD}}(h_n) = \hat{\tau}_{0,1}(h_n)$ and $\hat{\tau}_{\texttt{SKRD}}(h_n) = \hat{\tau}_{1,2}(h_n)$.

Lemma II.17. *Suppose Assumptions III.1–II.3 hold with $S \geq p+1$, and $\nu \leq p$. If $h_n \to 0$ and $nh_n \to \infty$, then*

$$\mathbb{E}[(\hat{\tau}_{\nu,p}(h_n) - \tau_\nu)^2 | \mathcal{X}_n] = h_n^{2(p+1-\nu)} \ \left[\mathsf{B}^2_{\nu,p,p+1} + o_p(1)\right] + n^{-1} h_n^{-1-2\nu} \ \left[\mathsf{V}_{\nu,p} + o_p(1)\right],$$

where $\mathsf{B}_{\nu,p,r} = (\mu_+^{(r)} - (-1)^{\nu+r}\mu_-^{(r)}) e_\nu' \Gamma_p^{-1} \vartheta_{p,r}/r!$ *and* $\mathsf{V}_{\nu,p} = (\sigma_-^2 + \sigma_+^2)\nu!^2 e_\nu' \Gamma_p^{-1} \Psi_p \Gamma_p^{-1} e_\nu / f$.

If, in addition, $\mathsf{B}_{\nu,p,p+1} \neq 0$, then the (asymptotic) MSE-optimal bandwidth is

$$h_{\texttt{MSE},\nu,p} = C_{\texttt{MSE},\nu,p}^{\frac{1}{2p+3}} \ n^{-\frac{1}{2p+3}}, \qquad C_{\texttt{MSE},\nu,p} = \frac{(1+2\nu)\mathsf{V}_{\nu,p}}{2(p+1-\nu)\mathsf{B}^2_{\nu,p,p+1}}.$$

This lemma is a generalization of Imbens and Kalyanaraman (2012) to include kink RD designs, among other possibilities. It justifies a set of MSE-optimal (infeasible) choices for h_n and b_n: $h_n = h_{\texttt{MSE},0,1}$ and $b_n = h_{\texttt{MSE},2,2}$ for Theorem II.4, and $h_n = h_{\texttt{MSE},1,2}$ and $b_n = h_{\texttt{MSE},3,3}$ for Theorem II.13.

Remark II.18 (Bandwidths validity). The MSE-optimal bandwidth choices for the sharp designs are fully compatible with our confidence intervals because they satisfy the rate-restrictions in Theorems II.4–II.13. For example, $n \min\{h_{\texttt{MSE},0,1}, b_{\texttt{MSE},2,2}\} \to \infty$ and $n \min\{h_{\texttt{MSE},0,1}^5, b_{\texttt{MSE},2,2}^5\} \max\{h_{\texttt{MSE},0,1}^2, b_{\texttt{MSE},2,2}^2\} \to 0$ in Theorem II.4.

Remark II.19 (Estimated bandwidths). Section B.1.4 in the appendix describes new data-driven direct plug-in (DPI) bandwidth selectors for sharp RD designs based on Lemma II.17. Following Imbens and Kalyanaraman (2012), our proposed bandwidths incorporate "regularization" to avoid small denominators. However, relative to the selectors proposed by Imbens and Kalyanaraman (2012), our bandwidth selectors have two distinct features: *(i)* our estimator of $\mathsf{V}_{\nu,p}$ that does not require a choice of pilot bandwidth and avoids estimating σ_+^2, σ_-^2 and f directly, and *(ii)* pilot bandwidths are chosen to be MSE-optimal and thus the final bandwidth selectors are of the ℓ-stage

DPI variety ((Wand and Jones, 1995, Section 3.6)). Our final bandwidth selectors are consistent and optimal in the sense of Li (1987); see Theorem B.5 in the appendix.

Remark II.20 (Optimal ρ_n). The MSE-optimal bandwidth choices imply $\rho_n \to 0$. In research underway we are investigating whether this is an optimal choice from a distributional approximation perspective. See Remarks II.7 and II.8, and Calonico, Cattaneo, and Farrell (2014) for related discussion.

2.4.2 Fuzzy Designs

Let $\varsigma_\nu = \tau_{Y,\nu}/\tau_{T,\nu}$ with $\tau_{Y,\nu} = \mu_{Y+}^{(\nu)} - \mu_{Y-}^{(\nu)}$ and $\tau_{T,\nu} = \mu_{T+}^{(\nu)} - \mu_{T-}^{(\nu)}$. In particular, $\tau_{\texttt{FRD}} = \varsigma_0$ and $\tau_{\texttt{FKRD}} = \varsigma_1$. The p-th order local-polynomial estimators are $\hat{\varsigma}_{\nu,p}(h_n) = \hat{\tau}_{Y,\nu}(h_n)/\hat{\tau}_{T,\nu}(h_n)$ with $\nu \leq p$, $\hat{\tau}_{Y,\nu}(h_n) = \hat{\mu}_{Y+,p}^{(\nu)}(h_n) - \hat{\mu}_{Y-,p}^{(\nu)}(h_n)$ and $\hat{\tau}_{T,\nu}(h_n) = \hat{\mu}_{T+,p}^{(\nu)}(h_n) - \hat{\mu}_{T-,p}^{(\nu)}(h_n)$; see Section B.1.1 in the appendix. In particular, $\hat{\tau}_{\texttt{FRD}}(h_n) = \hat{\varsigma}_{0,1}(h_n)$ and $\hat{\tau}_{\texttt{FKRD}}(h_n) = \hat{\varsigma}_{1,2}(h_n)$. The first-order linear approximation of $\hat{\varsigma}_{\nu,p}(h_n)$ is $\tilde{\varsigma}_{\nu,p}(h_n) = (\hat{\tau}_{Y,\nu,p}(h_n) - \tau_{Y,\nu})/\tau_{T,\nu} - \tau_{Y,\nu}(\hat{\tau}_{T,\nu,p}(h_n) - \tau_{T,\nu})/\tau_{T,\nu}^2$, which we employ to construct the (approximate) MSE objective function.

Lemma II.21. *Suppose Assumptions III.1–II.14 hold with $S \geq p+1$, and $\nu \leq p$. If $h_n \to 0$ and $nh_n \to \infty$, then*

$$\mathbb{E}[(\tilde{\varsigma}_{\nu,p}(h_n))^2|\mathcal{X}_n] = h_n^{2(p+1-\nu)} \ \left[\mathsf{B}_{\mathtt{F},\nu,p,p+1}^2 + o_p(1)\right] + \frac{1}{nh_n^{1+2\nu}} \ \left[\mathsf{V}_{\mathtt{F},\nu,p} + o_p(1)\right],$$

where

$$\begin{aligned}
\mathsf{B}_{\mathtt{F},\nu,p,r} &= ((\mu_{Y+}^{(r)} - (-1)^{\nu+r}\mu_{Y-}^{(r)})/\tau_{T,\nu} - \tau_{Y,\nu}(\mu_{T+}^{(r)} - (-1)^{\nu+r}\mu_{T-}^{(r)})/\tau_{T,\nu}^2)e_\nu'\Gamma_p^{-1}\vartheta_{p,r}/r! \\
\mathsf{V}_{\mathtt{F},\nu,p} &= ((\sigma_{YY-}^2 + \sigma_{YY+}^2)/\tau_{T,\nu}^2 \\
&- 2\tau_{Y,\nu}(\sigma_{YT-}^2 + \sigma_{YT+}^2)/\tau_{T,\nu}^3 + \tau_{Y,\nu}^2(\sigma_{TT-}^2 + \sigma_{TT+}^2)/\tau_{T,\nu}^4)\nu!^2 e_\nu'\Gamma_p^{-1}\Psi_p\Gamma_p^{-1}e_\nu/f
\end{aligned}$$

If, in addition, $\mathsf{B}_{\mathtt{F},\nu,p,p+1} \neq 0$*, then the (asymptotic) MSE-optimal bandwidth is*

$$h_{\mathtt{MSE},\mathtt{F},\nu,p} = C_{\mathtt{MSE},\mathtt{F},\nu,p}^{\frac{1}{2p+3}}\, n^{-\frac{1}{2p+3}}, \qquad C_{\mathtt{MSE},\mathtt{F},\nu,p} = \frac{(2\nu+1)\mathsf{V}_{\mathtt{F},\nu,p}}{2(p+1-\nu)\mathsf{B}_{\mathtt{F},\nu,p,p+1}^2}.$$

Valid bandwidth choices of h_n and b_n for Theorems II.15–II.16 are also readily available using Lemma II.21: $h_n = h_{\mathtt{MSE},\mathtt{F},0,1}$ and $b_n = h_{\mathtt{MSE},\mathtt{F},2,2}$ for Theorem II.15, and $h_n = h_{\mathtt{MSE},\mathtt{F},1,2}$ and $b_n = h_{\mathtt{MSE},\mathtt{F},3,3}$ for Theorem II.16. This lemma generalizes Imbens and Kalyanaraman (2012) to account for fuzzy kink RD designs. Feasible versions can be developed along the lines of Section B.1.4 in the appendix. Importantly, just as in the sharp RD cases (Remark II.18), these MSE-optimal bandwidth choices are fully compatible with our asymptotics.

2.5 Standard Errors

The exact formulas for the new variances $\mathsf{V}_{\mathtt{SRD}}^{\mathtt{bc}}(h_n,b_n)$ [sharp RD], $\mathsf{V}_{\mathtt{SKRD}}^{\mathtt{bc}}(h_n,b_n)$ [sharp kink RD], $\mathsf{V}_{\mathtt{FRD}}^{\mathtt{bc}}(h_n,b_n)$ [fuzzy RD] and $\mathsf{V}_{\mathtt{FKRD}}^{\mathtt{bc}}(h_n,b_n)$ [fuzzy kink RD] in Theorems II.4–II.16, respectively, are straightforward to derive but notationally cumbersome. They all have the same structure because they are derived by computing the conditional variance of (linear combinations of weighted) linear least-squares estimators. The only unknowns in these variance matrices are (depending on the setting under consideration, sharp or fuzzy RD designs) the diagonal matrices: $\Psi_{YY+,p,q}(h_n,b_n)$, $\Psi_{YT+,p,q}(h_n,b_n)$, $\Psi_{TT+,p,q}(h_n,b_n)$, $\Psi_{YY-,p,q}(h_n,b_n)$, $\Psi_{YT-,p,q}(h_n,b_n)$ and $\Psi_{TT-,p,q}(h_n,b_n)$, with $p,q \in \mathbb{N}_+$ and the generic notation

$$\begin{aligned}
\Psi_{UV+,p,q}(h_n,b_n) &= \sum\nolimits_{i=1}^{n} \mathbf{1}(X_i \geq 0) K_{h_n}(X_i) K_{b_n}(X_i) r_p(X_i/h_n) r_q(X_i/b_n)' \sigma_{UV+}^2(X_i)/n, \\
\Psi_{UV-,p,q}(h_n,b_n) &= \sum\nolimits_{i=1}^{n} \mathbf{1}(X_i < 0) K_{h_n}(X_i) K_{b_n}(X_i) r_p(X_i/h_n) r_q(X_i/b_n)' \sigma_{UV-}^2(X_i)/n,
\end{aligned}$$

where $\sigma^2_{UV+}(x) = \mathbb{C}ov[U(1), V(1)|X = x]$ and $\sigma^2_{UV-}(x) = \mathbb{C}ov[U(0), V(0)|X = x]$, and U and V are placeholders for either Y or T. This generality is required to handle the fuzzy designs, where the covariances between Y_i and T_i arise naturally. Theorems B.2 and B.4 in the appendix give the exact standard error formulas, showing how the matrices $\Psi_{UV+,p,q}(h_n, b_n)$ and $\Psi_{UV-,p,q}(h_n, b_n)$ are employed.

The $(p+1)\times(q+1)$ matrices $\Psi_{UV+,p,q}(h_n, b_n)$ and $\Psi_{UV-,p,q}(h_n, b_n)$ are a generalization of the middle matrix in the traditional Huber-Eicker-White heteroskedasticity-robust standard error formula for linear models, and thus an analogue of these standard error estimator can be constructed by plugging in the corresponding estimated residuals. This choice, although simple and convenient, may not perform well in finite-samples because it implicitly employs the bandwidth choices used to construct the estimates of the underlying regression functions. As an alternative, following Abadie and Imbens (2006), we propose standard error estimators based on nearest-neighbor estimators with a fixed tuning parameter, which may be more robust in finite-samples. Specifically, we define:

$$\begin{aligned}
\hat{\Psi}_{UV+,p,q}(h_n, b_n) &= \sum\nolimits_{i=1}^{n} \mathbf{1}(X_i \geq 0) K_{h_n}(X_i) K_{b_n}(X_i) r_p(X_i/h_n) r_q(X_i/h_n)' \hat{\sigma}^2_{UV+}(X_i)/n, \\
\hat{\Psi}_{UV-,p,q}(h_n, b_n) &= \sum\nolimits_{i=1}^{n} \mathbf{1}(X_i < 0) K_{h_n}(X_i) K_{b_n}(X_i) r_p(X_i/h_n) r_q(X_i/h_n)' \hat{\sigma}^2_{UV-}(X_i)/n,
\end{aligned}$$

with

$$\begin{aligned}
\hat{\sigma}^2_{UV+}(X_i) &= \mathbf{1}(X_i \geq 0) \frac{J}{J+1} \left(U_i - \sum\nolimits_{j=1}^{J} U_{\ell_{+,j}(i)}/J \right) \left(V_i - \sum\nolimits_{j=1}^{J} V_{\ell_{+,j}(i)}/J \right), \\
\hat{\sigma}^2_{UV-}(X_i) &= \mathbf{1}(X_i < 0) \frac{J}{J+1} \left(U_i - \sum\nolimits_{j=1}^{J} U_{\ell_{-,j}(i)}/J \right) \left(V_i - \sum\nolimits_{j=1}^{J} V_{\ell_{-,j}(i)}/J \right),
\end{aligned}$$

where $\ell^+_j(i)$ is the j-th closest unit to unit i among $\{X_i : X_i \geq 0\}$ and $\ell^-_j(i)$ is the j-th closest unit to unit i among $\{X_i : X_i < 0\}$. ("Local sample covariances" could be used instead; see Abadie and Imbens (2010).)

In the supplemental appendix (Calonico, Cattaneo, and Titiunik (2014d)), we show that these estimators are asymptotically valid for any choice of $J \in \mathbb{N}_+$, because they are approximately conditionally unbiased (even though inconsistent for fixed nearest-neighbors $J \geq 1$). This justifies employing $\hat{\Psi}_{UV+,p,q}(h_n, b_n)$ and $\hat{\Psi}_{UV-,p,q}(h_n, b_n)$ in place of $\Psi_{UV+,p,q}(h_n, b_n)$ and $\Psi_{UV-,p,q}(h_n, b_n)$ to construct the estimators $\hat{\mathsf{V}}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$, $\hat{\mathsf{V}}^{\mathtt{bc}}_{\mathtt{SKRD}}(h_n, b_n)$, $\hat{\mathsf{V}}^{\mathtt{bc}}_{\mathtt{FRD}}(h_n, b_n)$ and $\hat{\mathsf{V}}^{\mathtt{bc}}_{\mathtt{FKRD}}(h_n, b_n)$. For example, in Theorem II.4, feasible confidence intervals are

$$\hat{I}^{\mathtt{rbc}}_{\mathtt{SRD}}(h_n, b_n) = \left[\hat{\tau}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n) \pm \Phi^{-1}_{1-\alpha/2} \sqrt{\hat{\mathsf{V}}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)} \right],$$

where $\hat{\mathsf{V}}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n, b_n)$ is constructed using $\hat{\Psi}_{YY+,1,1}(h_n, b_n)$, $\hat{\Psi}_{YY+,1,2}(h_n, b_n)$, $\hat{\Psi}_{YY+,2,1}(h_n, b_n)$, $\hat{\Psi}_{YY+,2,2}(h_n, b_n)$, $\hat{\Psi}_{YY-,1,1}(h_n, b_n)$, $\hat{\Psi}_{YY-,1,2}(h_n, b_n)$, $\hat{\Psi}_{YY-,2,1}(h_n, b_n)$ and $\hat{\Psi}_{YY-,2,2}(h_n, b_n)$. The other confidence intervals are constructed analogously.

2.6 Simulation Evidence

We report the main results of a Monte Carlo experiment. We conducted $10,000$ replications, and for each replication we generated a random sample $\{(X_i, \varepsilon_i)' : i = 1, ..., n\}$ with size $n = 500$, $X_i \sim 2\mathcal{B}(2,4) - 1$ with $\mathcal{B}(p_1, p_2)$ denoting a beta distribution with parameters p_1 and p_2, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$ with $\sigma_\varepsilon = 0.1295$. Three regression functions are considered (Figure 2.1), denoted $\mu_1(x)$, $\mu_2(x)$ and $\mu_3(x)$, and labeled Model 1, 2 and 3, respectively. The outcome is generated as $Y_i = \mu_j(X_i) + \varepsilon_i$, $i = 1, 2, \cdots, n$, for each regression model $j = 1, 2, 3$. The exact functional form of $\mu_1(x)$ and $\mu_2(x)$ was obtained from the data in Lee (2008) and Ludwig and Miller (2007), respectively, while $\mu_3(x)$ was chosen to exhibit more curvature. All other features of the simulation study are held fixed, matching exactly the data generating process in Imbens and Kalyanaraman (2012). For further details see (Calonico, Cattaneo, and Titiunik, 2014d, Section 3).

Figure 2.1: Regression Functions for Models 1–3 in simulations.

We consider confidence intervals for $\tau_{\texttt{SRD}}$ (sharp RD), employing a local-linear RD estimator ($p = 1$) with local-quadratic bias-correction ($q = 2$), denoted $\hat{\tau}^{\texttt{rbc}}_{\texttt{SRD}}(h_n, b_n)$ as in Section 2.2. We report empirical coverage and interval length of conventional (based on $T_{\texttt{SRD}}(h_n)$) and robust (based on $T^{\texttt{rbc}}_{\texttt{SRD}}(h_n, b_n)$) 95% confidence intervals for different bandwidth choices:

$$\hat{I}_{\texttt{SRD}}(h_n) = \left[\hat{\tau}_{\texttt{SRD}}(h_n) \pm 1.96 \sqrt{\hat{\mathsf{V}}_{\texttt{SRD}}(h_n)}\right]$$

$$\hat{I}^{\texttt{rbc}}_{\texttt{SRD}}(h_n, b_n) = \left[\hat{\tau}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n) \pm 1.96 \sqrt{\hat{\mathsf{V}}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)}\right],$$

where the estimators $\hat{\mathsf{V}}_{\texttt{SRD}}(h_n)$ and $\hat{\mathsf{V}}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)$ are constructed using the nearest-neighbor procedure discussed in Section 2.5 with $J = 3$. For comparison, we also report infeasible confidence intervals employing infeasible standard errors ($\mathsf{V}_{\texttt{SRD}}(h_n)$ and $\mathsf{V}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)$), and those constructed using the standard "plug-in estimated residuals" approach, which we denote $\check{\mathsf{V}}_{\texttt{SRD}}(h_n)$ and $\check{\mathsf{V}}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)$.

Table 1 presents the main simulation results. The main bandwidth h_n is chosen in four different ways: (i) infeasible MSE-optimal choice $h_{\texttt{MSE},0,1}$, denoted $h_{\texttt{MSE}}$; (ii) plug-in, regularized MSE-optimal selector as described in (Imbens and Kalyanaraman, 2012, Section 4.1), denoted $\hat{h}_{\texttt{IK}}$; (iii) cross-validation as described in (Imbens and Kalyanaraman, 2012, Section 4.5), denoted $\hat{h}_{\texttt{CV}}$; and (iv) plug-in choice proposed in

Section 2.4 (Remark II.19), denoted $\hat{h}_{\texttt{CCT}}$. Similarly, to choose the pilot bandwidth b_n, we construct modified versions of the choices enumerated above, with the exception of $\hat{h}_{\texttt{CV}}$ because cross-validation is not readily available for derivative estimation; these choices are denoted $b_{\texttt{MSE}}$, $\hat{b}_{\texttt{IK}}$ and $\hat{b}_{\texttt{CCT}}$, respectively. For further results, including other bandwidth selectors and test statistics, see (Calonico, Cattaneo, and Titiunik, 2014d, Section 3).

The simulation results show that the robust confidence intervals lead to important improvements in empirical coverage (EC) with moderate increments in average empirical interval length (IL). The empirical coverage of the interval estimator $I^{\texttt{rbc}}_{\texttt{SRD}}(h_n, b_n)$ exhibits an improvement of about 10-15 percentage points on average with respect to the conventional interval $I_{\texttt{SRD}}(h_n)$, depending on the particular model, standard error estimator and bandwidth choices considered. As expected, the feasible versions of the confidence intervals exhibit slightly more empirical coverage distortion and longer intervals than their infeasible counterparts. The conventional plug-in residual standard error estimators ($\check{\mathsf{V}}_{\texttt{SRD}}(h_n)$ and $\check{\mathsf{V}}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)$) tend to exhibit more undercoverage in our simulations than the proposed fixed-neighbor standard error estimators ($\hat{\mathsf{V}}_{\texttt{SRD}}(h_n)$ and $\hat{\mathsf{V}}^{\texttt{bc}}_{\texttt{SRD}}(h_n, b_n)$). The choice $\rho_n = 1$ is not only simple and intuitive (Remark II.10), but also performed well in our simulations. Although not the main goal of this chapter, we also found that our two-stage direct plug-in rule selector of h_n performs well relative to the other plug-in selectors, and on par with the cross-validation bandwidth selector.

2.7 Conclusion

We introduced new confidence interval estimators for several regression-discontinuity estimands that enjoy demonstrably superior robustness properties. The results cover the sharp (level or kink) and fuzzy (level or kink) RD designs. Our confidence intervals were constructed using an alternative asymptotic theory for bias-corrected local

polynomial estimators in the context of RD designs, which leads to a different asymptotic variance in general and thus justifies a new standard-error estimator. We found that the resulting data-driven confidence intervals performed very well in simulations, suggesting in particular that they provide a robust (to the choice of bandwidths) alternative when compared to the conventional confidence intervals routinely employed in empirical work.

Table 2.1: Empirical Coverage and Average Interval Length of different 95% Confidence Intervals

	Conventional Approach							Robust Approach							Bandwidths	
		EC (%)			IL				EC (%)			IL				
		V	$\hat{\mathsf{V}}$	$\check{\mathsf{V}}$	V	$\hat{\mathsf{V}}$	$\check{\mathsf{V}}$		V^{bc}	$\hat{\mathsf{V}}^{\mathsf{bc}}$	$\check{\mathsf{V}}^{\mathsf{bc}}$	V^{bc}	$\hat{\mathsf{V}}^{\mathsf{bc}}$	$\check{\mathsf{V}}^{\mathsf{bc}}$	h_n	b_n
Model 1																
	$I_{\mathtt{SRD}}(h_{\mathtt{MSE}})$	93.5	92.9	91.0	0.225	0.230	0.213	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_{\mathtt{MSE}}, b_{\mathtt{MSE}})$	94.5	93.8	92.2	0.273	0.279	0.258	0.166	0.251
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}})$	82.3	83.0	80.5	0.149	0.152	0.145	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}}, \hat{b}_{\mathtt{IK}})$	93.2	93.0	92.2	0.276	0.282	0.269	0.399	0.349
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{CV}})$	80.8	81.5	79.1	0.145	0.149	0.141	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CV}}, \hat{b}_{\mathtt{CV}})$	91.8	91.5	90.0	0.213	0.217	0.206	0.428	0.428
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}})$	90.7	90.4	88.4	0.202	0.205	0.193	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}}, \hat{b}_{\mathtt{CCT}})$	92.9	92.5	90.9	0.238	0.242	0.227	0.211	0.345
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_{\mathtt{MSE}}, h_{\mathtt{MSE}})$	94.7	93.2	92.0	0.339	0.343	0.315	0.166	0.166
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}}, \hat{h}_{\mathtt{IK}})$	92.8	92.5	91.3	0.218	0.223	0.212	0.399	0.399
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}}, \hat{h}_{\mathtt{CCT}})$	95.0	93.6	92.6	0.301	0.303	0.284	0.211	0.211
Model 2																
	$I_{\mathtt{SRD}}(h_{\mathtt{MSE}})$	92.4	91.7	86.7	0.315	0.342	0.283	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_{\mathtt{MSE}}, b_{\mathtt{MSE}})$	94.8	94.1	90.5	0.342	0.372	0.308	0.082	0.189
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}})$	10.5	13.3	13.6	0.199	0.215	0.223	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}}, \hat{b}_{\mathtt{IK}})$	93.7	94.3	95.9	0.290	0.320	0.340	0.213	0.222
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{CV}})$	76.6	78.5	73.1	0.263	0.288	0.251	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CV}}, \hat{b}_{\mathtt{CV}})$	94.6	94.1	91.8	0.401	0.451	0.377	0.124	0.124
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}})$	86.7	87.6	80.6	0.298	0.326	0.265	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}}, \hat{b}_{\mathtt{CCT}})$	94.2	94.2	90.5	0.323	0.355	0.288	0.095	0.220
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_{\mathtt{MSE}}, h_{\mathtt{MSE}})$	95.0	93.5	89.5	0.490	0.542	0.427	0.082	0.082
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}}, \hat{h}_{\mathtt{IK}})$	94.0	94.2	95.8	0.297	0.329	0.346	0.213	0.213
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}}, \hat{h}_{\mathtt{CCT}})$	95.1	94.2	90.8	0.461	0.517	0.400	0.095	0.095
Model 3																
	$I_{\mathtt{SRD}}(h_{\mathtt{MSE}})$	85.8	85.8	84.1	0.179	0.184	0.175	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_{\mathtt{MSE}}, b_{\mathtt{MSE}})$	94.7	94.2	93.6	0.235	0.240	0.230	0.260	0.322
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}})$	84.3	84.4	82.8	0.184	0.189	0.179	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}}, \hat{b}_{\mathtt{IK}})$	95.0	94.4	93.6	0.223	0.228	0.218	0.247	0.392
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{CV}})$	93.1	92.7	90.9	0.219	0.224	0.208	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CV}}, \hat{b}_{\mathtt{CV}})$	94.9	93.4	92.3	0.329	0.333	0.307	0.177	0.177
	$I_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}})$	91.3	90.8	89.1	0.214	0.217	0.203	$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}}, \hat{b}_{\mathtt{CCT}})$	94.9	94.1	92.5	0.245	0.249	0.233	0.186	0.341
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(h_{\mathtt{MSE}}, h_{\mathtt{MSE}})$	94.9	94.0	93.4	0.266	0.271	0.259	0.260	0.260
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{IK}}, \hat{h}_{\mathtt{IK}})$	94.9	94.0	93.4	0.274	0.279	0.265	0.247	0.247
								$I^{\mathtt{rbc}}_{\mathtt{SRD}}(\hat{h}_{\mathtt{CCT}}, \hat{h}_{\mathtt{CCT}})$	95.3	93.7	92.6	0.320	0.322	0.300	0.186	0.186

Notes: (i) EC = empirical coverage in percentage points; (ii) IL = empirical average interval length; (iii) columns "Bandwidths" report the population and average estimated bandwidth choices, for main bandwidth h_n and pilot bandwidth b_n; (iv) $\mathsf{V} = \mathsf{V}(h_n)$ and $\mathsf{V}^{\mathsf{bc}} = \mathsf{V}^{\mathsf{bc}}(h_n, b_n)$ denote infeasible variance estimators using the population variance of the residuals, $\hat{\mathsf{V}} = \hat{\mathsf{V}}(h_n)$ and $\hat{\mathsf{V}}^{\mathsf{bc}} = \hat{\mathsf{V}}^{\mathsf{bc}}(h_n, b_n)$ denote variance estimators constructed using nearest-neighbor standard errors with $J = 3$, and $\check{\mathsf{V}} = \check{\mathsf{V}}(h_n)$ and $\check{\mathsf{V}}^{\mathsf{bc}} = \check{\mathsf{V}}^{\mathsf{bc}}(h_n, b_n)$ denote variance estimators constructed using the conventional plug-in estimated residuals.

CHAPTER III

Optimal Data-Driven Regression Discontinuity Plots

3.1 Introduction

The regression discontinuity (RD) design, originally introduced by Thistlethwaite and Campbell (1960), is by now among the most popular quasi-experimental empirical strategies to estimate (local) causal treatment effects in Economics, Political Science and many other social, behavioral and natural sciences. In this research design, for each unit $i = 1, 2, \cdots, n$, researchers observe an outcome variable Y_i and a continuous covariate X_i, and units are assigned to treatment or control depending on whether their observed covariate exceeds a known cutoff. Provided the units of analysis cannot systematically sort around the cutoff, the RD design employs observations just below and just above the cutoff as control and treatment groups to conduct inference on the (local) causal effect of the treatment. The underlying idea, and crucial assumption, is that units around the cutoff do not differ in their unobservable characteristics, thereby offering valid counterfactual comparisons between control and treatment groups. For recent reviews on the RD design, including references to a large number of empirical applications employing RD designs, see van der Klaauw (2008), Cook (2008), Imbens and Lemieux (2008), Lee and Lemieux (2010) and Dinardo and Lee (2011).

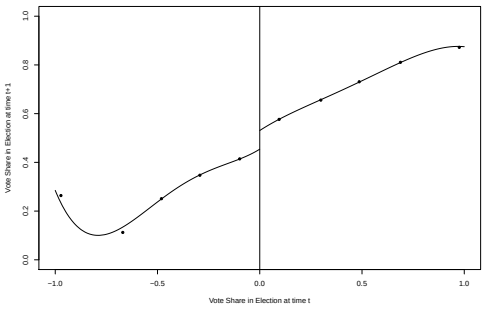
A key feature of the RD design is its simplicity and transparency. The empirical analysis relies on simple and easy to interpret identifying assumptions to study the effect of a policy, intervention or other treatment for units near the threshold, involving only a univariate outcome Y_i and a univariate continuous covariate X_i (which determines treatment assignment). Estimation and inference of RD treatment effects is usually conducted using local polynomial estimators, and great attention has been devoted to these estimators in the recent methodological RD literature (see, for example, Hahn, Todd, and van der Klaauw, 2001; Porter, 2003; Imbens and Kalyanaraman, 2012; Calonico, Cattaneo, and Titiunik, 2014c, and references therein). Other approaches are also possible, such as those employing randomization inference methods (Cattaneo, Frandsen, and Titiunik, 2014). No matter the inference approach employed in empirical work, formal exploratory data analysis and graphical falsification tests are essential when employing RD designs. These methods have been strongly advocated in the literature because they play an important role in both presentation and validation of RD research designs (e.g., Imbens and Lemieux (2008, Section 3) and Lee and Lemieux (2010, Section 4.1)).

So called RD plots are nowadays used in almost all RD empirical applications to illustrate the research design. These popular plots are constructed using two main ingredients. First, the plot shows two smooth polynomial approximations of the underlying conditional expectations of the outcome variable Y_i given the observed covariate X_i, for control and treatment units separately. These polynomial fits seek to present graphically the behavior of the underlying conditional expectations in a smooth fashion and from a global perspective. The second ingredient in the RD plots concerns a collection of local sample means of the outcome variable: first, the support of the covariate X_i is partitioned into disjoint bins for control and treatment units separately, and then sample means of the outcome variable Y_i are computed for each bin using, in each case, only observations that have covariate X_i within each bin. This

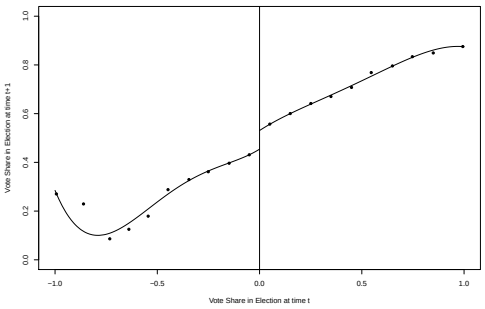
collection of local sample means are then plotted on top of the smooth polynomial fits, with the goal of (i) highlighting potential discontinuities in the underlying conditional expectations and (ii) providing a sense of the local behavior of the data for different values of covariate X_i. Figure 3.1 shows four examples of RD plots employing the data of Lee (2008), and using different choices for the number of bins. In this empirical example, Lee studies the incumbency advantage in U.S. House elections, and his identification strategy is based on the discontinuity generated by the rule that assigns electoral victory to the party that obtains the most votes. The forcing variable is the margin of victory in a given election—the difference in vote share between the Democratic candidate and her strongest opponent— and the threshold is $\bar{x} = 0$, since the the party wins the election when its margin of victory is positive and loses otherwise. The outcome variable is the Democratic vote share in the following U.S House election. We further discuss this empirical application in Section 3.5.

While RD plots are a well established and commonly used tool in empirical analysis of RD designs, their formal properties remain unknown. In particular, these plots are constructed using an *ad hoc* choice of the partitions' size (i.e., the number of bins used to construct the partitions), making the procedure less automatic and more subjective than is ideal for a tool whose main role is to provide objective evidence about the plausibility of the design's main assumptions. Given the absence of concrete guidance on these choices, practitioners typically experiment and select an arbitrary number of bins, which may misrepresent the actual behavior of the data. In this chapter, we study the properties of the most common RD plot used in the literature, one that employs an evenly-spaced binning of the data, and propose an integrated mean-square error (IMSE) optimal choice for the number of bins. We then propose several data-driven, nonparametric implementations of this IMSE-optimal partition size selector and show that they are consistent under simple and easy-to-interpret assumptions. The resulting optimal data-driven selector provides the first fully auto-

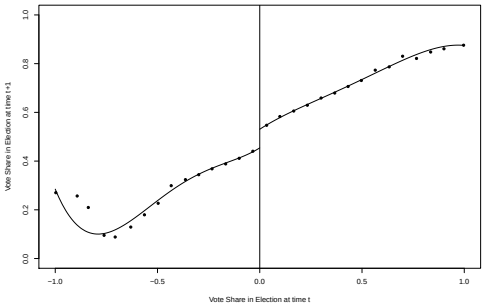
Figure 3.1: RD Plots - House Elections Data from Lee (2008).



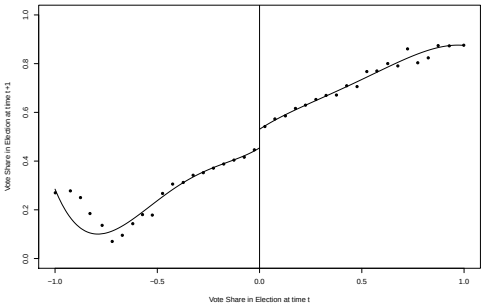
(a) 5 bins on each side.



(b) 10 bins on each side.



(c) 15 bins on each side.



(d) 20 bins on each side.

matic and objective benchmark in the RD literature, offering concrete guidance for empirical work employing RD plots.

In addition to studying RD plots with evenly-spaced bins, we also introduce an alternative RD plot based on quantile-spaced binning. This approach forces each bin to have approximately the same number of observations, a feature that may be appealing when the data is too sparse. This alternative partitioning scheme for the construction of the local sample means may be viewed as (covariate) design adaptive. For this case, we also derive an expansion of the IMSE, propose the corresponding optimal choice of number of bins, and develop data-driven nonparametric consistent implementations thereof.

Our main implementations employ spacings estimation techniques to construct the data-driven IMSE-optimal partition size choices because these estimators do not require additional tuning parameter choices, and thus are more robust in applications. However, this technique requires continuity of the outcome variable, and hence is not applicable in all possible empirical settings. To handle non-continuous outcomes, we also propose and formally analyze IMSE-optimal partition size data-driven choices employing nonparametric polynomial estimators, which can be used broadly under mild assumptions.

Finally, we also analyze the performance of our automatic RD plots numerically. First, we apply our results to two empirical illustrations studying incumbency advantage in the U.S., and find that our optimal data-driven RD-plots perform well when using real data. Second, we study the finite-sample properties of our results in a Monte Carlo experiment employing several data generating processes, and find that our RD-plots tuning parameter selectors perform extremely well. Third, we compare numerically the two RD plotting alternatives analyzed in this chapter: evenly-spaced vs. quantile-spaced. Our results highlight the fact that neither approach dominates the other in general, because features of the underlying (unknown) data generating

process (i.e., distribution of X_i and shapes of the conditional expectation and conditional heteroskedasticity) ultimately determine which RD plot is best from a IMSE perspective. Nonetheless, we offer some intuitive discussion on the relative merits of each approach.

The rest of the chapter is organized as follows. Section 3.2 presents the RD design and reviews basic results and concepts, including a generic formal description of the RD plots. Section 3.3 introduces the popular evenly-spaced RD plot, derives a weighted IMSE expansion and presents our results for this case, while Section 3.4 proceeds analogously but for the alternative RD plot based on quantile-spaced bins. Section 3.5 compares the two RD plots approach using our IMSE expansions, and also showcases how the exploratory data devices perform numerically using simulated and real data. Section 3.6 describes some potential extensions of our work and concludes. Companion `R` and `STATA` software packages are described in Calonico, Cattaneo, and Titiunik (2014e,b).

3.2 Setup and RD plots

In the regression discontinuity design, the observed data is a random sample $(Y_i, X_i)'$, $i = 1, 2, \ldots, n$, from a large population, with X_i a continuous random variable with (possibly restricted) support $[x_l, x_u]$ and density $f(x)$. All units with a value of the observed "score" or "forcing" variable X_i greater than a known threshold $\bar{x}$ are assigned to the treatment group ($T_i = 1$), while all units with $X_i < \bar{x}$ are assigned to the control group ($T_i = 0$). Thus, under perfect compliance, treatment assignment is defined as $T_i = \mathbb{1}(X_i \geq \bar{x})$ with $\mathbb{1}(\cdot)$ denoting the indicator function. As is common in the program evaluation literature (e.g., Imbens and Wooldridge (2009)), we employ potential outcomes notation to characterize the two underlying counterfactual states (control or treatment). Letting $Y_i(1)$ and $Y_i(0)$ denote the potential outcome with

and without treatment, respectively, the observed outcome is

$$Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i = \begin{cases} Y_i(0) & \text{if } X_i < \bar{x} \\ Y_i(1) & \text{if } X_i \geq \bar{x} \end{cases} .$$

The most popular parameter of interest is the average treatment effect at the threshold, given by $\tau_{\texttt{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = \bar{x}]$. This parameter is nonparametrically identifiable under a mild continuity condition (Hahn, Todd, and van der Klaauw, 2001), and RD estimators employing local polynomial techniques have become the default choice in the literature (Imbens and Kalyanaraman, 2012; Calonico, Cattaneo, and Titiunik, 2014c, and references therein). In the so called sharp RD design, T_i is a deterministic function of treatment assignment (perfect compliance), while in the so called fuzzy RD design treatment take-up and treatment assignment may differ. This distinction, however, is irrelevant for our purposes because we do not focus on estimation and inference for RD treatment effects, but rather on the RD plots commonly encountered in empirical work. These plots may be used for presentation and falsification of both sharp and fuzzy RD research designs. These RD plots are described in great detail in the upcoming section, but first we introduce the main notation and assumptions employed throughout the chapter.

We set

$$\mu_-(x) = \mathbb{E}[Y_i(0)|X_i = x], \qquad \sigma^2_-(x) = \mathbb{V}[Y_i(0)|X_i = x],$$

$$\mu_+(x) = \mathbb{E}[Y_i(1)|X_i = x], \qquad \sigma^2_+(x) = \mathbb{V}[Y_i(1)|X_i = x],$$

and impose the following assumption through the chapter.

Assumption III.1. *For $x_l, x_u \in \mathbb{R}$ with $x_l < \bar{x} < x_u$, and all $x \in [x_l, x_u]$:*

(a) $\mathbb{E}[Y_i^4|X_i = x]$ is bounded, and $f(x)$ is continuous and bounded away from zero.

(b) $\mu_-(x)$ and $\mu_+(x)$ are S-times continuously differentiable.

(c) $\sigma^2_-(x)$ and $\sigma^2_+(x)$ are continuous and bounded away from zero.

Part (a) in Assumption III.1 imposes existence of moments and requires that the running variable X_i be continuously distributed. Part (b) imposes smoothness on the underlying regression functions, while part (c) requires that the conditional variance be continuous; all these functions may be different at either side of the threshold. Notice that $\mu_-(x) = \mathbb{E}[Y_i|X_i = x]$ for all $x < \bar{x}$ and $\mu_+(x) = \mathbb{E}[Y_i|X_i = x]$ for all $x \geq \bar{x}$, enabling (consistent) estimation of these conditional expectations for control and treatment units, respectively.

3.2.1 RD Plots

The main features of an RD design are easily summarized employing RD plots [Imbens and Lemieux (2008, Section 3) and Lee and Lemieux (2010, Section 4.1)]. As mentioned in the Introduction, these plots include two main ingredients: (i) smooth polynomial estimation, and (ii) local sample-means estimation. We now formalize the underlying estimation approaches used to construct the RD plots, which provides the basis for our analysis. Our main focus is on tuning parameter selection for the construction of the collection of local sample means under two distinct partitioning schemes: evenly-spaced and quantile-spaced partitions of $[x_l, \bar{x})$ and $[\bar{x}, x_u]$.

3.2.1.1 Global Polynomial Estimation

In the RD plots the unknown functions $\mu_-(x) = \mathbb{E}[Y_i(0)|X_i = x]$ and $\mu_+(x) = \mathbb{E}[Y_i(1)|X_i = x]$ are estimated using global polynomials for control and treatment observations separately. To formalize this approach, let $k \in \mathbb{Z}_+$ and $\mathbf{r}_k(x) = (1, x, x^2, \cdots, x^k)'$, and define

$$\hat{\mu}_{-,k}(x) = \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{-,k}, \qquad \hat{\boldsymbol{\beta}}_{-,k} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x})(Y_i - \mathbf{r}_k(x)'\boldsymbol{\beta})^2,$$

$$\hat{\mu}_{+,k}(x) = \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{+,k}, \qquad \hat{\boldsymbol{\beta}}_{+,k} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x})(Y_i - \mathbf{r}_k(x)'\boldsymbol{\beta})^2,$$

with $p \in \mathbb{Z}_{++} = \{1, 2, \cdots\}$. In words, $\hat{\mu}_{-,k}(x)$ and $\hat{\mu}_{+,k}(x)$ are k-th order polynomial fits of Y_i on X_i employing only control and treatment units, respectively.

These polynomial regressions may be viewed as a nonparametric approach, usually called series or (linear) sieve estimation, for the approximation of the underlying population conditional expectations when $k = k_n \to \infty$ as $n \to \infty$ (see, e.g., Newey (1997b), Chen (2007b) and Belloni, Chernozhukov, Chetverikov, and Kato (2013) for reviews). Below we will exploit this interpretation explicitly to construct consistent plug-in rules for the optimal tuning parameter choices. Employing results from the nonparametric literature, it is possible to select k_n using some data-driven approach such as (plug-in) IMSE minimization or cross-validation. In practice, however, $k = 4$ or $k = 5$ are almost always the preferred choices. Either way, we do not discuss further the choice of k for RD plots because this is a well understood problem. Instead, our main focus is on choosing the partition size for the local means as discussed next, a result that is not currently available in the literature.

3.2.1.2 Local Mean Estimation

The second ingredient in the RD plots are a collection of local sample means of the outcome variable computed over a disjoint partition of the support of the running variable, for control and treatment units separately. To describe this construction formally, we employ ideas from the nonparametric literature on partitioning estimators (for further details see Cattaneo and Farrell, 2013b, and references therein).

Set $\mathbb{1}_A(x) = \mathbb{1}(x \in A)$ to save notation. The partitioning estimators (of order 1), sometimes called binning estimators or local-mean estimators, are formally described as follows:

$$\hat{\mu}_-(x; J_{-,n}) = \sum_{j=1}^{J_{-,n}} \mathbb{1}_{P_{-,j}}(x)\bar{Y}_{-,j}, \qquad \bar{Y}_{-,j} = \frac{\mathbb{1}(N_{-,j} > 0)}{N_{-,j}} \sum_{i=1}^{n} \mathbb{1}_{P_{-,j}}(X_i)Y_i$$

$$\hat{\mu}_+(x;J_{+,n}) = \sum_{j=1}^{J_{+,n}} \mathbb{1}_{P_{+,j}}(x)\bar{Y}_{+,j}, \qquad \bar{Y}_{+,j} = \frac{\mathbb{1}(N_{+,j}>0)}{N_{+,j}} \sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i)Y_i$$

with

$$N_{-,j} = \sum_{i=1}^{n} \mathbb{1}_{P_{-,j}}(X_i), \quad N_- = \sum_{j=1}^{J_{-,n}} N_{-,j}, \quad N_{+,j} = \sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i), \quad N_+ = \sum_{j=1}^{J_{+,n}} N_{+,j},$$

and where $\mathcal{P}_{-,n} = \{P_{-,j} : j = 1, 2, \cdots, J_{-,n}\}$ and $\mathcal{P}_{+,n} = \{P_{+,j} : j = 1, 2, \cdots, J_{+,n}\}$ are generic disjoint partitions of the support of the running variable X_i, which vary with the sample size n. More precisely,

$$[x_l, \bar{x}] = \bigcup_{j=1}^{J_{-,n}} P_{-,j}, \qquad P_{-,j} = \begin{cases} [x_l \ , \ p_{-,1}) & j = 1 \\ [p_{-,j-1} \ , \ p_{-,j}) & j = 2, \cdots, J_{+,n} - 1 \\ [p_{-,J_{-,n}-1} \ , \ \bar{x}) & j = J_{-,n} \end{cases}$$

and

$$[\bar{x}, x_u] = \bigcup_{j=1}^{J_{+,n}} P_{+,j}, \qquad P_{+,j} = \begin{cases} [\bar{x} \ , \ p_{+,1}) & j = 1 \\ [p_{+,j-1} \ , \ p_{+,j}) & j = 2, \cdots, J_{+,n} - 1 \\ [p_{+,J_{+,n}-1} \ , \ x_u] & j = J_{+,n} \end{cases}$$

with $J_{-,n}, J_{+,n} \in \mathbb{Z}_{++}$ denoting the partition sizes for control and treatment groups, respectively.

The estimators $\hat{\mu}_-(x;J_{-,n})$ and $\hat{\mu}_+(x;J_{+,n})$ collect the sample means of the outcomes Y_i for observations with covariate X_i taking values within each bin in the partitions $\mathcal{P}_{-,n}$ and $\mathcal{P}_{+,n}$, and may be interpreted as nonparametric estimators of $\mu_-(x)$ and $\mu_+(x)$, respectively. As for other nonparametric procedures, these binning-type estimators involve a choice of tuning and smoothing parameters. In this case, $(J_{-,n}, J_{+,n})$ may be regarded as the tuning parameters (e.g., similar to a bandwidth for conventional kernel estimators) and $(\mathcal{P}_{-,n}, \mathcal{P}_{+,n})$ may be viewed as the smoothing

parameters (e.g., similar to the shape of kernel function for conventional kernel estimators). Under Assumption III.1, and provided a well-behaved partitioning scheme is used, it is not difficult to show that $\hat{\mu}_-(x; J_{-,n}) \to_{\mathbb{P}} \mu_-(x)$ and $\hat{\mu}_+(x; J_{+,n}) \to_{\mathbb{P}} \mu_+(x)$, provided that $J_{-,n} \to \infty$ and $J_{+,n} \to \infty$ as $n \to \infty$ and some regularity conditions hold.

The behavior of these estimators is dependent on how the partitions are constructed and, as mentioned above, this chapter considers two approaches for choosing the partitions: evenly-spaced partitions and quantile-spaced partitions. Given a chosen partitioning scheme, the parameters $J_{-,n}$ and $J_{+,n}$ control the rate of approximation of the partitioning estimators, capturing the usual bias-variance trade-off: smaller $(J_{-,n}, J_{+,n})$ imply more variance but less bias (more smaller bins), while larger $(J_{-,n}, J_{+,n})$ imply less variance but more bias (fewer larger bins). The main contribution of this chapter is to derive optimal choices of $(J_{-,n}, J_{+,n})$ based on an IMSE objective function for each of the two partitioning schemes, and to develop consistent data-driven implementations thereof.

As we briefly discuss in Section 3.6, these choices can also be used to conduct inference and to construct falsification tests in the context of RD designs. We plan to formally investigate the properties of these inferential procedures in upcoming research work.

3.3 Evenly-Spaced RD Plots

In this section we consider evenly-spaced (ES) bins for the construction of the partitioning scheme underlying the RD plots. Thus, we set

$$p_{-,j} = x_l + j \cdot \frac{\bar{x} - x_l}{J_{-,n}} \qquad \text{and} \qquad p_{+,j} = \bar{x} + j \cdot \frac{x_u - \bar{x}}{J_{+,n}},$$

leading to the evenly-spaced partitioning estimators denoted by $\hat{\mu}_{\mathtt{ES},-}(x; J_{-,n})$ and $\hat{\mu}_{\mathtt{ES},+}(x; J_{+,n})$, with (nonrandom) partitioning schemes denoted by $\mathcal{P}_{\mathtt{ES},-,n}$ and $\mathcal{P}_{\mathtt{ES},+,n}$, respectively.

3.3.1 Optimal Choice of ES Partition Size

To select the number of bins $J_{-,n}$ and $J_{+,n}$ we consider an approximation of the IMSE loss function of these estimators:

$$\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n}) = \int_{x_l}^{\bar{x}} \mathbb{E}\left[(\hat{\mu}_{\mathtt{ES},-}(x; J_{-,n}) - \mu_-(x))^2 \middle| \mathbf{X}_n \right] w(x) dx,$$

$$\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n}) = \int_{\bar{x}}^{x_u} \mathbb{E}\left[(\hat{\mu}_{\mathtt{ES},+}(x; J_{+,n}) - \mu_+(x))^2 \middle| \mathbf{X}_n \right] w(x) dx,$$

where $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)'$. The following theorem gives our main result. Throughout the chapter all limits are taken as $n \to \infty$ unless otherwise stated.

Theorem III.2. *Suppose Assumption III.1 holds with $S \geq 2$, and $w : [x_l, x_u] \mapsto \mathbb{R}_+$ is continuous.*

(−) If $J_{-,n} \log(J_{-,n})/n \to 0$ and $J_{-,n} \to \infty$, then

$$\mathsf{IMSE}_{\mathtt{ES},-}(J_{n,-}) = \frac{J_{-,n}}{n} \mathscr{V}_{\mathtt{ES},-}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{-,n}^2} \mathscr{B}_{\mathtt{ES},-}\{1 + o_{\mathbb{P}}(1)\},$$

$$\mathscr{V}_{\mathtt{ES},-} = \frac{1}{\bar{x} - x_l} \int_{x_l}^{\bar{x}} \frac{\sigma_-^2(x)}{f(x)} w(x) dx \qquad \text{and} \qquad \mathscr{B}_{\mathtt{ES},-} = \frac{(\bar{x} - x_l)^2}{12} \int_{x_l}^{\bar{x}} \left(\mu_-^{(1)}(x) \right)^2 w(x) dx,$$

where $\mu_-^{(1)}(x) = \partial \mu_-(x)/\partial x$.

(+) If $J_{+,n} \log(J_{+,n})/n \to 0$ and $J_{+,n} \to \infty$, then

$$\mathsf{IMSE}_{\mathtt{ES},+}(J_{n,+}) = \frac{J_{+,n}}{n} \mathscr{V}_{\mathtt{ES},+}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{+,n}^2} \mathscr{B}_{\mathtt{ES},+}\{1 + o_{\mathbb{P}}(1)\},$$

$$\mathscr{V}_{\mathtt{ES},+} = \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \frac{\sigma_+^2(x)}{f(x)} w(x) dx \qquad \textit{and} \qquad \mathscr{B}_{\mathtt{ES},+} = \frac{(x_u - \bar{x})^2}{12} \int_{\bar{x}}^{x_u} \left(\mu_+^{(1)}(x)\right)^2 w(x) dx,$$

where $\mu_+^{(1)}(x) = \partial \mu_+(x)/\partial x$.

This theorem gives the result for a family of IMSE loss functions, depending on the choice of weight function $w(x)$. This result remains valid if $w(x) = w_+(x)\mathbb{1}(x \geq \bar{x}) + w_-(x)\mathbb{1}(x < \bar{x})$, thus allowing for $w(x)$ discontinuous at $\bar{x}$, though for notational simplicity we do not discuss this case. In general, assuming that $\mathscr{B}_{\mathtt{ES},-} \neq 0$ and $\mathscr{B}_{\mathtt{ES},+} \neq 0$, the expansions of $\mathsf{IMSE}_{\mathtt{ES},-}(J_{n,-})$ and $\mathsf{IMSE}_{\mathtt{ES},+}(J_{n,+})$ can be used to derive optimal choices of $J_{-,n}$ and $J_{+,n}$:

$$J_{\mathtt{ES},-,n} = \left\lfloor \left(\frac{2\mathscr{B}_{\mathtt{ES},-}}{\mathscr{V}_{\mathtt{ES},-}} \right)^{1/3} n^{1/3} \right\rfloor \qquad \text{and} \qquad J_{\mathtt{ES},+,n} = \left\lfloor \left(\frac{2\mathscr{B}_{\mathtt{ES},+}}{\mathscr{V}_{\mathtt{ES},+}} \right)^{1/3} n^{1/3} \right\rfloor \tag{3.1}$$

with $\lfloor x \rfloor$ denoting the smallest integer part of $x \in \mathbb{R}_{++}$. (This "optimal" choice implies a slight undersmoothing in finite samples.)

3.3.2 Data-Driven Implementations of $J_{\mathtt{ES},-,n}$ and $J_{\mathtt{ES},+,n}$

Employing some reference model, we could easily construct rule-of-thumb estimates of the unknown constants $\mathscr{V}_{\mathtt{ES},-}$, $\mathscr{B}_{\mathtt{ES},-}$, $\mathscr{V}_{\mathtt{ES},+}$ and $\mathscr{B}_{\mathtt{ES},+}$ to estimate empirically the IMSE-optimal size of evenly-spaced partitions given in (3.1), for a given choice of weighting function $w(x)$—see, e.g., Wand and Jones (1995) for further discussion in the context of kernel-based estimation. In this chapter, we propose easy-to-implement consistent nonparametric estimators of $J_{\mathtt{ES},-,n}$ and $J_{\mathtt{ES},+,n}$ instead. First, we outline a general approach allowing for a user-chosen known weighting function $w(x)$. We then also discuss the special case of $w(x) = f(x)$ in a follow-up remark because this choice simplifies the estimation approach. In all cases, we estimate $\mu_-^{(1)}(x)$ and $\mu_+^{(1)}(x)$ using global polynomial approximations, trying to mimic as close as possible current empirical practices: these polynomial approximations are already available as part of

the RD plots. Our approaches are not only theoretically justified, but also arguably simple, easy-to-interpret and more robust than the usual nonparametric alternatives in some cases, as we discussed further below.

Taking $w(x)$ as given, we estimate the constants $\mathscr{V}_{\mathtt{ES},-}$, $\mathscr{B}_{\mathtt{ES},-}$, $\mathscr{V}_{\mathtt{ES},+}$ and $\mathscr{B}_{\mathtt{ES},+}$ using ideas related to spacings estimators (see, e.g., Ghosh and Jammalamadaka, 2001; Lewbel and Schennach, 2007; Baryshnikov, Penrose, and Yurich, 2009, and references therein). These estimators are closely related to nearest neighbor estimators with fixed neighbors (e.g., Abadie and Imbens, 2006, 2010), and are more robust than other nonparametric estimators such as kernel-based estimators because they do not require additional tuning parameter choices in their implementation. To describe the spacings estimators, we need to introduce notation for order statistics and concomitants. For a collection of continuous random variables $\{(Z_i, W_i) : i = 1, 2, \cdots, n\}$ we let $W_{(i)}$ be the i-th order statistic of W_i and $Z_{[i]}$ its corresponding concomitant. That is, $W_{(1)} < W_{(2)} < \cdots < W_{(n)}$ and $(Z_{[i]}, W_{(i)}) = (Z_i, W_{(i)})$ for all $i = 1, 2, \cdots, n$. For further details on order statistics and their associated concomitants see David and Nagaraja (1998, 2003).

Letting $\{(Y_{-,i}, X_{-,i}) : i = 1, 2, \cdots, N_-\}$ and $\{(Y_{+,i}, X_{+,i}) : i = 1, 2, \cdots, N_+\}$ be the subsamples of control ($X_i < \bar{x}$) and treatment ($X_i \geq \bar{x}$) units, respectively, and with the above notation, we propose the following generic estimators:

$$\hat{\mathscr{V}}_{\mathtt{ES},-} = \frac{1}{\bar{x} - x_l} \frac{n}{4} \sum_{i=2}^{N_-} (X_{-,(i)} - X_{-,(i-1)})^2 (Y_{-,[i]} - Y_{-,[i-1]})^2 w(\bar{X}_{-,(i)}), \tag{3.2}$$

$$\hat{\mathscr{B}}_{\mathtt{ES},-} = \frac{(\bar{x} - x_l)^2}{12} \sum_{i=2}^{N_-} (X_{-,(i)} - X_{-,(i-1)}) \left(\hat{\mu}^{(1)}_{-,k}(\bar{X}_{-,[i]}) \right)^2 w(\bar{X}_{-,(i)}), \tag{3.3}$$

and

$$\hat{\mathscr{V}}_{\mathtt{ES},+} = \frac{1}{x_u - \bar{x}} \frac{n}{4} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^2 (Y_{+,[i]} - Y_{+,[i-1]})^2 w(\bar{X}_{+,(i)}), \tag{3.4}$$

$$\hat{\mathscr{B}}_{\mathtt{ES},+} = \frac{(x_u - \bar{x})^2}{12} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)}) \left(\hat{\mu}^{(1)}_{+,k}(\bar{X}_{+,[i]}) \right)^2 w(\bar{X}_{+,(i)}), \tag{3.5}$$

with

$$\bar{X}_{-,(i)} = \frac{X_{-,(i)} + X_{-,(i-1)}}{2}, \qquad i = 2, 3, \cdots, N_-, \qquad \hat{\mu}^{(1)}_{-,k}(x) = \mathbf{r}^{(1)}_k(x)' \hat{\boldsymbol{\beta}}_{-,k},$$

$$\bar{X}_{+,(i)} = \frac{X_{+,(i)} + X_{+,(i-1)}}{2}, \qquad i = 2, 3, \cdots, N_+, \qquad \hat{\mu}^{(1)}_{+,k}(x) = \mathbf{r}^{(1)}_k(x)' \hat{\boldsymbol{\beta}}_{+,k},$$

and $\mathbf{r}^{(1)}_k(x) = \partial \mathbf{r}_k(x)/\partial x = (0, 1, 2x, 3x^2, \cdots, kx^{k-1})'$. The intuition behind these constructions comes from observing that, conditional on N_+,

$$X_{+,(i)} - X_{+,(i-1)} \approx \frac{1}{N_+ f_+(\bar{X}_{-,(i)})}, \qquad f_+(x) = \frac{\mathbb{1}(x \geq \bar{x}) f(x)}{P_+}, \qquad P_+ = \mathbb{P}[X_i \geq \bar{x}],$$

and

$$\mathbb{E}[(Y_{+,[i]} - Y_{+,[i-1]})^2 | X_{+,(1)}, \cdots, X_{+,(N_+)}] \approx \sigma^2_+(X_{+,(i)}) + \sigma^2_+(X_{+,(i-1)}) \approx 2\sigma^2_+(\bar{X}_{+,[i]}),$$

which, after plugging in, leads to the results in Theorem III.3 below when combined with an appropriate limit theorem for the resulting averages. Lemma C.2 in the appendix gives more general results along these lines. As mentioned above, these estimators are particularly well suited for our purposes, and are arguably more robust in practice, because they (i) avoid explicit estimation of the density $f(x)$ appearing in the denominators and (ii) do not require specific choices of tuning parameters (e.g. bandwidths in kernel-based estimation). For these reasons, and given their simple implementation, we recommend employing the above spacings-based estimators whenever possible.

To summarize, in the case of ES partitions, our proposed selectors are

$$\hat{J}_{\mathtt{ES},-,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{ES},-}}{\hat{\mathscr{V}}_{\mathtt{ES},-}} \right)^{1/3} n^{1/3} \right\rfloor \qquad \text{and} \qquad \hat{J}_{\mathtt{ES},+,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{ES},+}}{\hat{\mathscr{V}}_{\mathtt{ES},+}} \right)^{1/3} n^{1/3} \right\rfloor, \quad (3.6)$$

using the estimators in (3.2)-(3.3) and (3.4)-(3.5), respectively. The following theorem shows that, when the polynomial fits are viewed as nonparametric approximations with $k = k_n \to \infty$, these partition size selectors are nonparametric consistent.

Theorem III.3. *Suppose Assumption III.1 holds with $S \geq 5$, $w : [x_l, x_u] \mapsto \mathbb{R}_+$ is continuous, and $Y_i(0)$ and $Y_i(1)$ are continuously distributed. If $k_n^7/n \to 0$ and $k_n \to \infty$, then*

$$\frac{\hat{J}_{\mathtt{ES},-,n}}{J_{\mathtt{ES},-,n}} \to_{\mathbb{P}} 1 \qquad \textit{and} \qquad \frac{\hat{J}_{\mathtt{ES},+,n}}{J_{\mathtt{ES},+,n}} \to_{\mathbb{P}} 1.$$

This theorem gives formal justification for employing $\hat{J}_{\mathtt{ES},-,n}$ and $\hat{J}_{\mathtt{ES},+,n}$ in applications, whenever the outcome variable is continuous and the weight function $w(\cdot)$ is known. A particularly convenient and easy-to-implement choice of the latter is $w(x) = 1$, but other choices are also covered by theorem.

Remark III.4 (Discontinuous Outcomes). When $Y_i(0)$ and $Y_i(1)$ are not continuously distributed, the concomitant-based estimation method becomes invalid. In this case, we need to employ other more standard nonparametric techniques. For example, assuming that $\mathbb{E}[Y_i(t)^2|X_i = x]$, $t = 0, 1$, are twice continuously differentiable, we can use the following estimators:

$$\check{\mathscr{V}}_{\mathtt{ES},-} = \frac{1}{\bar{x} - x_l} \frac{n}{2} \sum_{i=2}^{N_-} (X_{-,(i)} - X_{-,(i-1)})^2 \hat{\sigma}_-^2(\bar{X}_{-,(i)}) w(\bar{X}_{-,(i)}), \qquad \hat{\sigma}_-^2(x) = \hat{\mu}_{-,k,2}(x) - (\hat{\mu}_{-,k,1}(x))^2,$$

$$\check{\mathscr{V}}_{\mathtt{ES},+} = \frac{1}{x_u - \bar{x}} \frac{n}{2} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^2 \hat{\sigma}_+^2(\bar{X}_{+,(i)}) w(\bar{X}_{+,(i)}), \qquad \hat{\sigma}_+^2(x) = \hat{\mu}_{+,k,2}(x) - (\hat{\mu}_{+,k,1}(x))^2,$$

where

$$\hat{\mu}_{-,k,p}(x) = \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{-,k,p}, \qquad \hat{\boldsymbol{\beta}}_{-,k,p} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x})(Y_i^p - \mathbf{r}_k(x)'\boldsymbol{\beta})^2,$$

$$\hat{\mu}_{+,k,p}(x) = \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{+,k,p}, \qquad \hat{\boldsymbol{\beta}}_{+,k,p} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x})(Y_i^p - \mathbf{r}_k(x)'\boldsymbol{\beta})^2,$$

and note that $\hat{\mu}_{-,k}(x) = \hat{\mu}_{-,k,1}(x)$ and $\hat{\mu}_{+,k}(x) = \hat{\mu}_{+,k,1}(x)$ with our notation.

We show in the appendix that the resulting partition-size selectors using the above estimators,

$$\check{J}_{\mathtt{ES},-,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{ES},-}}{\check{\mathscr{V}}_{\mathtt{ES},-}} \right)^{1/3} n^{1/3} \right\rfloor \qquad \text{and} \qquad \check{J}_{\mathtt{ES},+,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{ES},+}}{\check{\mathscr{V}}_{\mathtt{ES},+}} \right)^{1/3} n^{1/3} \right\rfloor, \quad (3.7)$$

are also consistent in the sense of Theorem III.3, under the conditions imposed in that theorem.

Remark III.5 (Density-Weighted IMSE). Taking $w(x) = f(x)$, with $f(x)$ unknown, leads to the simplified constants:

$$\mathscr{V}^{\mathtt{dw}}_{\mathtt{ES},-} = \frac{1}{\bar{x} - x_l} \int_{x_l}^{\bar{x}} \sigma_-^2(x)dx, \qquad \mathscr{V}^{\mathtt{dw}}_{\mathtt{ES},+} = \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \sigma_+^2(x)dx,$$

$$\mathscr{B}^{\mathtt{dw}}_{\mathtt{ES},-} = \frac{(\bar{x} - x_l)^2}{12}\mathbb{E}[\mathbb{1}(X_i < \bar{x})(\mu_-^{(1)}(X_i))^2] \qquad \mathscr{B}^{\mathtt{dw}}_{\mathtt{ES},+} = \frac{(x_u - \bar{x})^2}{12}\mathbb{E}[\mathbb{1}(X_i \geq \bar{x})(\mu_-^{(1)}(X_i))^2].$$

The biases can now be estimated by a simple plug-in procedure,

$$\hat{\mathscr{B}}^{\mathtt{dw}}_{\mathtt{ES},-} = \frac{(\bar{x} - x_l)^2}{12n} \sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x}) \left(\hat{\mu}^{(1)}_{-,k}(X_i) \right)^2, \quad \hat{\mathscr{B}}^{\mathtt{dw}}_{\mathtt{ES},+} = \frac{(x_u - \bar{x})^2}{12n} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x}) \left(\hat{\mu}^{(1)}_{+,k}(X_i) \right)^2.$$

The variances can be estimated employing either (i) spacings estimators,

$$\hat{\mathscr{V}}^{\mathtt{dw}}_{\mathtt{ES},-} = \frac{1}{\bar{x} - x_l}\frac{1}{2} \sum_{i=2}^{N_-} (X_{-,(i)} - X_{-,(i-1)})(Y_{-,[i]} - Y_{-,[i-1]})^2$$

and

$$\hat{\mathscr{V}}^{\texttt{dw}}_{\texttt{ES},+} = \frac{1}{x_u - \bar{x}} \frac{1}{2} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})(Y_{+,[i]} - Y_{+,[i-1]})^2,$$

or (ii) polynomial approximations,

$$\check{\mathscr{V}}^{\texttt{dw}}_{\texttt{ES},-} = \frac{1}{\bar{x} - x_l} \int_{x_l}^{\bar{x}} \hat{\sigma}^2_-(x)dx \qquad \text{and} \qquad \check{\mathscr{V}}^{\texttt{dw}}_{\texttt{ES},+} = \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \hat{\sigma}^2_+(x)dx,$$

using the notation introduced in Remark III.4. The results in the appendix can be used to show that the corresponding partition-size selectors are also consistent in the sense of Theorem III.3.

Computer software implementing all the approaches described above to construct ES-RD plots is available in `R` and `STATA`, as described in Calonico, Cattaneo, and Titiunik (2014e,b).

3.4 Quantile-Spaced RD Plots

In addition to the popular ES-RD plot, we also study an alternative plotting approach based on quantile-spaced (QS) bins. This approach takes into account the sparsity of the data, forcing each bin to have approximately the same number of observations. This feature may be appealing because with QS bins the variability of the local sample means will change across bins only due to nonconstant conditional variances (i.e., due to the presence of heteroskedasticity), but not due to different sample sizes in each bin (as it occurs with an evenly-spaced partition).

In this case, we construct the partitioning scheme as follows:

$$p_{-,j} = \hat{F}^{-1}_- \left(\frac{j}{J_{-,n}} \right) \qquad \text{and} \qquad p_{+,j} = \hat{F}^{-1}_+ \left(\frac{j}{J_{+,n}} \right),$$

with

$$\hat{F}_-^{-1}(y) = \inf\{x : \hat{F}_-(x) \geq y\}, \qquad \hat{F}_-(x) = \frac{1}{N_-}\sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x})\mathbb{1}(X_i \leq x),$$

$$\hat{F}_+^{-1}(y) = \inf\{x : \hat{F}_+(x) \geq y\}, \qquad \hat{F}_+(x) = \frac{1}{N_+}\sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x})\mathbb{1}(X_i \leq x).$$

In words, the QS-RD plot sets $p_{-,j}$ and $p_{+,j}$ to be the approximately $100(j/J_{-,n})$-th quantiles of the subsample $\{X_i : X_i < \bar{x}\}$ and the approximately $100(j/J_{+,n})$-th quantile of the subsample $\{X_i : X_i \geq \bar{x}\}$, respectively. This construction leads to the quantile-spaced partitioning estimators denoted by $\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n})$ and $\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n})$, with now random partitioning schemes denoted by $\mathcal{P}_{\mathtt{QS},-,n}$ and $\mathcal{P}_{\mathtt{QS},+,n}$, respectively.

3.4.1 Optimal Choice of QS Partition Size

We study again the integrated mean-square error loss functions of the QS-based estimators, which in this case are given by

$$\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n}) = \int_{x_l}^{\bar{x}} \mathbb{E}\left[(\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n}) - \mu_-(x))^2 \middle| \mathbf{X}_n\right] w(x)dx$$

and

$$\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n}) = \int_{\bar{x}}^{x_u} \mathbb{E}\left[(\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) - \mu_+(x))^2 \middle| \mathbf{X}_n\right] w(x)dx.$$

The following theorem gives the corresponding asymptotic expansion of the IMSE for the QS-RD plots, which we will use to develop optimal choices of $J_{-,n}$ and $J_{+,n}$.

Theorem III.6. *Suppose Assumption III.1 holds with $S \geq 2$, and $w : [x_l, x_u] \mapsto \mathbb{R}_+$ is continuous.*

(−) If $J_{-,n}\log(J_{-,n})/n \to 0$ and $J_{-,n}/\log(n) \to \infty$, then

$$\mathsf{IMSE}_{\mathtt{QS},-}(J_{n,-}) = \frac{J_{-,n}}{n}\mathscr{V}_{\mathtt{QS},-}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{-,n}^2}\mathscr{B}_{\mathtt{QS},-}\{1 + o_{\mathbb{P}}(1)\},$$

$$\mathscr{V}_{\mathtt{QS},-} = \frac{1}{P_-}\int_{x_l}^{\bar{x}} \sigma_-^2(x)w(x)dx \qquad \textit{and} \qquad \mathscr{B}_{\mathtt{QS},-} = \frac{P_-^2}{12}\int_{x_l}^{\bar{x}} \left(\frac{\mu_-^{(1)}(x)}{f(x)}\right)^2 w(x)dx$$

where $P_- = \mathbb{P}[X_i < \bar{x}]$.

(+) If $J_{+,n}\log(J_{+,n})/n \to 0$ *and* $J_{+,n}/\log(n) \to \infty$*, then*

$$\mathsf{IMSE}_{\mathtt{QS},+}(J_{n,+}) = \frac{J_{+,n}}{n}\mathscr{V}_{\mathtt{QS},+}\{1+o_{\mathbb{P}}(1)\} + \frac{1}{J_{+,n}^2}\mathscr{B}_{\mathtt{QS},+}\{1+o_{\mathbb{P}}(1)\},$$

$$\mathscr{V}_{\mathtt{QS},+} = \frac{1}{P_+}\int_{\bar{x}}^{x_u} \sigma_+^2(x)w(x)dx \qquad \textit{and} \qquad \mathscr{B}_{\mathtt{QS},+} = \frac{P_+^2}{12}\int_{\bar{x}}^{x_u} \left(\frac{\mu_+^{(1)}(x)}{f(x)}\right)^2 w(x)dx,$$

where $P_+ = \mathbb{P}[X_i \geq \bar{x}]$.

The conclusion in this theorem is similar to Theorem III.2, but its proof is different because the estimators are constructed using a random partitioning scheme. The partitioning scheme used in the ES-RD plots ($\mathcal{P}_{\mathtt{ES},-,n}$ and $\mathcal{P}_{\mathtt{ES},+,n}$) requires $J_{-,n} \to \infty$ and $J_{+,n} \to \infty$ but could lead to empty bins in finite samples (this possibility disappears asymptotically; see Lemma C.1 in the appendix). In contrast, the partitioning schemes underlying the QS-RD plots ($\mathcal{P}_{\mathtt{QS},-,n}$ and $\mathcal{P}_{\mathtt{QS},+,n}$) guarantee roughly the same number of observations ($\approx N_-/J_{-,n}$ and $\approx N_+/J_{+,n}$) in each bin. The slightly stronger rate conditions $J_{-,n}/\log(n) \to \infty$ and $J_{+,n}/\log(n) \to \infty$ are imposed to ensure consistency of the sample quantiles functions at the appropriate rate; see Mason (1984) for further details.

For the QS-RD plots, the expansions of $\mathsf{IMSE}_{\mathtt{QS},-}(J_{n,-})$ and $\mathsf{IMSE}_{\mathtt{QS},+}(J_{n,+})$ imply the following optimal choice of partition sizes:

$$J_{\mathtt{QS},-,n} = \left\lfloor \left(\frac{2\mathscr{B}_{\mathtt{QS},-}}{\mathscr{V}_{\mathtt{QS},-}}\right)^{1/3} n^{1/3} \right\rfloor \qquad \text{and} \qquad J_{\mathtt{QS},+,n} = \left\lfloor \left(\frac{2\mathscr{B}_{\mathtt{QS},+}}{\mathscr{V}_{\mathtt{QS},+}}\right)^{1/3} n^{1/3} \right\rfloor \tag{3.8}$$

3.4.2 Data-Driven Implementations of $J_{\mathtt{QS},-,n}$ and $J_{\mathtt{QS},+,n}$

Paralleling the discussion in Section 3.3, we propose consistent estimators for $J_{\mathtt{QS},-,n}$ and $J_{\mathtt{QS},+,n}$ using the idea of spacings estimators, which are simple, more robust and easy-to-implement but require continuous outcomes; Remark III.8 below discusses the case of non-continuous outcomes. We retain all the notation introduced for the implementation of ES-RD plots, including that of control and treatment subsamples, order statistics and their concominats.

Our proposed estimators of the optimal QS partition sizes are the following:

$$\hat{\mathscr{V}}_{\mathtt{QS},-} = \frac{n}{2N_-}\sum_{i=2}^{N_-}(X_{-,(i)} - X_{-,(i-1)})(Y_{-,[i]} - Y_{-,[i-1]})^2 w(\bar{X}_{-,(i)}), \tag{3.9}$$

$$\hat{\mathscr{B}}_{\mathtt{QS},-} = \frac{N_-^2}{72}\sum_{i=2}^{N_-}(X_{-,(i)} - X_{-,(i-1)})^3 \left(\hat{\mu}^{(1)}_{-,k}(\bar{X}_{-,(i)})\right)^2 w(\bar{X}_{-,(i)}), \tag{3.10}$$

and

$$\hat{\mathscr{V}}_{\mathtt{QS},+} = \frac{n}{2N_+}\sum_{i=2}^{N_+}(X_{+,(i)} - X_{+,(i-1)})(Y_{+,[i]} - Y_{+,[i-1]})^2 w(\bar{X}_{+,(i)}), \tag{3.11}$$

$$\hat{\mathscr{B}}_{\mathtt{QS},+} = \frac{N_+^2}{72}\sum_{i=2}^{N_+}(X_{+,(i)} - X_{+,(i-1)})^3 \left(\hat{\mu}^{(1)}_{+,k}(\bar{X}_{+,(i)})\right)^2 w(\bar{X}_{+,(i)}), \tag{3.12}$$

with, as introduced above, $\bar{X}_{-,(i)} = (X_{-,(i)} + X_{-,(i-1)})/2$, $i = 2, 3, \cdots, N_-$, $\hat{\mu}^{(1)}_{-,k}(x) = \mathbf{r}^{(1)}_k(x)'\hat{\boldsymbol{\beta}}_{-,k}$, $\bar{X}_{+,(i)} = (X_{+,(i)} + X_{+,(i-1)})/2$, $i = 1, 2, \cdots, N_+$, $\hat{\mu}^{(1)}_{+,k}(x) = \mathbf{r}^{(1)}_k(x)'\hat{\boldsymbol{\beta}}_{+,k}$, and $\mathbf{r}^{(1)}_k(x) = \partial\mathbf{r}_k(x)/\partial x$.

Therefore, in the QS partitions case, our data-driven, IMSE-optimal selectors take the form:

$$\hat{J}_{\mathtt{QS},-,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{QS},-}}{\hat{\mathscr{V}}_{\mathtt{QS},-}}\right)^{1/3} n^{1/3} \right\rfloor \qquad \text{and} \qquad \hat{J}_{\mathtt{QS},+,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{QS},+}}{\hat{\mathscr{V}}_{\mathtt{QS},+}}\right)^{1/3} n^{1/3} \right\rfloor, \tag{3.13}$$

using the estimators in (3.9)-(3.10) and (3.11)-(3.12), respectively. As in the case of

Theorem III.7 for ES-RD plots, the following theorem shows that these automatic partition-size selectors are nonparametric consistent if the polynomial fits are viewed as nonparametric approximations with $k = k_n \to \infty$.

Theorem III.7. *Suppose Assumption III.1 holds with $S \geq 5$, $w : [x_l, x_u] \mapsto \mathbb{R}_+$ is continuous, and $Y_i(0)$ and $Y_i(1)$ are continuously distributed. If $k_n^7/n \to 0$ and $k_n \to \infty$, then*

$$\frac{\hat{J}_{\mathtt{QS},-,n}}{J_{\mathtt{QS},-,n}} \to_{\mathbb{P}} 1 \qquad \textit{and} \qquad \frac{\hat{J}_{\mathtt{QS},+,n}}{J_{\mathtt{QS},+,n}} \to_{\mathbb{P}} 1.$$

In practice, the choice $w(x) = 1$ is arguably the simplest one, and this is the one we implement by default in our companion software.

Remark III.8 (Non-continuous Outcomes). As mentioned in Remark III.4, the concomitant-based estimation approach cannot be used when $Y_i(0)$ and $Y_i(1)$ are not continuously distributed. For the latter cases, alternatively, we can use the series polynomial estimation approach already introduced above. Assuming that $\mathbb{E}[Y_i(t)^2|X_i = x]$, $t = 0, 1$, are twice continuously differentiable, we may use the following estimators:

$$\check{\mathscr{V}}_{\mathtt{QS},-} = \frac{n}{N_-}\sum_{i=2}^{N_-}(X_{-,(i)} - X_{-,(i-1)})\hat{\sigma}_-^2(\bar{X}_{-,(i)})w(\bar{X}_{-,(i)}),$$

$$\check{\mathscr{V}}_{\mathtt{QS},+} = \frac{n}{N_+}\sum_{i=2}^{N_+}(X_{+,(i)} - X_{+,(i-1)})\hat{\sigma}_-^2(\bar{X}_{+,(i)})w(\bar{X}_{+,(i)}),$$

where $\hat{\sigma}_-^2(x)$ and $\hat{\sigma}_+^2(x)$ are the polynomial approximations discussed in Remark III.4. The corresponding data-driven partition-size selectors in this case are

$$\check{J}_{\mathtt{QS},-,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{QS},-}}{\check{\mathscr{V}}_{\mathtt{QS},-}}\right)^{1/3} n^{1/3} \right\rfloor \qquad \text{and} \qquad \check{J}_{\mathtt{QS},+,n} = \left\lfloor \left(\frac{2\hat{\mathscr{B}}_{\mathtt{QS},+}}{\check{\mathscr{V}}_{\mathtt{QS},+}}\right)^{1/3} n^{1/3} \right\rfloor, \quad (3.14)$$

which we show in the appendix are also consistent in the sense of Theorem III.7, provided the conditions in that theorem hold.

Remark III.9 (Density-Weighted IMSE). Taking $w(x) = f(x)$ in the case of QS-RD Plots leads to the following constants:

$$\mathscr{V}^{\mathtt{dw}}_{\mathtt{QS},-} = \frac{1}{P_-}\mathbb{E}[\mathbb{1}(X_i < \bar{x})\sigma^2_-(X_i)], \qquad \mathscr{V}^{\mathtt{dw}}_{\mathtt{QS},+} = \frac{1}{P_+}\mathbb{E}[\mathbb{1}(X_i \geq \bar{x})\sigma^2_+(X_i)],$$

$$\mathscr{B}^{\mathtt{dw}}_{\mathtt{QS},-} = \frac{P^2_-}{12}\int_{x_l}^{\bar{x}} \frac{\left(\mu^{(1)}_-(x)\right)^2}{f(x)}dx, \qquad \mathscr{B}^{\mathtt{dw}}_{\mathtt{QS},+} = \frac{P^2_+}{12}\int_{\bar{x}}^{x_u} \frac{\left(\mu^{(1)}_+(x)\right)^2}{f(x)}dx.$$

These constants, which are not as simple as in the ES-RD Plot case, may also be estimated using either the spacings approach or the polynomial approximations approach. The results in the appendix can be used to show that the resulting partition-size selectors are consistent in the sense of Theorem III.7, under appropriate assumptions (c.f., Remark III.5).

In Calonico, Cattaneo, and Titiunik (2014e,b) we also describe computer software implementations in `R` and `STATA` of several EQ-RD Plots, using the results discussed above.

3.5 Numerical Results

This section reports numerical evidence on the performance of our proposed methods employing real data from two empirical applications, and data from a Monte Carlo experiment. We also compare numerically the two partitioning schemes studied in this chapter (evenly-spaced and quantile-spaced) in terms of their asymptotic IMSE.

3.5.1 Empirical Illustration

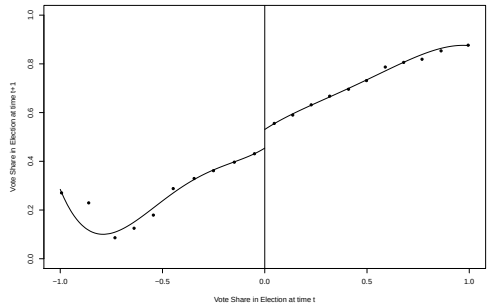
We illustrate our proposed methods using data from two RD empirical applications. We first look at the data from Lee (2008) already mentioned in the Introduction. As previously discussed, Lee studies the incumbency advantage in U.S. House elections; the forcing variable is the margin of victory of the Democratic party in a

given U.S. House election, the threshold is $\bar{x} = 0$, and the outcome variable is the Democratic vote share in the following U.S. House election, which occurs two years later. The unit of observation is the U.S. House district. All U.S. House elections between 1948 and 2008 are included, with the exception of years when district boundaries change; the dataset we employ has a total of $n = 6,558$ complete district-year observations —see Lee (2008) for details.

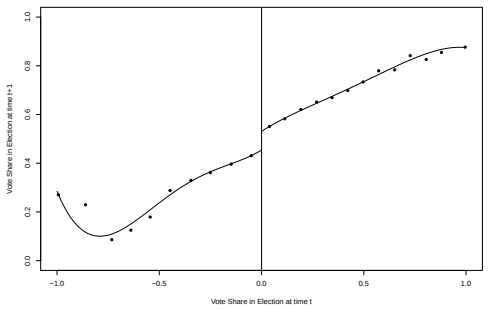
Second, we employ an extract of the dataset constructed by Cattaneo, Frandsen, and Titiunik (2014), who study several measures of incumbency advantage in U.S. Senate elections for the period 1914–2010. In particular, we focus here on the RD effect of the Democratic party winning a U.S. Senate seat on the vote share obtained in the following election for that same seat. This empirical illustration is analogous to the one presented by Lee (2008) for U.S. House elections: the running variable is the state-level margin of victory of the Democratic party in an election for a Senate seat, the threshold is $\bar{x} = 0$ and the outcome is the vote share of the Democratic party in the following election for the same Senate seat in the state, which occurs six years later. The unit of observation is the state, and the data set has a total of $n = 1,297$ state-year complete observations.

The resulting data-driven RD plots using the above empirical illustrations are presented in Figures 3.2 and 3.3, respectively. These figures are constructed using the command/function `rdbinselect` in our companion software packages (Calonico, Cattaneo, and Titiunik, 2014e,b). Using the notation introduced above, the command estimates the number of optimal bins for control and treatment units given in formulas (3.6), (3.7), (3.13) and (3.14), while the global polynomial are constructed using a 4-th degree polynomial (i.e., $\hat{\mu}_{-,4}(x)$ and $\hat{\mu}_{+,4}(x)$). The default bin choices are explicitly constructed to approximate the underlying regression function. As the figures show, the local, binned sample means indeed seem to approximate well the underlying regression function (taking the global polynomial fit as benchmark). It is

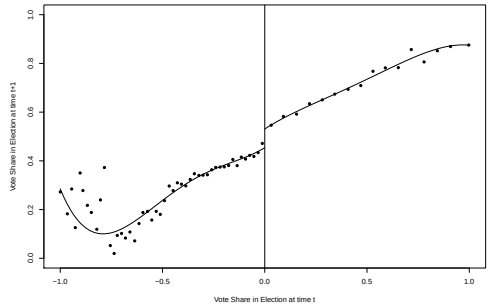
Figure 3.2: Optimal Data-Driven RD Plots for House Elections Data



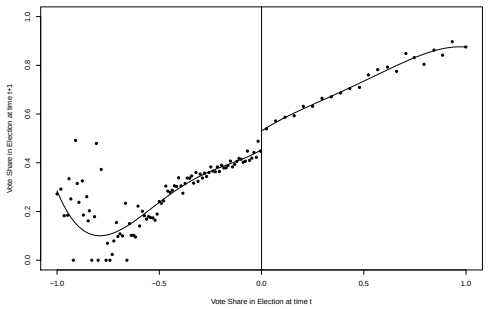
(a) ES RD-Plot, $\hat{J}_{\texttt{ES},-,n} = 10$, $\hat{J}_{\texttt{ES},+,n} = 11$.



(b) ES RD-Plot, $\check{J}_{\texttt{ES},-,n} = 10$, $\check{J}_{\texttt{ES},+,n} = 13$.



(c) QS RD-Plot, $\hat{J}_{\texttt{QS},-,n} = 48$, $\hat{J}_{\texttt{QS},+,n} = 16$.



(d) QS RD-Plot, $\check{J}_{\texttt{QS},-,n} = 95$, $\check{J}_{\texttt{QS},+,n} = 22$.

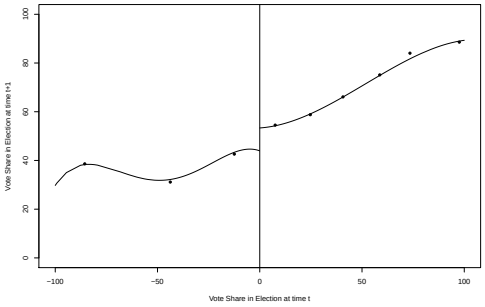
also interesting to note that the QS RD plots tend to have more bins than the ES RD plots in these empirical applications.

3.5.2 Simulations

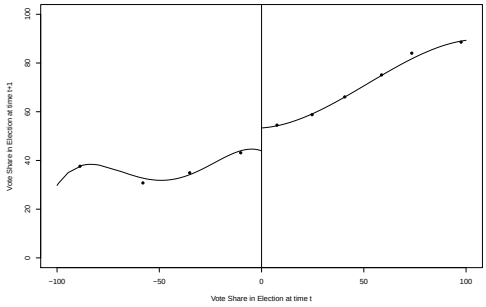
We report the results from a Monte Carlo experiment to study the finite-sample behavior of our proposed methods. We consider several data generating processes, which vary in the distribution of the running variable, the conditional variance and the distribution of the unobserved error term in the regression function.

Specifically, the data is generated as i.i.d. draws, $\{(Y_i, X_i)' : i = 1, 2, ..., n\}$

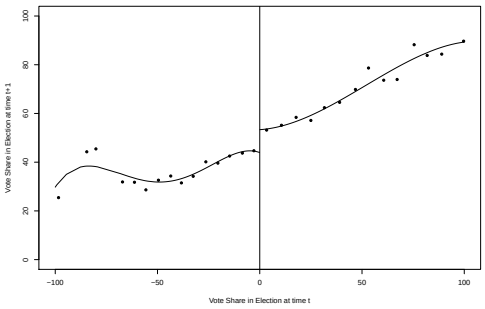
Figure 3.3: Optimal Data-Driven RD Plots for Senate Elections Data



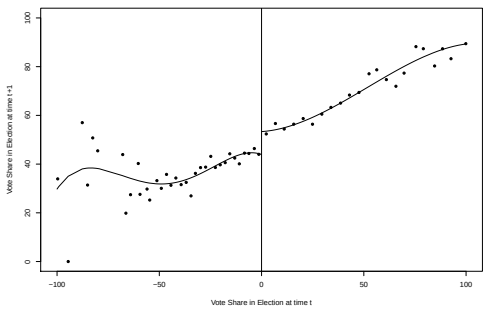
(a) ES RD-Plot, $\hat{J}_{\mathtt{ES},-,n} = 3$, $\hat{J}_{\mathtt{ES},+,n} = 6$.



(b) ES RD-Plot, $\check{J}_{\mathtt{ES},-,n} = 4$, $\check{J}_{\mathtt{ES},+,n} = 6$.

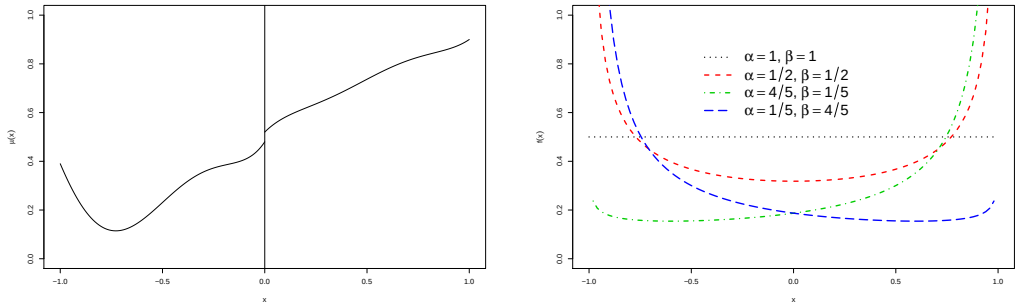


(c) QS RD-Plot, $\hat{J}_{\mathtt{QS},-,n} = 17$, $\hat{J}_{\mathtt{QS},+,n} = 14$.



(d) QS RD-Plot, $\check{J}_{\mathtt{QS},-,n} = 42$, $\check{J}_{\mathtt{QS},+,n} = 22$.

Figure 3.4: Data Generating Processes



(a) Regression function, $\mu(x)$. (b) X_i's distribution, $\mathcal{B}(p_1, p_2)$.

following

$$Y_i = \mu(X_i) + \varepsilon_i, \qquad X_i \sim (2\mathcal{B}(p_1, p_2) - 1), \qquad \varepsilon_i \sim \sigma(X_i)\mathcal{F},$$

where

$$\mu(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{if } x \geq 0 \end{cases},$$

$\mathcal{B}(p_1, p_2)$ denotes a Beta distribution with parameters p_1 and p_2, $\sigma(x)$ is either equal to 1 (homoskedasticity) or equal to $\exp(-|x|/2)$ (heteroskedasticity), and $\mathcal{F}$ is either $\mathcal{N}(0,1)$ or $(\chi_4 - 4)/\sqrt{8}$. The functional form of $\mu(x)$ is obtained by fitting a 5-th order global polynomial with different coefficients for control and treatment units separately using the original data of Lee (2008), after discarding observations with past vote share differences greater than 0.99 and less than -0.99. Figure 3.4 plots the regression function $\mu(x)$ and the two choices for the density of X_i. This simulation setup generalizes the one considered in Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014c).

Our Monte Carlo experiment considers 8 models that combine different choices of (p_1, p_2), $\sigma(x)$ and $\mathcal{F}$, as described in Table 3.1. For each model in Table 3.1, we

Table 3.1: Data Generating Processes

Model	p_1	p_2	$\sigma(x)$	$\mathcal{F}$
1	1	1	1	$\mathcal{N}(0,1)$
2	1/2	1/2	1	$\mathcal{N}(0,1)$
3	1/5	4/5	$\exp(-\lvert x\rvert/2)$	$\mathcal{N}(0,1)$
4	4/5	1/5	$\exp(-\lvert x\rvert/2)$	$\mathcal{N}(0,1)$
5	1/5	4/5	1	$(\chi_4 - 4)/\sqrt{8}$
6	4/5	1/5	1	$(\chi_4 - 4)/\sqrt{8}$
7	1/5	4/5	$\exp(-\lvert x\rvert/2)$	$(\chi_4 - 4)/\sqrt{8}$
8	4/5	1/5	$\exp(-\lvert x\rvert/2)$	$(\chi_4 - 4)/\sqrt{8}$

set $n = 1,000$ and generate $5,000$ simulations to compute the IMSEs of both ES and QS partitioning schemes, with $w(x) = 1$ or $w(x) = f(x)$, and different choices of partition sizes. In each case considered, we also computed the different estimators of the corresponding optimal-partition size introduced in the chapter. [We experimented with other sample sizes and data generating processes and, in all cases, we found qualitative similar results to those reported here.]

The simulation results are presented in Tables C.1-C.8, which corresponds to simulation models 1–8, respectively. Each table includes results for both ES and QS partitioning organized in two distinct panels as follows.

• Panel A: Reports results for normalized IMSEs using both nonrandom partition sizes and estimated partition sizes for ES and QS RD-Plots. All IMSEs are normalized relative to the IMSE evaluated at the optimal partition-size choice. The first part of this panel reports (normalized) IMSEs with $w(x) = 1$ across different values of partition size J_n, for each RD plot, where the grid of J_n is centered at the optimal (infeasible) choice in each case. The second part of the panel reports (normalized) IMSEs with $w(x) = 1$ when the partition size is estimated using either the spacings-based ($\hat{J}_{\cdot,\cdot,n}$) or the polynomial-based ($\check{J}_{\cdot,\cdot,n}$) approaches. Finally, for completeness, the last part of this panel reports the (normalized) IMSE with $w(x) = f(x)$, for both ES and QS RD-plots, when using the optimal (infeasible) partition size as well as when using the estimated spacings-based partition size.

• Panel B: Reports several features of the empirical distribution across simulations of the different estimators of the optimal partition size. Specifically, our simulations consider 6 distinct estimators: (i) spacings-based for ES partitions with $w(x) = 1$ (see (3.6)), (ii) polynomial-based for ES partitions with $w(x) = 1$ (see (3.7)), (iii) spacings-based for ES partitions with $w(x) = f(x)$ (see Remark III.5), (iv) spacings-based for QS partitions with $w(x) = 1$ (see (3.13)), (v) polynomial-based for QS partitions with $w(x) = 1$ (see (3.14)), (vi) spacings-based for QS partitions with $w(x) = f(x)$ (see Remark III.9).

In sum, our Monte Carlo results reported in Tables C.1-C.8 are meant to capture the finite-sample performance of Theorems III.2 and III.6 (Panel A), and the finite-sample performance of Theorems III.3 and III.7 and the other consistency results discussed in the Remarks above (Panel B). In term of actual results, our simulation findings are very encouraging. First, in all cases the IMSEs are minimized at the corresponding optimal choice of partition size, suggesting that the Theorems III.2 and III.6 provide a good finite-sample approximation. Second, in all cases our proposed estimators of the optimal partition size perform quite well, exhibiting a concentrated finite-sample distribution centered at the population optimal choice for partition size. Put together, these results suggest that our proposed optimal data-driven tuning parameter choices for constructing RD-plots perform excellent in samples of moderate size.

3.5.3 Comparison of Partitioning Schemes

We proposed two alternative ways of constructing RD plots, one employing ES partitioning while the other employing QS partitioning. While developing a general theory for optimal partitioning scheme selection is well beyond the scope of this chapter, we can employ our IMSE expansions to compare the two partitioning schemes theoretically in order to assess their relative IMSE-optimality properties. [Here IMSE-

optimality is understood as point estimation optimality in the IMSE sense.]

Without loss of generality we focus on the IMSE for the treatment group ("+" subindex). Assuming the regularity conditions imposed above hold, we obtain (up to the floor operator for selecting the optimal partition sizes):

$$\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n}) = \frac{\sqrt[3]{3}}{4}\mathsf{C}_{\mathtt{ES},+}n^{-2/3}\{1+o_{\mathbb{P}}(1)\}, \quad \mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n}) = \frac{\sqrt[3]{3}}{4}\mathsf{C}_{\mathtt{QS},+}n^{-2/3}\{1+o_{\mathbb{P}}(1)\},$$

where

$$\mathsf{C}_{\mathtt{ES},+} = \left(\int_{\bar{x}}^{x_u} \left(\mu_+^{(1)}(x)\right)^2 w(x)dx\right)^{1/3} \left(\int_{\bar{x}}^{x_u} \frac{\sigma_+^2(x)}{f(x)} w(x)dx\right)^{2/3},$$

$$\mathsf{C}_{\mathtt{QS},+} = \left(\int_{\bar{x}}^{x_u} \left(\frac{\mu_+^{(1)}(x)}{f(x)}\right)^2 w(x)dx\right)^{1/3} \left(\int_{\bar{x}}^{x_u} \sigma_+^2(x)w(x)dx\right)^{2/3}.$$

Thus, in order to compare the performance of the partition-size selectors for ES and QS RD plots we need to compare the two DGP constants $\mathsf{C}_{\mathtt{ES},+}$ and $\mathsf{C}_{\mathtt{QS},+}$. It follows that when $f(x) \propto \kappa$ (i.e., the running variable is uniformly distributed), then $\mathsf{C}_{\mathtt{ES},+} = \mathsf{C}_{\mathtt{QS},+}$ and therefore both partitioning schemes have equal (asymptotic) IMSE when the corresponding optimal partition size is used. Unfortunately, when the density $f(x)$ is not constant on the support $[x_l, x_u]$ it is not possible to obtain a unique ranking between $\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})$ and $\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})$. Heuristically, the QS RD-plots should perform better in cases where the data is sparse because the estimated quantile-spaced partition should adapt to this situation better, but we have been unable to provide a formal ranking along these lines.

Nonetheless, in Table C.9 we explore the ranking between the two partitioning schemes using the eight data generating processes discussed in our simulation study (Table 3.1). As expected, this table shows that when $f(x)$ is uniform both IMSE are equal, while when $f(x)$ is not uniform either IMSE may dominate the other. This depends on the shape of the regression function (different for control and treat-

ment sides) and conditional heteroskedasticity in the underlying true data generating process.

3.6 Conclusions

This chapter introduced several optimal data-driven partition-size selectors for RD Plots, focusing on both the popular and commonly used evenly-spaced RD plot and also on an alternative quantile-space RD plot. The resulting selectors lead to practical RD plots that are constructed in an automatic and objective way using the available data. More generally, they also provide a benchmark for empirical work employing RD plots: because the IMSE-optimal choices of number of bins are obtained balancing (integrated) squared-bias and variance of a partitioning estimator of the underlying conditional expectations, empirical researcher may use the selectors presented in this chapter to construct undersmoothed (more bins) or oversmoothed (fewer bins) RD plots.

In addition to improve the construction of RD plots in empirical applications, the selectors introduced in this chapter could be used to conduct data-driven inference in the RD design. Using the data-driven bins, it is possible to construct confidence intervals and testing procedures for interesting hypotheses concerning the underlying regression functions $\mu_-(x)$ and $\mu_+(x)$. For example, as discussed in Imbens and Lemieux (2008), it is possible to use the partitioning estimation approach together with the optimal partition size selectors introduced herein to test for "discontinuities" of $\mu_-(x)$ and $\mu_+(x)$ as a form of falsification test of the RD design. In research underway we plan to investigate this and related inference problems employing partitioning estimation for RD designs.

APPENDICES

APPENDIX A

Appendix to Chapter 1

A.1 Appendix Chapter 1

Proof Theorem I. Let

$$\delta_{ics}(p,z) = \int \frac{\partial m(p,z,\mu,\varepsilon)}{\partial p} \mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{ics},Z_{ics}}(\mu,\varepsilon|p,z)$$

Next, note that:

$$\begin{aligned}
\frac{\partial \mathbb{E}(Y_{ics}|P_{cs},Z_{ics},V_s)}{\partial P_{cs}} &= \frac{\partial}{\partial p}\left[\int m(p,z,\mu,\varepsilon)\,\mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics},V_s}(\mu,\varepsilon|p,z,v)\right] \\
&= \int \frac{\partial}{\partial p}\left[m(p,z,\mu,\varepsilon)\,\mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics},V_s}(\mu,\varepsilon|p,z,v)\right] \\
&= \int \frac{\partial}{\partial p}\left[m(p,z,\mu,\varepsilon)\,\mathbf{d}F_{\varepsilon_{ics}|\mu_s,P_{cs},Z_{ics},V_s}(\varepsilon|\mu,p,z,v)\,\mathbf{d}F_{\mu_s|P_{cs},Z_{ics},V_s}(\mu|p,z,v)\right] \\
&= \int \frac{\partial}{\partial p}\left[m(p,z,\mu,\varepsilon)\,\mathbf{d}F_{\varepsilon_{ics}|\mu_s,Z_{ics},V_s}(\varepsilon|\mu,z,v)\,\mathbf{d}F_{\mu_s|Z_{ics},V_s}(\mu|z,v)\right] \\
&= \int \frac{\partial m(p,z,\mu,\varepsilon)}{\partial p}\mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics},V_s}(\mu,\varepsilon|p,z,v) \\
&\quad + \underbrace{\int m(p,z,\mu,\varepsilon)\frac{\partial}{\partial p}\left[\mathbf{d}F_{\varepsilon_{ics}|\mu_s,Z_{ics},V_s}(\varepsilon|\mu,z,v)\,\mathbf{d}F_{\mu_s|Z_{ics},V_s}(\mu|z,v)\right]}_{=0} \\
&= \int \frac{\partial m(p,z,\mu,\varepsilon)}{\partial p}\mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics},V_s}(\mu,\varepsilon|p,z,v)
\end{aligned}$$

Then,

$$\begin{aligned}\int \frac{\partial \mathbb{E}\left(Y_{ics}|P_{cs},Z_{ics},V_s\right)}{\partial P_{cs}}\mathbf{d}F_{V_s|P_{cs},Z_{ics}}\left(v|p,z\right) &= \int \frac{\partial m\left(p,z,\mu,\varepsilon\right)}{\partial p}\mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics},V_s}\left(\mu,\varepsilon|p,z,v\right)\mathbf{d}F_{V_s|} \\ &= \int \frac{\partial m\left(p,z,\mu,\varepsilon\right)}{\partial p}\mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics}}\left(\mu,\varepsilon|p,z\right) \\ &= \delta_{ics}\left(p,z\right)\end{aligned}$$

□

From this result, identification of the density weighted average derivatives follows directly:

$$\begin{aligned}\delta_{ics}^{\omega} &= \int\int \frac{\partial \mathbb{E}\left(Y_{ics}|P_{cs},Z_{ics},V_s\right)}{\partial P_{ics}}\omega(p,z)\mathbf{d}F_{V_s|P_{cs},Z_{ics}}\left(v|p,z\right)\mathbf{d}F_{P_{cs},,Z_{ics}}\left(p,v\right) \\ &= \mathbb{E}\left[\frac{\partial \mathbb{E}\left(Y_{ics}|P_{cs},Z_{ics},V_s\right)}{\partial P_{ics}}\omega(p,z)\right]\end{aligned}$$

Proof Theorem II. Let

$$\begin{aligned}\Delta\left(p'',p'\right) &= \int \left[m\left(p'',z,\mu,\varepsilon\right)-m\left(p',z,\mu,\varepsilon\right)\right]\mathbf{d}F_{Z_{ics},\mu_s,\varepsilon_{ics}|P_{cs}}\left(z,\mu,\varepsilon|p'\right) \\ &\equiv A-B\end{aligned}$$

First, note that

$$B=\int m\left(p',z,\mu,\varepsilon\right)\mathbf{d}F_{Z_{ics},\mu_s,\varepsilon_{ics}|P_{cs}}\left(z,\mu,\varepsilon|p'\right)=\mathbb{E}\left(Y_{ics}|P_{cs}=p'\right)$$

and so is identified directly from the data. For the other term:

$$\begin{aligned}
A &= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{Z_{ics},\mu_s,\varepsilon_{ics}|P_{cs}}\left(z, \mu, \varepsilon|p'\right) \\
&= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{Z_{ics},\mu_s,\varepsilon_{ics}|P_{cs},V_s}\left(v, z, \mu, \varepsilon|p'\right) \mathbf{d}F_{V_s|P_{cs}}\left(v|p'\right) \\
&= \underbrace{\int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p', v, z\right)}_{(1)} \mathbf{d}F_{Z_{ics},|P_{cs},V_s}\left(z|p', v\right) \mathbf{d}F_{V_s|P_{cs}}\left(v|p'\right)
\end{aligned}$$

where

$$\begin{aligned}
(1) &= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p', v, z\right) \\
&= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{\varepsilon_{ics}|\mu_s,P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p', v, z\right) \mathbf{d}F_{\mu_s|P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p', v, z\right) \\
&= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{\varepsilon_{ics}|\mu_s,V_s,Z_{ics}}\left(\mu, \varepsilon|v, z\right) \mathbf{d}F_{\mu_s|V_s,Z_{ics}}\left(\mu, \varepsilon|v, z\right) \\
&= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{\varepsilon_{ics}|\mu_s,P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p'', v, z\right) \mathbf{d}F_{\mu_s|P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p'', v, z\right) \\
&= \int m\left(p'', z, \mu, \varepsilon\right) \mathbf{d}F_{\mu_s,\varepsilon_{ics}|P_{cs},V_s,Z_{ics}}\left(\mu, \varepsilon|p'', v, z\right) \\
&= \mathbb{E}\left(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v\right)
\end{aligned}$$

so, finally

$$\begin{aligned}
A &= \int \mathbb{E}\left(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v\right) \mathbf{d}F_{Z_{ics},|P_{cs},V_s}\left(z|p', v\right) \mathbf{d}F_{V_s|P_{cs}}\left(v|p'\right) \\
&= \int \mathbb{E}\left(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v\right) \mathbf{d}F_{Z_{ics},V_s|P_{cs}}\left(z, v|p'\right)
\end{aligned}$$

□

Proof Theorem III. Let $\phi(y)$ be any function of y. Then,

$$\begin{aligned}\mathbb{E}\left[\phi(Y_{ics}^*)\right] &= \mathbb{E}\left[\mathbb{E}\left[\phi(m\left(P_{ics}^*, Z_{ics}, \mu_s, \varepsilon_{ics}\right))|P_{ics}^*, Z_{ics}\right]\right] \\ &= \int \underbrace{\int \phi(m\left(p,z,\mu,\varepsilon\right))dF_{\mu_s,\varepsilon_{ics}|P_{ics}^*,Z_{ics}}\left(\mu,\varepsilon|p,z\right)}_{(1)} dF_{P_{cs}^*,Z_{ics}}\left(p,z\right)\end{aligned}$$

where

$$(1) = \int \underbrace{\int \phi(m\left(p,z,\mu,\varepsilon\right))dF_{\mu_s,\varepsilon_{ics}|P_{ics}^*,Z_{ics},V_s}\left(\mu,\varepsilon|p,z,v\right)}_{(2)} dF_{V_s|P_{ics}^*,Z_{ics}}\left(v|p,z\right)$$

and

$$\begin{aligned}(2) &= \int \phi(m\left(p,z,\mu,\varepsilon\right))dF_{\varepsilon_{ics}|\mu_s,P_{ics}^*,Z_{ics},V_s}\left(\varepsilon|\mu,p,z,v\right) dF_{\mu_s|P_{ics}^*,Z_{ics},V_s}\left(\mu|p,z,v\right) \\ &= \int \phi(m\left(p,z,\mu,\varepsilon\right))dF_{\varepsilon_{ics}|\mu_s,Z_{ics},V_s}\left(\varepsilon|\mu,z,v\right) dF_{\mu_s|Z_{ics},V_s}\left(\mu|z,v\right) \\ &= \int \phi(m\left(p,z,\mu,\varepsilon\right))dF_{\varepsilon_{ics}|\mu_s,P_{ics},Z_{ics},V_s}\left(\varepsilon|\mu,p,z,v\right) dF_{\mu_s|P_{ics},Z_{ics},V_s}\left(\mu|p,z,v\right) \\ &= \int \phi(m\left(p,z,\mu,\varepsilon\right))dF_{\mu_s,\varepsilon_{ics}|P_{cs},Z_{ics},V_s}\left(\mu,\varepsilon|p,z,v\right) = \mathbb{E}\left[Y_{ics}|P_{ics},Z_{ics},V_s\right]\end{aligned}$$

Then, finally

$$\begin{aligned}\mathbb{E}\left[\phi(Y_{ics}^*)\right] &= \int\int \mathbb{E}\left[\phi(Y_{ics})|P_{ics},Z_{ics},V_s\right] dF_{V_s|P_{ics}^*,Z_{ics}}\left(v|p,z\right) dF_{P_{cs}^*,Z_{ics}}\left(p,z\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[\phi(Y_{ics})|P_{ics},Z_{ics},V_s\right]\right]\end{aligned}$$

□

APPENDIX B

Appendix to Chapter 2

B.1 Appendix Chapter 2

In this appendix we summarize our main results for arbitrary order of local polynomials. Here p denotes the order of main RD estimator, while q denotes the order in the bias correction. All the results stated in this appendix are proven in the online supplemental appendix (Calonico, Cattaneo, and Titiunik (2014d)).

B.1.1 Local Polynomial Estimators and Other Notation

For $\nu, p \in \mathbb{N}$ with $\nu \leq p$, the p-th order local polynomial estimators of the ν-th order derivatives $\mu_{Y+}^{(\nu)}$ and $\mu_{Y-}^{(\nu)}$ are $\hat{\mu}_{Y+,p}^{(\nu)}(h_n) = \nu! e_\nu' \hat{\beta}_{Y+,p}(h_n)$ and $\hat{\mu}_{Y-,p}^{(\nu)}(h_n) = \nu! e_\nu' \hat{\beta}_{Y-,p}(h_n)$, where

$$\hat{\beta}_{Y+,p}(h_n) = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \mathbf{1}(X_i \geq 0)(Y_i - r_p(X_i)'\beta)^2 K_{h_n}(X_i)$$

$$\hat{\beta}_{Y-,p}(h_n) = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \mathbf{1}(X_i < 0)(Y_i - r_p(X_i)'\beta)^2 K_{h_n}(X_i)$$

where $r_p(x) = (1, x, \cdots, x^p)'$, e_ν is the conformable $(\nu+1)$-th unit vector (e.g., $e_1 = (0,1,0)'$ if $p = 2$), $K_h(u) = K(u/h)/h$, and h_n is a positive bandwidth sequence. (We drop the evaluation point of functions at $\bar{x} = 0$ to simplify notation.) Let $Y = (Y_1, \cdots, Y_n)'$, $X_p(h) = [r_p(X_1/h), \cdots, r_p(X_n/h)]'$, $S_p(h) = [(X_1/h)^p, \cdots, (X_n/h)^p]'$, $W_+(h) = diag(\mathbf{1}(X_1 \geq 0)K_h(X_1), \cdots, \mathbf{1}(X_n \geq 0)K_h(X_n))$, $W_-(h) = diag(\mathbf{1}(X_1 < 0)K_h(X_1), \cdots, \mathbf{1}(X_n < 0)K_h(X_n))$, $\Gamma_{+,p}(h) = X_p(h)'W_+(h)X_p(h)/n$ and $\Gamma_{-,p}(h) = X_p(h)'W_-(h)X_p(h)/n$, with $diag(a_1, ..., a_n)$ denoting the $(n \times n)$ diagonal matrix with diagonal elements $a_1, ..., a_n$. It follows that

$$\hat{\beta}_{Y+,p}(h_n) = H_p(h_n)\Gamma_{+,p}^{-1}(h_n)X_p(h_n)'W_+(h_n)Y/n$$

and

$$\hat{\beta}_{Y-,p}(h_n) = H_p(h_n)\Gamma_{+,p}^{-1}(h_n)X_p(h_n)'W_+(h_n)Y/n$$

with $H_p(h) = diag(1, h^{-1}, \cdots, h^{-p})$. We set $\hat{\mu}_{Y+,p}(h_n) = \hat{\mu}^{(0)}_{Y+,p}(h_n)$ and $\hat{\mu}_{Y+,p}(h_n) = \hat{\mu}^{(0)}_{Y+,p}(h_n)$ and, whenever possible, we also drop the outcome variable subindex notation. Under conditions given below,

$$\hat{\beta}_{+,p}(h_n) \to_p \beta_{+,p} = (\mu_+, \mu_+^{(1)}/1!, \mu_+^{(2)}/2!, \cdots, \mu_+^{(p)}/p!)'$$

$$\hat{\beta}_{-,p}(h_n) \to_p \beta_{-,p} = (\mu_-, \mu_-^{(1)}/1!, \mu_-^{(2)}/2!, \cdots, \mu_-^{(p)}/p!)'$$

implying that local polynomial regression estimates consistently the level of the unknown regression function (μ_+ and μ_-) as well as its first p derivatives (up to a known scale).

We also employ the following notation: $\vartheta_{+,p,q}(h) = X_p(h)'W_+(h)S_q(h)/n$ and $\vartheta_{-,p,q}(h) = X_p(h)'W_-(h)S_q(h)/n$, and $\Psi_{UV+,p,q}(h,b) = X_p(h)'W_+(h)\Sigma_{UV}W_+(b)X_q(b)/n$ and $\Psi_{UV-,p,q}(h,b) = X_p(h)'W_-(h)\Sigma_{UV}W_-(b)X_q(b)/n$ with $\Sigma_{UV} = diag(\sigma^2_{UV}(X_1), \cdots, \sigma^2_{UV}(X_n))$ with $\sigma^2_{UV}(X_i) = \mathbb{C}ov[U_i, V_i|X_i]$, where U and V placeholders for Y or T. We set

$\Psi_{UV+,p}(h) = \Psi_{UV+,p,p}(h,h)$ and $\Psi_{UV-,p}(h) = \Psi_{UV-,p,p}(h,h)$ for brevity, and drop the outcome variable subindex notation whenever possible.

B.1.2 Sharp RD Designs

As in the main text, in this section we drop the notational dependence on the outcome variable Y. The general estimand is $\tau_\nu = \mu_+^{(\nu)} - \mu_-^{(\nu)}$ with $\mu_+^{(\nu)} = \nu! e_\nu' \beta_{+,p}$ and $\mu_-^{(\nu)} = \nu! e_\nu' \beta_{-,p}$, $\nu \leq p$. Recall that $\tau_{\texttt{SRD}} = \tau_0$ and $\tau_{\texttt{SKRD}} = \tau_1$. For any $\nu \leq p$, the conventional p-th order local polynomial RD estimator is $\hat{\tau}_{\nu,p}(h_n) = \hat{\mu}_{+,p}^{(\nu)}(h_n) - \hat{\mu}_{-,p}^{(\nu)}(h_n)$ with $\hat{\mu}_{+,p}^{(\nu)}(h_n) = \nu! e_\nu' \hat{\beta}_{+,p}(h_n)$ and $\hat{\mu}_{-,p}^{(\nu)}(h_n) = \nu! e_\nu' \hat{\beta}_{-,p}(h_n)$. Recall that $\hat{\tau}_{\texttt{SRD}}(h_n) = \hat{\tau}_{0,1}(h_n)$ and $\hat{\tau}_{\texttt{SKRD}}(h_n) = \hat{\tau}_{1,2}(h_n)$.

The following lemma describes the asymptotic bias, variance and distribution of $\hat{\tau}_{\nu,p}(h_n)$.

Lemma B.1. *Suppose Assumptions III.1–II.3 hold with $S \geq p+2$, and $nh_n \to \infty$. Let $r \in \mathbb{N}$ and $\nu \leq p$.*

(B) *If $h_n \to 0$, then $\mathbb{E}[\hat{\tau}_{\nu,p}(h_n)|\mathcal{X}_n] = \tau_\nu + h_n^{p+1-\nu}\mathsf{B}_{\nu,p,p+1}(h_n) + h_n^{p+2-\nu}\mathsf{B}_{\nu,p,p+2}(h_n) + o_p(h_n^{p+2-\nu})$, where*

$$\begin{aligned}
\mathsf{B}_{\nu,p,r}(h_n) &= \mu_+^{(r)}\mathcal{B}_{+,\nu,p,r}(h_n)/r! - \mu_-^{(r)}\mathcal{B}_{-,\nu,p,r}(h_n)/r! \\
\mathcal{B}_{+,\nu,p,r}(h_n) &= e_\nu'\Gamma_{+,p}^{-1}(h_n)\vartheta_{+,p,r}(h_n) = e_\nu'\Gamma_p^{-1}\vartheta_{p,r} + o_p(1) \\
\mathcal{B}_{-,\nu,p,r}(h_n) &= e_\nu'\Gamma_{-,p}^{-1}(h_n)\vartheta_{-,p,r}(h_n) = (-1)^{\nu+r}e_\nu'\Gamma_p^{-1}\vartheta_{p,r} + o_p(1)
\end{aligned}$$

(V) *If $h_n \to 0$, then $\mathsf{V}_{\nu,p}(h_n) = \mathbb{V}[\hat{\tau}_{\nu,p}(h_n)|\mathcal{X}_n] = \mathcal{V}_{+,\nu,p}(h_n) + \mathcal{V}_{-,\nu,p}(h_n)$ with:*

$$\begin{aligned}
\mathcal{V}_{+,\nu,p}(h_n) &= n^{-1}h_n^{-2\nu}\nu!^2 e_\nu'\Gamma_{+,p}^{-1}(h_n)\Psi_{+,p}(h_n)\Gamma_{+,p}^{-1}(h_n)e_\nu \\
&= n^{-1}h_n^{-1-2\nu}\sigma_+^2\nu!^2 e_\nu'\Gamma_p^{-1}\Psi_p\Gamma_p^{-1}e_\nu/f\ \{1+o_p(1)\} \\
\mathcal{V}_{-,\nu,p}(h_n) &= n^{-1}h_n^{-2\nu}\nu!^2 e_\nu'\Gamma_{-,p}^{-1}(h_n)\Psi_{-,p}(h_n)\Gamma_{-,p}^{-1}(h_n)e_\nu \\
&= n^{-1}h_n^{-1-2\nu}\sigma_-^2\nu!^2 e_\nu'\Gamma_p^{-1}\Psi_p\Gamma_p^{-1}e_\nu/f\ \{1+o_p(1)\}
\end{aligned}$$

(D) *If* $nh_n^{2p+5} \to 0$, *then* $(\hat{\tau}_{\nu,p}(h_n) - \tau_\nu - h_n^{p+1-\nu}\mathsf{B}_{\nu,p,p+1}(h_n))/\sqrt{\mathsf{V}_{\nu,p}(h_n)} \to_d \mathcal{N}(0,1)$.

A q-th order ($p < q$) local polynomial bias-corrected estimator is

$$\hat{\tau}^{\mathsf{bc}}_{\nu,p,q}(h_n, b_n) \;=\; \hat{\tau}_p(h_n) - h_n^{p+1}\hat{\mathsf{B}}_{\nu,p,q}(h_n, b_n)$$

with $\hat{\mathsf{B}}_{\nu,p,q}(h_n, b_n) = (e'_{p+1}\hat{\beta}_{+,q}(b_n))\mathcal{B}_{+,\nu,p,p+1}(h_n) - (e'_{p+1}\hat{\beta}_{-,q}(b_n))\mathcal{B}_{-,\nu,p,p+1}(h_n)$. The following theorem gives the asymptotic bias, variance and distribution of $\hat{\tau}^{\mathsf{bc}}_{\nu,p,q}(h_n, b_n)$. Theorems II.4 and II.13 are special cases with $(\nu, p, q) = (0, 1, 2)$ and $(\nu, p, q) = (1, 2, 3)$, respectively.

Theorem B.2. *Suppose Assumptions III.1–II.3 hold with* $S \geq q+1$, *and* $n \min\{h_n, b_n\} \to \infty$. *Let* $\nu \leq p < q$.

(B) *If* $\max\{h_n, b_n\} \to 0$, *then*

$$\mathbb{E}[\hat{\tau}^{\mathsf{bc}}_{\nu,p,q}(h_n, b_n)|\mathcal{X}_n] = \tau + h_n^{p+2-\nu}\mathsf{B}_{\nu,p,p+2}(h_n)\,\{1+o_p(1)\} - h_n^{p+1-\nu}b_n^{q-p}\mathsf{B}^{\mathsf{bc}}_{\nu,p,q}(h_n, b_n)\,\{1+o_p(1)\}$$

where

$$\mathsf{B}^{\mathsf{bc}}_{\nu,p,q}(h, b) = [\mu_+^{(q+1)}\mathcal{B}_{+,p+1,q,q+1}(b)\mathcal{B}_{+,\nu,p,p+1}(h) - \mu_-^{(q+1)}\mathcal{B}_{-,p+1,q,q+1}(b)\mathcal{B}_{-,\nu,p,p+1}(h)]/[(q+1)!(p+1)!]$$

(V) $\mathsf{V}^{\mathsf{bc}}_{\nu,p,q}(h_n, b_n) = \mathbb{V}[\hat{\tau}^{\mathsf{bc}}_{\nu,p,q}(h_n, b_n)|\mathcal{X}_n] = \mathcal{V}^{\mathsf{bc}}_{+,\nu,p,q}(h_n, b_n) + \mathcal{V}^{\mathsf{bc}}_{-,\nu,p,q}(h_n, b_n)$ *with*

$$
\begin{aligned}
\mathcal{V}^{\mathtt{bc}}_{+,\nu,p,q}(h,b) &= \mathcal{V}_{+,\nu,p}(h) \\
&- 2h^{p+1-\nu}\mathcal{C}_{+,\nu,p,q}(h,b)\mathcal{B}_{+,\nu,p,p+1}(h)/(p+1)! \\
&+ h^{2(p+1-\nu)}\mathcal{V}_{+,p+1,q}(b)\mathcal{B}^2_{+,\nu,p,p+1}(h)/(p+1)!^2 \\
\mathcal{V}^{\mathtt{bc}}_{-,\nu,p,q}(h,b) &= \mathcal{V}_{-,\nu,p}(h) \\
&- 2h^{p+1-\nu}\mathcal{C}_{-,\nu,p,q}(h,b)\mathcal{B}_{-,\nu,p,p+1}(h)/(p+1)! \\
&+ h^{2(p+1-\nu)}\mathcal{V}_{-,p+1,q}(b)\mathcal{B}^2_{-,\nu,p,p+1}(h)/(p+1)!^2
\end{aligned}
$$

$\mathcal{C}_{+,\nu,p,q}(h,b) = n^{-1}h^{-\nu}b^{-p-1}\nu!(p+1)!e_\nu'\Gamma^{-1}_{+,p}(h)\Psi_{+,p,q}(h,b)\Gamma^{-1}_{+,q}(b)e_{p+1}$,
$\mathcal{C}_{-,\nu,p,q}(h,b) = n^{-1}h^{-\nu}b^{-p-1}\nu!(p+1)!e_\nu'\Gamma^{-1}_{-,p}(h)\Psi_{-,p,q}(h,b)\Gamma^{-1}_{-,q}(b)e_{p+1}$.

(D) *If* $n\min\{h_n^{2p+3}, b_n^{2p+3}\}\max\{h_n^2, b_n^{2(q-p)}\} \to 0$, *and* $\kappa\max\{h_n, b_n\} < \kappa_0$, *then* $T^{\mathtt{rbc}}_{\nu,p,q}(h_n,b_n) = (\hat{\tau}^{\mathtt{bc}}_{\nu,p,q}(h_n,b_n) - \tau_\nu)/\sqrt{\mathsf{V}^{\mathtt{bc}}_{\nu,p,q}(h_n,b_n)} \to_d \mathcal{N}(0,1)$.

Thus, $\mathsf{V}^{\mathtt{bc}}_{\mathtt{SRD}}(h_n,b_n) = \mathsf{V}^{\mathtt{bc}}_{0,1,2}(h_n,b_n)$ and $\mathsf{V}^{\mathtt{bc}}_{\mathtt{SKRD}}(h_n,b_n) = \mathsf{V}^{\mathtt{bc}}_{1,2,3}(h_n,b_n)$ in Theorems II.4 and II.13, respectively.

B.1.3 Fuzzy RD Designs

The ν-th fuzzy RD estimand is $\varsigma_\nu = \tau_{Y,\nu}/\tau_{T,\nu}$ with $\tau_{Y,\nu} = \mu^{(\nu)}_{Y+} - \mu^{(\nu)}_{Y-}$ and $\tau_{T,\nu} = \mu^{(\nu)}_{T+} - \mu^{(\nu)}_{T-}$, provided that $\nu \leq S$. Note that $\tau_{\mathtt{FRD}} = \varsigma_0$ and $\tau_{\mathtt{FKRD}} = \varsigma_1$. The fuzzy RD estimator based on the p-th order local polynomial estimators $\hat{\tau}_{Y,\nu,p}(h_n)$ and $\hat{\tau}_{T,\nu,p}(h_n)$ is $\hat{\varsigma}_{\nu,p}(h_n) = \hat{\tau}_{Y,\nu,p}(h_n)/\hat{\tau}_{T,\nu,p}(h_n)$ with $\hat{\tau}_{Y,\nu,p}(h_n) = \hat{\mu}^{(\nu)}_{Y+,p}(h_n) - \hat{\mu}^{(\nu)}_{Y-,p}(h_n)$ and $\hat{\tau}_{T,\nu,p}(h_n) = \hat{\mu}^{(\nu)}_{T+,p}(h_n) - \hat{\mu}^{(\nu)}_{T-,p}(h_n)$, where $\hat{\mu}^{(\nu)}_{Y+,p}(h_n) = \nu!e_\nu'\hat{\beta}_{Y+,p}(h_n)$, $\hat{\mu}^{(\nu)}_{Y-,p}(h_n) = \nu!e_\nu'\hat{\beta}_{Y-,p}(h_n)$, $\hat{\mu}^{(\nu)}_{T+,p}(h_n) = \nu!e_\nu'\hat{\beta}_{T+,p}(h_n)$ and $\hat{\mu}^{(\nu)}_{T-,p}(h_n) = \nu!e_\nu'\hat{\beta}_{T-,p}(h_n)$. Note that $\hat{\tau}_{\mathtt{FRD}}(h_n) = \hat{\varsigma}_{0,1}(h_n)$ and $\hat{\tau}_{\mathtt{FKRD}}(h_n) = \hat{\varsigma}_{1,2}(h_n)$.

The following lemma gives an analogue of Lemma B.1 for fuzzy RD designs. Note that $\hat{\varsigma}_{\nu,p}(h_n) - \varsigma_\nu = \tilde{\varsigma}_{\nu,p}(h_n) + R_n$ with $\tilde{\varsigma}_{\nu,p}(h_n) = (\hat{\tau}_{Y,\nu,p}(h_n) - \tau_{Y,\nu})/\tau_{T,\nu} - \tau_{Y,\nu}(\hat{\tau}_{T,\nu,p}(h_n) -$

$\tau_{T,\nu})/\tau_{T,\nu}^2$ and $R_n = \tau_{Y,\nu}(\hat{\tau}_{T,\nu,p}(h_n)-\tau_{T,\nu})^2/(\tau_{T,\nu}^2\hat{\tau}_{T,\nu,p}(h_n))-(\hat{\tau}_{Y,\nu,p}(h_n)-\tau_{Y,\nu})(\hat{\tau}_{T,\nu,p}(h_n)-\tau_{T,\nu})/(\tau_{T,\nu}\hat{\tau}_{T,\nu,p}(h_n))$.

Lemma B.3. *Suppose Assumptions III.1–II.14 hold with $S \geq p+2$, and $nh_n \to \infty$. Let $r \in \mathbb{N}$ and $\nu \leq p$.*

(R) *If $h_n \to 0$ and $nh_n^{1+2\nu} \to \infty$, then $R_n = O_p(n^{-1}h_n^{-1-2\nu} + h_n^{2p+2-2\nu})$.*

(B) *If $h_n \to 0$, then $\mathbb{E}[\tilde{\varsigma}_{\nu,p}(h_n)|\mathcal{X}_n] = h_n^{p+1-\nu}\mathsf{B}_{\mathtt{F},\nu,p,p+1}(h_n) + h_n^{p+2-\nu}\mathsf{B}_{\mathtt{F},\nu,p,p+2}(h_n) + o_p(h_n^{p+2-\nu})$, where*

$\mathsf{B}_{\mathtt{F},\nu,p,r}(h_n) = \mathsf{B}_{Y,\nu,p,r}(h_n)/\tau_{T,\nu} - \tau_{Y,\nu}\mathsf{B}_{T,\nu,p,r}(h_n)/\tau_{T,\nu}^2$,

$\mathsf{B}_{Y,\nu,p,r}(h_n) = \mu_{Y+}^{(r)}\mathcal{B}_{+,\nu,p,r}(h_n)/r! - \mu_{Y-}^{(r)}\mathcal{B}_{-,\nu,p,r}(h_n)/r!$,

$\mathsf{B}_{T,\nu,p,r}(h_n) = \mu_{T+}^{(r)}\mathcal{B}_{+,\nu,p,r}(h_n)/r! - \mu_{T-}^{(r)}\mathcal{B}_{-,\nu,p,r}(h_n)/r!$.

(V) *If $h_n \to 0$, then $\mathsf{V}_{\mathtt{F},\nu,p}(h_n) = \mathbb{V}[\tilde{\varsigma}_{\nu,p}(h_n)|\mathcal{X}_n] = \mathsf{V}_{\mathtt{F},+,\nu,p}(h_n) + \mathsf{V}_{\mathtt{F},-,\nu,p}(h_n)$ with*

$\mathsf{V}_{\mathtt{F},+,\nu,p}(h_n) = (1/\tau_{T,\nu}^2)\mathcal{V}_{YY+,\nu,p}(h_n) - (2\tau_{Y,\nu}/\tau_{T,\nu}^3)\mathcal{V}_{YT+,\nu,p}(h_n) + (\tau_{Y,\nu}^2/\tau_{T,\nu}^4)\mathcal{V}_{TT+,\nu,p}(h_n)$,

$\mathsf{V}_{\mathtt{F},-,\nu,p}(h_n) = (1/\tau_{T,\nu}^2)\mathcal{V}_{YY-,\nu,p}(h_n) - (2\tau_{Y,\nu}/\tau_{T,\nu}^3)\mathcal{V}_{YT-,\nu,p}(h_n) + (\tau_{Y,\nu}^2/\tau_{T,\nu}^4)\mathcal{V}_{TT-,\nu,p}(h_n)$,

where, for $U = Y, T$ and $V = Y, T$

$$\begin{aligned}
\mathcal{V}_{UV+,\nu,p}(h_n) &= n^{-1}h_n^{-2\nu}\nu!^2 e_\nu'\Gamma_{+,p}^{-1}(h_n)\Psi_{UV+,p}(h_n)\Gamma_{+,p}^{-1}(h_n)e_\nu \\
&= n^{-1}h_n^{-1-2\nu}\sigma_{UV+}^2\nu!^2 e_\nu'\Gamma_p^{-1}\Psi_p\Gamma_p^{-1}e_\nu/f\ \{1+o_p(1)\} \\
\mathcal{V}_{UV-,\nu,p}(h_n) &= n^{-1}h_n^{-2\nu}\nu!^2 e_\nu'\Gamma_{-,p}^{-1}(h_n)\Psi_{UV-,p}(h_n)\Gamma_{-,p}^{-1}(h_n)e_\nu \\
&= n^{-1}h_n^{-1-2\nu}\sigma_{UV-}^2\nu!^2 e_\nu'\Gamma_p^{-1}\Psi_p\Gamma_p^{-1}e_\nu/f\ \{1+o_p(1)\}
\end{aligned}$$

(D) *If $nh_n^{2p+5} \to 0$ and $nh_n^{1+2\nu} \to \infty$, then*

$$(\hat{\varsigma}_{\nu,p}(h_n) - \varsigma_\nu - h_n^{p+1-\nu}\mathsf{B}_{\mathtt{F},\nu,p,p+1}(h_n))/\sqrt{\mathsf{V}_{\mathtt{F},\nu,p}(h_n)} \to_d \mathcal{N}(0,1)$$

The following theorem gives an analogue of Theorem B.2 for fuzzy RD designs; Theorems II.15 and II.16 are special cases with $(\nu,p,q) = (0,1,2)$ and $(\nu,p,q) = (1,2,3)$, respectively. This theorem summarizes the asymptotic bias, variance and

distribution of the bias-corrected fuzzy RD estimator: $\hat{\varsigma}^{\mathsf{bc}}_{\nu,p,q}(h_n,b_n) = \hat{\varsigma}_{\nu,p}(h_n) - h_n^{p+1-\nu}\hat{\mathsf{B}}_{\mathtt{F},\nu,p,q}(h_n,b_n)$,

$\hat{\mathsf{B}}_{\mathtt{F},\nu,p,q}(h_n,b_n) = [(e'_{p+1}\hat{\beta}_{Y+,q}(b_n))\mathcal{B}_{+,\nu,p,p+1}(h_n) - (e'_{p+1}\hat{\beta}_{Y-,q}(b_n))\mathcal{B}_{-,\nu,p,p+1}(h_n)]/\hat{\tau}_{T,\nu,p}(h_n)$
$-\hat{\tau}_{Y,\nu,p}(h_n)[(e'_{p+1}\hat{\beta}_{T+,q}(b_n))\mathcal{B}_{+,\nu,p,p+1}(h_n) - (e'_{p+1}\hat{\beta}_{T-,q}(b_n))\mathcal{B}_{-,\nu,p,p+1}(h_n)]/\hat{\tau}_{T,\nu,p}(h_n)^2$.

Linearizing the estimator we obtain: $\hat{\varsigma}^{\mathsf{bc}}_{\nu,p,q}(h_n,b_n) - \varsigma_\nu = \tilde{\varsigma}^{\mathsf{bc}}_{\nu,p,q}(h_n,b_n) + R_n - R^{\mathsf{bc}}_n$,

$\tilde{\varsigma}^{\mathsf{bc}}_{\nu,p,q}(h_n,b_n) = (\hat{\tau}^{\mathsf{bc}}_{Y,\nu,p,q}(h_n,b_n) - \tau_{Y,\nu})/\tau_{T,\nu} - \tau_{Y,\nu}(\hat{\tau}^{\mathsf{bc}}_{T,\nu,p,q}(h_n,b_n) - \tau_{T,\nu})/\tau^2_{T,\nu}$,

$R_n = \tau_{Y,\nu}(\hat{\tau}_{T,\nu,p}(h_n) - \tau_{T,\nu})^2/(\tau^2_{T,\nu}\hat{\tau}_{T,\nu,p}(h_n)) - (\hat{\tau}_{Y,\nu,p}(h_n) - \tau_{Y,\nu})(\hat{\tau}_{T,\nu,p}(h_n) - \tau_{T,\nu})/(\tau_{T,\nu}\hat{\tau}_{T,\nu,p}(h_n))$,

$R^{\mathsf{bc}}_n = h_n^{p+1-\nu}(\hat{\mathsf{B}}_{\mathtt{F},\nu,p,q}(h_n,b_n) - \check{\mathsf{B}}_{\mathtt{F},\nu,p,q}(h_n,b_n))$,

$\check{\mathsf{B}}_{\mathtt{F},\nu,p,q}(h_n,b_n) = [(e'_{p+1}\hat{\beta}_{Y+,q}(b_n))\mathcal{B}_{+,\nu,p,p+1}(h_n) - (e'_{p+1}\hat{\beta}_{Y-,q}(b_n))\mathcal{B}_{-,\nu,p,p+1}(h_n)]/\tau_{T,\nu}$
$-\tau_{Y,\nu}[(e'_{p+1}\hat{\beta}_{T+,q}(b_n))\mathcal{B}_{+,\nu,p,p+1}(h_n) - (e'_{p+1}\hat{\beta}_{T-,q}(b_n))\mathcal{B}_{-,\nu,p,p+1}(h_n)]/\tau^2_{T,\nu}$.

Theorem B.4. *Suppose Assumptions III.1–II.14 hold with $S \geq p+2$, and $n\min\{h_n,b_n\} \to \infty$. Let $\nu \leq p < q$.*

($\boldsymbol{R}^{\mathsf{bc}}$) *If $h_n \to 0$ and $nh_n^{1+2\nu} \to \infty$, and provided that $\kappa b_n < \kappa_0$, then $R^{\mathsf{bc}}_n = O_p(n^{-1/2}h_n^{p+1/2} + h_n^{2p+2-2\nu})O_p(1 + n^{-1/2}b_n^{-3/2-p})$.*

(B) *If $\max\{h_n,b_n\} \to 0$, then*

$$\mathbb{E}[\tilde{\varsigma}^{\mathsf{bc}}_{\nu,p,q}(h_n,b_n)|\mathcal{X}_n] = h_n^{p+2-\nu}\mathsf{B}_{\mathtt{F},\nu,p,p+2}(h_n)\{1+o_p(1)\} + h_n^{p+1-\nu}b_n^{q-p}\mathsf{B}^{\mathsf{bc}}_{\mathtt{F},\nu,p,q}(h_n,b_n)\{1+o_p(1)\}$$

where $\mathsf{B}^{\mathsf{bc}}_{\mathtt{F},\nu,p,q}(h,b) = \mathsf{B}^{\mathsf{bc}}_{Y,\nu,p,q}(h_n,b_n)/\tau_{T,\nu} - \tau_{Y,\nu}\mathsf{B}^{\mathsf{bc}}_{T,\nu,p,q}(h_n,b_n)/\tau^2_{T,\nu}$

$$\begin{aligned}
\mathsf{B}^{\mathsf{bc}}_{Y,\nu,p,q}(h,b) &= [\mu^{(q+1)}_{Y+}\mathcal{B}_{+,p+1,q,q+1}(b)\mathcal{B}_{+,\nu,p,p+1}(h) - \mu^{(q+1)}_{Y-}\mathcal{B}_{-,p+1,q,q+1}(b)\mathcal{B}_{-,\nu,p,p+1}(h)]/[(q+1)!(p+\\
\mathsf{B}^{\mathsf{bc}}_{T,\nu,p,q}(h,b) &= [\mu^{(q+1)}_{T+}\mathcal{B}_{+,p+1,q,q+1}(b)\mathcal{B}_{+,\nu,p,p+1}(h) - \mu^{(q+1)}_{T-}\mathcal{B}_{-,p+1,q,q+1}(b)\mathcal{B}_{-,\nu,p,p+1}(h)]/[(q+1)!(p+
\end{aligned}$$

(V) $\mathsf{V}^{\mathtt{bc}}_{\mathtt{F},\nu,p,q}(h_n,b_n) = \mathbb{V}[\hat{\varsigma}^{\mathtt{bc}}_{\nu,p,q}(h_n,b_n)|\mathcal{X}_n] = \mathsf{V}^{\mathtt{bc}}_{\mathtt{F},+,\nu,p,q}(h_n,b_n) + \mathsf{V}^{\mathtt{bc}}_{\mathtt{F},-,\nu,p,q}(h_n,b_n)$ *with*

$$\begin{aligned}
\mathsf{V}^{\mathtt{bc}}_{\mathtt{F},+,\nu,p,q}(h,b) &= \mathsf{V}_{\mathtt{F},+,\nu,p}(h) - 2h^{p+1-\nu}\mathcal{C}_{\mathtt{F},+,\nu,p,q}(h,b)\mathcal{B}_{+,\nu,p,p+1}(h)/(p+1)! \\
&+ h^{2p+2-2\nu}\mathsf{V}_{\mathtt{F},+,p+1,q}(b)\mathcal{B}^2_{+,\nu,p,p+1}(h)/(p+1)!^2 \\
\mathsf{V}^{\mathtt{bc}}_{\mathtt{F},-,\nu,p,q}(h,b) &= \mathsf{V}_{\mathtt{F},-,\nu,p}(h) - 2h^{p+1-\nu}\mathcal{C}_{\mathtt{F},-,\nu,p,q}(h,b)\mathcal{B}_{-,\nu,p,p+1}(h)/(p+1)! \\
&+ h^{2p+2-2\nu}\mathsf{V}_{\mathtt{F},-,p+1,q}(b)\mathcal{B}^2_{-,\nu,p,p+1}(h)/(p+1)!^2
\end{aligned}$$

$\mathcal{C}_{\mathtt{F},+,\nu,p,q}(h,b) = (1/\tau^2_{T,\nu})\mathcal{C}_{YY+,\nu,p,q}(h,b) - (2\tau_{Y,\nu}/\tau^3_{T,\nu})\mathcal{C}_{YT+,\nu,p,q}(h,b) + (\tau^2_{Y,\nu}/\tau^4_{T,\nu})\mathcal{C}_{TT+,\nu,p,q}(h,b),$

$\mathcal{C}_{\mathtt{F},-,\nu,p,q}(h,b) = (1/\tau^2_{T,\nu})\mathcal{C}_{YY-,\nu,p,q}(h,b) - (2\tau_{Y,\nu}/\tau^3_{T,\nu})\mathcal{C}_{YT-,\nu,p,q}(h,b) + (\tau^2_{Y,\nu}/\tau^4_{T,\nu})\mathcal{C}_{TT-,\nu,p,q}(h,b),$

where, for $U = Y, T$ *and* $V = Y, T$,

$\mathcal{C}_{UV+,\nu,p,q}(h,b) = n^{-1}h^{-\nu}b^{-p-1}\nu!(p+1)!e'_\nu\Gamma^{-1}_{+,p}(h)\Psi_{UV+,p,q}(h,b)\Gamma^{-1}_{+,q}(b)e_{p+1},$

$\mathcal{C}_{UV-,\nu,p,q}(h,b) = n^{-1}h^{-\nu}b^{-p-1}\nu!(p+1)!e'_\nu\Gamma^{-1}_{-,p}(h)\Psi_{UV-,p,q}(h,b)\Gamma^{-1}_{-,q}(b)e_{p+1}.$

(D) *If* $n\min\{h_n^{2p+3}, b_n^{2p+3}\}\max\{h_n^2, b_n^{2(q-p)}\} \to 0$ *and* $n\min\{h_n^{1+2\nu}, b_n\} \to \infty$, *and* $h_n \to 0$ *and* $\kappa b_n < \kappa_0$, *then*

$T^{\mathtt{rbc}}_{\mathtt{F},\nu,p,q}(h_n,b_n) = (\hat{\varsigma}^{\mathtt{bc}}_{\nu,p,q}(h_n,b_n) - \varsigma_\nu)/\sqrt{\mathsf{V}^{\mathtt{bc}}_{\mathtt{F},\nu,p,q}(h_n,b_n)} \to_d \mathcal{N}(0,1).$

Thus, $\mathsf{V}^{\mathtt{bc}}_{\mathtt{FRD}}(h_n,b_n) = \mathsf{V}^{\mathtt{bc}}_{\mathtt{F},0,1,2}(h_n,b_n)$ and $\mathsf{V}^{\mathtt{bc}}_{\mathtt{FKRD}}(h_n,b_n) = \mathsf{V}^{\mathtt{bc}}_{\mathtt{F},1,2,3}(h_n,b_n)$ in Theorems II.15 and II.16, respectively.

B.1.4 Sharp RD Bandwidth Selectors

For any $\nu \leq p$, let $\hat{\mathsf{V}}_{\nu,p}(h_n) = \hat{\mathcal{V}}_{+,\nu,p}(h_n) + \hat{\mathcal{V}}_{-,\nu,p}(h_n)$, with

$$\begin{aligned}
\hat{\mathcal{V}}_{+,\nu,p}(h_n) &= \nu!^2 e'_\nu\hat{\Gamma}^{-1}_{+,p}(h_n)\hat{\Psi}_{YY+,p}(h_n)\hat{\Gamma}^{-1}_{+,p}(h_n)e_\nu/nh_n^{2\nu} \\
\hat{\mathcal{V}}_{-,\nu,p}(h_n) &= \nu!^2 e'_\nu\hat{\Gamma}^{-1}_{-,p}(h_n)\hat{\Psi}_{YY-,p}(h_n)\hat{\Gamma}^{-1}_{-,p}(h_n)e_\nu/nh_n^{2\nu}
\end{aligned}$$

where $\hat{\Psi}_{YY+,p}(h_n)$ and $\hat{\Psi}_{YY-,p}(h_n)$ as in Section 2.5.

Plug-in Bandwidths Selectors. Fix $p, q \in \mathbb{N}$ with $q \geq p+1$. Let $\mathcal{B}_{\nu,p} = e'_\nu\Gamma_p^{-1}\vartheta_{p,p+1}$.

Step 0: Initial Bandwidths (v_n, c_n).

(i) Suppose $v_n \to_p 0$ and $nv_n \to_p \infty$. In particular, let $v_n = 2.58 \cdot \omega \cdot n^{-1/5}$ with $\omega = \min\{S_X,\ IQR_X/1.349\}$, where S_X^2 denotes the sample variance of X_i, and IQR_X is the interquartile range of X_i.

(ii) Suppose $c_n \to_p 0$ and $nc_n^{2q+3} \to_p \infty$. In particular, let $c_n = \hat{C}_{q+1,q+1}^{1/(2q+5)}\ n^{-1/(2q+5)}$ with $\hat{C}_{q+1,q+1} = (2q+3)nv_n^{2q+3}\hat{\mathsf{V}}_{q+1,q+1}(v_n)/\{2\mathcal{B}_{q+1,q+1}^2(e'_{q+2}\check{\beta}_{+,q+2}+e'_{q+2}\check{\beta}_{-,q+2})^2\}$, where $\check{\beta}_{+,p}$ and $\check{\beta}_{-,p}$ denote the estimated coefficients of a $(p+1)$-th order global polynomial fit at either side of the threshold; i.e., $\check{\beta}_{+,p} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \mathbf{1}(X_i \geq 0)(Y_i - r_p(X_i)'\beta)^2$ and $\check{\beta}_{-,p} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \mathbf{1}(X_i < 0)(Y_i - r_p(X_i)'\beta)^2$.

Step 1: Pilot Bandwidth b_n. Compute $\hat{b}_{p+1,q} = \hat{C}_{p+1,q}^{1/(2q+3)}\ n^{-1/(2q+3)}$ with $\hat{C}_{p+1,q} = (2p+3)nv_n^{2p+3}\hat{\mathsf{V}}_{p+1,q}(v_n)/\{2(q-p)\mathcal{B}_{p+1,q}^2[(e'_{q+1}\hat{\beta}_{+,q+1}(c_n)-(-1)^{p+q+2}e'_{q+1}\hat{\beta}_{-,q+1}(c_n))^2+3\hat{\mathsf{V}}_{q+1,q+1}(c_n)]\}$.

Step 2: Main Bandwidth h_n. Let $b_n = \hat{b}_{p+1,q}$, and compute $\hat{h}_{\nu,p} = \hat{C}_{\nu,p}^{1/(2p+3)}\ n^{-1/(2p+3)}$ with $\hat{C}_{\nu,p} = (2\nu+1)nv_n\hat{\mathsf{V}}_{p,0}(v_n)/\{2(p+1-\nu)\mathcal{B}_{\nu,p}^2[(e'_{p+1}\hat{\beta}_{+,q}(b_n)-(-1)^{\nu+p+1}e'_{p+1}\hat{\beta}_{-,q}(b_n))^2+3\hat{\mathsf{V}}_{p+1,q}(b_n)]\}$.

Theorem B.5. *Suppose Assumptions III.1–II.3 hold with $S \geq q+1$ and $p < q$. In addition, suppose $e'_{q+2}\check{\beta}_{+,q+2} + e'_{q+2}\check{\beta}_{-,q+2} \to_p c \neq 0$ and $\nu \leq p$. Let $\mathsf{MSE}_{\nu,p}(h_n) = \mathbb{E}[(\hat{\tau}_{\nu,p}(h_n) - \tau_\nu)^2|\mathcal{X}_n]$ to save notation.*

*(**Step 1**) If $\mathsf{B}_{p+1,q,q+1} \neq 0$, then $\hat{b}_{p+1,q}/b_{\mathtt{MSE},p+1,q} \to_p 1$ and*
$\mathsf{MSE}_{p+1,q}(\hat{b}_{p+1,q})/\mathsf{MSE}_{p+1,q}(b_{\mathtt{MSE},p+1,q}) \to_p 1$.

*(**Step 2**) If $\mathsf{B}_{\nu,p,p+1} \neq 0$, then $\hat{h}_{\nu,p}/h_{\mathtt{MSE},\nu,p} \to_p 1$ and*
$\mathsf{MSE}_{\nu,p}(\hat{h}_{\nu,p})/\mathsf{MSE}_{\nu,p}(h_{\mathtt{MSE},\nu,p}) \to_p 1$.

APPENDIX C

Appendix to Chapter 3

C.1 Appendix Chapter 3

We state and prove results only for the treatment group (subindex "+") because all the proofs for the control group are analogous. We offer short proofs of the main results, and provide references to the underlying results not reproduced here to conserve space. Recall that the lower and upper end points of $P_{+,j}$ are denoted, respectively, by $p_{+,j-1}$ and $p_{+,j}$ for $j = 1, 2, \cdots, J_{+,n}$, which are nonrandom under ES partitioning (Section 3.3) and random under QS partitioning (Section 3.4). Let $\bar{p}_{+,j} = (p_{+,j} + p_{+,j-1})/2$ be the middle point of bin $P_{+,j}$. Throughout the appendix C denotes a positive, bounded constant that may take different values in different places.

We provide three lemmas that will be used to prove our main results. The first lemma holds for any nonrandom partition $\mathcal{P}_{+,n}$ satisfying $C_1 J_{+,n} \leq \min_{1\leq j\leq J_{+,n}} |p_{+,j} - p_{+,j-1}| \leq \max_{1\leq j\leq J_{+,n}} |p_{+,j} - p_{+,j-1}| \leq C_2 J_{+,n}$, for fixed positive constants C_1 and C_2. Thus, it holds for $\mathcal{P}_{\mathtt{ES},+,n}$ in particular.

Lemma C.1. *Let Assumption III.1 hold. Consider $\mathcal{P}_{\mathtt{ES},+,n}$ with $J_{+,n}\log(J_{+,n})/n \to 0$ and $J_{+,n} \to \infty$.*

(i) $\max_{1\leq j\leq J_{+,n}} |\mathbb{1}(N_{+,j}>0)-1| = o_{\mathbb{P}}(1)$.

(ii) $\max_{1\leq j\leq J_{+,n}} |n^{-1}N_{+,j} - \mathbb{P}[X_i\in P_{+,n,j}]| = o_{\mathbb{P}}(J_{+,n}^{-1})$.

(iii) $\max_{1\leq j\leq J_{+,n}} \left| n^{-1}\sum_{i=1}^n \mathbb{1}_{P_{+,n,j}}(X_i)\frac{X_i-\bar{p}_{+,j}}{p_{+,j}-p_{+,j-1}} - \mathbb{E}\left[\mathbb{1}_{P_{+,n,j}}(X_i)\frac{X_i-\bar{p}_{+,j}}{p_{+,j}-p_{+,j-1}}\right]\right| = o_{\mathbb{P}}(J_{+,n}^{-1})$.

(iv) $\max_{1\leq j\leq J_{+,n}} \left| \mathbb{E}\left[\mathbf{1}_{P_{+,n,j}}(X_i)\frac{X_i-\bar{p}_{+,n,j}}{p_{+,n,j}-p_{+,n,j-1}}\right]\right| = o(J_{+,n}^{-1})$.

Proof of Lemma C.1. Parts (i)-(iii) follow by Hoeffding exponential inequality, while part (iv) follows by change of variables and standard bounding arguments. See Cattaneo and Farrell (2013b) for details. ■

Lemma C.1(i) shows that $\mathbb{1}(N_{+,j}>0)\to_{\mathbb{P}} 1$ uniformly in j, which guarantees that the estimators for the ES partitioning scheme are well-behaved in large samples.

Our second lemma characterizes the properties of the random partitioning scheme based on quantile estimates. These results will be used when handling the partitioning scheme $\mathcal{P}_{\mathtt{QS},+,n}$: recall that $p_{+,j} = \hat{F}_+^{-1}(j/J_{+,n})$ in this case, $j=1,2,\cdots,J_{+,n}$, and thus set $q_{+,j} = F_+^{-1}(j/J_{+,n})$ with $F_+^{-1}(y) = \inf\{x : F_+(x)\geq y\}$ with $F_+(x) = \mathbb{P}[X_i\leq x, X_i\geq \bar{x}]/\mathbb{P}[X_i\geq\bar{x}] = F(x|X_i\geq\bar{x})$.

Lemma C.2. *Let Assumption III.1 hold. Consider $\mathcal{P}_{\mathtt{QS},+,n}$ with $J_{+,n}\log(J_{+,n})/n\to 0$ and $J_{+,n}/\log(n)\to\infty$.*

(i) $\max_{1\leq j\leq J_{+,n}} |N_{+,j}/N_+ - 1/J_{+,n}| = o_{\mathbb{P}}(J_{+,n}^{-1})$.

(ii) $\max_{1\leq j\leq J_{+,n}} |p_{+,j}-p_{+,j-1}-(q_{+,n,j}-q_{+,j-1})| = o_{\mathbb{P}}(J_{+,n}^{-1})$.

Proof of Lemma C.2. Because the sample size N_+ is random, we employ the following result: if $N_+\to_{\text{as}}\infty$ and $Z_n\to_{\text{as}} Z_\infty$, then $Z_{N_+}\to_{\text{as}} Z_\infty$. In our case, $N_+ = \sum_{i=1}^n \mathbb{1}(X_i\geq\bar{x})$ and thus $N_+/n\to_{\text{as}} P_+$. Hence, it suffices to assume $N_+\to\infty$ is not random, but we need to prove the statements in an almost sure sense. The rest of the proof takes limits as $N_+\to\infty$.

Part (i) now follows from properties of distribution function and quantile processes (e.g., Shorack and Wellner, 2009). Using continuity and boundedness of $f(x)$, we have

$$\begin{aligned} N_{+,j} &= \sum_{i=1}^{n} \mathbb{1}\left(\hat{F}_+^{-1}\left(\frac{j-1}{J_{+,n}}\right) \leq X_i < \hat{F}_+^{-1}\left(\frac{j}{J_{+,n}}\right)\right) \\ &= N_+\hat{F}_+\left(\hat{F}_+^{-1}\left(\frac{j}{J_{+,n}}\right)\right) - N_+\hat{F}_+\left(\hat{F}_+^{-1}\left(\frac{j-1}{J_{+,n}}\right)\right)\{1+o_{\text{as}}(1)\} = \frac{N_+}{J_{+,n}}\{1+o_{\text{as}}(1)\}, \end{aligned}$$

uniformly in $j = 1, 2, \cdots, J_{+,n}$, under the rate restrictions imposed.

Similarly, part (ii) follows from properties of the modulus of continuity of the sample quantile process (e.g., Shorack and Wellner, 2009, Chapter 14). We have

$$\begin{aligned} &\max_{1\leq j\leq J_{+,n}} |p_{+,j} - p_{+,j-1} - (q_{+,n,j} - q_{+,j-1})| \\ &= \max_{1\leq j\leq J_{+,n}} \left|\hat{F}_+^{-1}\left(\frac{j}{J_{+,n}}\right) - F_+^{-1}\left(\frac{j}{J_{+,n}}\right) - \left(\hat{F}_+^{-1}\left(\frac{j-1}{J_{+,n}}\right) - F_+^{-1}\left(\frac{j-1}{J_{+,n}}\right)\right)\right| = o_{\text{as}}(J_{+,n}^{-1}), \end{aligned}$$

under the rate restrictions imposed. ■

Our final lemma gives the main convergence results for the spacings estimators used to construct data-driven choices of partition sizes.

Lemma C.3. *Let Assumption III.1 hold. Suppose $Y_i(1)$ is continuously distributed and $g : [\bar{x}, x_u] \to \mathbb{R}_+$ is continuous. Set $k \in \mathbb{Z}_+$.*

(i) $N_+^{k-1}\sum_{i=2}^{N_+}(X_{+,(i)} - X_{+,(i-1)})^k g(\bar{X}_{+,(i)}) \to_{\mathbb{P}} k!P_+^{k-1}\int_{\bar{x}}^{x_u} f(x)^{1-k}g(x)dx.$

(ii) $N_+^{k-1}\sum_{i=2}^{N_+}(X_{+,(i)} - X_{+,(i-1)})^k (Y_{+,[i]} - Y_{+,[i-1]})^2 g(\bar{X}_{+,(i)}) \to_{\mathbb{P}} k!P_+^{k-1}2\int_{\bar{x}}^{x_u} f(x)^{1-k}\sigma_+^2(x)g(x)dx.$

Proof of Lemma C.3. We prove the result assuming that N_+ is nonrandom, and thus limits are taken as $N_+ \to \infty$. Set $U_i = F_+(X_{+,i}) \sim \text{Uniform}(0,1)$ and $U_{(i)} = F_+(X_{+,(i)})$, $i = 1, \cdots, N_+$. Recall that $\{N_+(U_{(i)} - U_{(i-1)}) : i = 2, \cdots, N_+\} =_{\text{d}} \{E_i/\bar{E} : i = 2, \cdots, N_+\}$, where $\{E_i : i = 2, \cdots, N_+\}$ i.i.d. random variables with $E_i \sim \text{Exponential}(1)$ and $\bar{E} = \sum_{i=2}^{N_+} E_i/N_+$, and where $Z_1 =_{\text{d}} Z_2$ denotes that

Z_1 and Z_2 have the same probability law. Set $\bar{u}_i = (i-1/2)/N_+$ and recall that $\max_{2\le i\le N_+}\sup_{U_{(i-1)}\le u\le U_{(i)}}|u-\bar{u}_i| \to_{\mathbb{P}} 0$.

For part (i), using the above, $N_+^{-1}\sum_{i=2}^{N_+} E_i^k \to_{\mathbb{P}} \mathbb{E}[E_i^k] = k!$, and uniform continuity of $g(\cdot)$ and $f(\cdot)$,

$$\begin{aligned}
&N_+^{k-1}\sum_{i=2}^{N_+}(X_{+,(i)}-X_{+,(i-1)})^k g(\bar{X}_{+,(i)}) \\
&\quad= \frac{1}{N_+}\sum_{i=2}^{N_+}(N_+(U_{(i)}-U_{(i-1)}))^k \frac{g(F_+^{-1}(u_{n,i}))}{f_+(F_+^{-1}(u_{n,i}))^k}\{1+o_{\mathbb{P}}(1)\} \\
&\quad=_{\mathrm{d}} \frac{1}{N_+}\sum_{i=2}^{N_+}\left(\frac{E_i}{\bar{E}}\right)^k \frac{g(F_+^{-1}(u_{n,i}))}{f_+(F_+^{-1}(u_{n,i}))^k}\{1+o_{\mathbb{P}}(1)\} \\
&\quad= \frac{1}{N_+}\sum_{i=2}^{N_+}\mathbb{E}[E_i^k]\frac{g(F_+^{-1}(u_{n,i}))}{f_+(F_+^{-1}(u_{n,i}))^k}\{1+o_{\mathbb{P}}(1)\} \\
&\quad\to_{\mathbb{P}} k!\int_0^1 \frac{g(F_+^{-1}(u))}{f_+(F_+^{-1}(u))^p}du,
\end{aligned}$$

and the result follows by change of variables and because $f_+(x) = f(x)\mathbb{1}(x\ge\bar{x})/P_+$. This result implies, in particular, $\sum_{i=2}^{N_+}(X_{+,(i)}-X_{+,(i-1)})^k g(\bar{X}_{+,(i)}) = O_{\mathbb{P}}(N_+^{1-k})$.

For part (ii), let $\mathbf{X}_{(+)} = (X_{+,(1)}, X_{+,(2)}, \cdots, X_{+,(N_+)})$. Recall that $(Y_{+,[1]}, Y_{+,[2]}, \cdots, Y_{+,[N_+]})$ are independent conditional on $\mathbf{X}_{(+)}$ and $\mathbb{E}[g(Y_{+,[i]})|\mathbf{X}_{(+)}] = \mathbb{E}[g(Y_{+,[i]})|X_{+,(i)}] = G(X_{+,(i)})$ with $G(x) = \mathbb{E}[g(Y_{+,i})|X_{+,i}=x]$. Therefore, $\mathbb{E}[(Y_{+,[i]}-Y_{+,[i-1]})^2|\mathbf{X}_{(+)}] = \sigma_+^2(X_{+,(i)})+\sigma_+^2(X_{+,(i-1)})+(\mathbb{E}[Y_{+,[i]}|\mathbf{X}_{(+)}]-\mathbb{E}[Y_{+,[i-1]}|\mathbf{X}_{(+)}])^2 = \sigma_+^2(X_{+,(i)})+\sigma_+^2(X_{+,(i-1)})+O_{\mathbb{P}}(N_+^{-2})$, uniformly in i. This gives

$$N_+^{k-1}\sum_{i=2}^{N_+}(X_{+,(i)}-X_{+,(i-1)})^k(Y_{+,[i]}-Y_{+,[i-1]})^2 g(\bar{X}_{+,(i)}) = T_1+T_2,$$

with

$$T_1 = N_+^{k-1}\sum_{i=2}^{N_+}(X_{+,(i)}-X_{+,(i-1)})^k(\sigma_+^2(X_{+,[i]})+\sigma_+^2(X_{+,[i-1]}))g(\bar{X}_{+,(i)})+o_{\mathbb{P}}(1),$$

$$T_2 = N_+^{k-1} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^k \left[(Y_{+,[i]} - Y_{+,[i-1]})^2 - \mathbb{E}[(Y_{+,[i]} - Y_{+,[i-1]})^2 | \mathbf{X}_{[+]}] \right] g(\bar{X}_{+,(i)}).$$

Noting that $\sigma_+^2(X_{+,(i)}) + \sigma_+^2(X_{+,(i-1)}) = 2\sigma_+^2(\bar{X}_{+,(i)})\{1 + o_{\mathbb{P}}(1)\}$, uniformly in i, it follows that $T_1 \to_{\mathbb{P}} k! P_+^{k-1} 2 \int_{\bar{x}}^{x_u} f(x)^{1-k} \sigma_+^2(x) g(x) dx$, as in part (i). Thus, it remains to show that $T_2 \to_{\mathbb{P}} 0$. To this end, first define $\tilde{Y}_i = (Y_{+,[i]} - Y_{+,[i-1]})^2 - \mathbb{E}[(Y_{+,[i]} - Y_{+,[i-1]})^2 | \mathbf{X}_{(+)}]$, and note that $\mathbb{E}[\tilde{Y}_i, \tilde{Y}_{i-s} | \mathbf{X}_{(+)}] = 0$ whenever $s \geq 2$, which implies

$$\begin{aligned}
\mathbb{V}[T_2 | \mathbf{X}_{(+)}] &\leq N_+^{2(k-1)} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^{2k} \mathbb{V}[\tilde{Y}_i | \mathbf{X}_{(+)}] g(\bar{X}_{+,(i)})^2 \\
&\quad + 2N_+^{2(k-1)} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^k (X_{+,(i-1)} - X_{+,(i-2)})^k \mathbb{E}[\tilde{Y}_i \tilde{Y}_{i-1} | \mathbf{X}_{(+)}] g(\bar{X}_{+,(i)}) g(\bar{X}_{+,(i-} \\
&\leq C N_+^{-1},
\end{aligned}$$

and the result follows by the dominated convergence theorem.

The random sample size case ($N_+ = \sum_{i=1}^n \mathbb{1}(X_i \geq \bar{x})$) can be handled, for example, using the approach described in Aras, Jammalamadaka, and Zhou (1989) and references therein. ■

Proof of Theorem III.2. Note that

$$\mathbb{E}[(\hat{\mu}_{\mathtt{ES},+}(x; J_{+,n}) - \mu_+(x))^2 | \mathbf{X}_n] = \mathbb{V}[\hat{\mu}_{\mathtt{ES},+}(x; J_{+,n}) | \mathbf{X}_n] + (\mathbb{E}[\hat{\mu}_{\mathtt{ES},+}(x; J_{+,n}) | \mathbf{X}_n] - \mu_+(x))^2.$$

For the variance part, we have

$$\mathbb{V}[\hat{\mu}_+(x; J_{+,n}) | \mathbf{X}_n] = \sum_{j=1}^{J_{+,n}} \frac{\mathbb{1}(N_{+,j} > 0) \mathbb{1}_{P_{+,j}}(x)}{N_{+,j}^2} \sum_{i=1}^n \mathbb{1}_{P_{+,j}}(X_i) \sigma_+^2(X_i),$$

and using uniform continuity of $w(\cdot)$ and $\sigma_+^2(\cdot)$ on $[\bar{x}, x_u]$ and Lemma C.1, we obtain

$$\begin{aligned}
&\int_{\bar{x}}^{x_u} \mathbb{V}[\hat{\mu}_+(x; J_{+,n})|\mathbf{X}_n] w(x) dx \\
&\quad = \sum_{j=1}^{J_{+,n}} \frac{\mathbb{1}(N_{+,j} > 0)}{N_{+,j}^2} \left(\int_{\bar{x}}^{x_u} \mathbb{1}_{P_{+,n,j}}(x) w(x) dx \right) \sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i) \sigma_+^2(X_i) \\
&\quad = \sum_{j=1}^{J_{+,n}} \frac{\mathbb{1}(N_{+,j} > 0)}{N_{+,j}} (p_{+,j} - p_{+,j-1}) \sigma_+^2(\bar{p}_{+,j}) w(\bar{p}_{+,j}) \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{1}{n} \sum_{j=1}^{J_{+,n}} \frac{\sigma_+^2(\bar{p}_{+,j}) w(\bar{p}_{+,j})}{f(\bar{p}_{+,j})} \{1 + o_{\mathbb{P}}(1)\},
\end{aligned}$$

because $\mathbb{P}[X_i \in P_{+,j}] = \int_{p_{+,j-1}}^{p_{+,j}} f(x) dx = (p_{+,j} - p_{+,j-1}) f(\bar{p}_{+,j}) \{1 + o(1)\}$ uniformly in j. Using properties of the Riemann integral it then follows that

$$\begin{aligned}
&\int_{\bar{x}}^{x_u} \mathbb{V}[\hat{\mu}_{\texttt{ES},+}(x; J_{+,n})|\mathbf{X}_n] w(x) dx \\
&\quad = \frac{J_{+,n}}{n} \frac{1}{x_u - \bar{x}} \sum_{j=1}^{J_{+,n}} (p_{+,j} - p_{+,j-1}) \frac{\sigma_+^2(\bar{p}_{+,j}) w(\bar{p}_{+,j})}{f(\bar{p}_{+,j})} \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{J_{+,n}}{n} \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \frac{\sigma_+^2(x)}{f(x)} w(x) dx \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{J_{+,n}}{n} \not{V}_{\texttt{ES},+} \{1 + o_{\mathbb{P}}(1)\},
\end{aligned}$$

because $p_{+,j+1} - p_{+,j} = (x_u - \bar{x})/J_{+,n}$ for the evenly-spaced partition.

Next, for the bias term, note that $\int_{\bar{x}}^{x_u} (\mathbb{E}[\hat{\mu}_+(x; J_n)|\mathbf{X}_n] - \mu_+(x))^2 w(x) dx = T_1 + T_2 + T_3$ with

$$T_1 = \int_{\bar{x}}^{x_u} T_1(x)^2 w(x) dx, \qquad T_2 = \int_{\bar{x}}^{x_u} T_2(x)^2 w(x) dx, \qquad T_3 = 2 \int_{\bar{x}}^{x_u} T_1(x) T_2(x) w(x) dx,$$

$$T_1(x) = \sum_{j=1}^{J_{+,n}} \mathbb{1}_{P_{+,j}}(x)(\mathbb{1}(N_{+,j} > 0)\mu_+(\bar{p}_{+,j}) - \mu_+(x)),$$

$$T_2(x) = \sum_{j=1}^{J_{+,n}} \mathbb{1}_{P_{+,j}}(x)\frac{\mathbb{1}(N_{+,j} > 0)}{N_{+,j}} \left(\sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i)(\mu_+(X_i) - \mu_+(\bar{p}_{+,j})) \right).$$

Using uniform continuity of $\mu_+(\cdot)$ and $w(\cdot)$ on $[\bar{x}, x_u]$ and Lemma C.1, we obtain

$$\begin{aligned}
T_1 &= \frac{1}{12}\sum_{j=1}^{J_{+,n}} \left(\mu_+^{(1)}(\bar{p}_{+,j})\right)^2 w(\bar{p}_{+,j}) \int_{\bar{x}}^{x_u} \mathbb{1}_{P_{+,j}}(x)(\bar{p}_{+,j} - x)^2 dx\{1 + o_{\mathbb{P}}(1)\} \\
&= \frac{1}{12}\sum_{j=1}^{J_{+,n}} (p_{+,j} - p_{+,j-1})^3 \left(\mu_+^{(1)}(\bar{p}_{+,j})\right)^2 w(\bar{p}_{+,j})\{1 + o_{\mathbb{P}}(1)\} \\
&= \frac{1}{J_{+,n}^2}\frac{(x_u - \bar{x})^2}{12} \int_{\bar{x}}^{x_u} \left(\mu_+^{(1)}(x)\right)^2 w(x) dx\{1 + o_{\mathbb{P}}(1)\} = J_{+,n}^{-2}\mathscr{B}_{\mathtt{ES},+}\ \{1 + o_{\mathbb{P}}(1)\},
\end{aligned}$$

because $\int_a^b ((a+b)/2 - x)^2 dx = (b-a)^3/12$ and $p_{+,j+1} - p_{+,j} = (x_u - \bar{x})/J_{+,n}$ for the evenly-spaced partition. This implies that $T_1 = O_{\mathbb{P}}(J_{+,n}^{-2})$. Thus, to finish the proof, we show that $T_2 = o_{\mathbb{P}}(J_{+,n}^{-2})$ and $T_3 = o_{\mathbb{P}}(J_{+,n}^{-2})$. For T_2, using uniform continuity of $\mu_+(\cdot)$ and $w(\cdot)$ on $[\bar{x}, x_u]$ and Lemma C.1 we have

$$|T_2| \leq C\sum_{j=1}^{J_{+,n}} \frac{\mathbb{1}(N_{+,j} > 0)}{J_{+,n}^2 N_{+,j}^2/n^2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i)\frac{X_i - \bar{p}_{+,j}}{p_{+,j} - p_{+,j-1}} \right)^2 \{1 + o_{\mathbb{P}}(1)\} = o_{\mathbb{P}}(J_{+,n}^{-2}),$$

while, for T_3, Cauchy-Swartz inequality implies $|T_3| \leq \sqrt{T_1}\sqrt{T_2} = O_{\mathbb{P}}(J_{+,n}^{-1})o_{\mathbb{P}}(J_{+,n}^{-1}) = o_{\mathbb{P}}(J_{+,n}^{-2})$.

Putting the results together we verify the result for $\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})$. ■

Proof of Theorem III.3. Using Lemma C.3 and $N_+/n \to_{\mathbb{P}} P_+$,

$$\begin{aligned}
\hat{\mathscr{V}}_{\mathtt{ES},+} &= \frac{1}{x_u - \bar{x}} \frac{n}{4} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^2 (Y_{+,[i]} - Y_{+,[i-1]})^2 w(\bar{X}_{+,(i)}) \\
&= \frac{1}{x_u - \bar{x}} \frac{N_+}{4P_+} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^2 (Y_{+,[i]} - Y_{+,[i-1]})^2 w(\bar{X}_{+,(i)}) + o_{\mathbb{P}}(1) \\
&= \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \frac{\sigma_+^2(x)}{f_+(x)} w(x) dx + o_{\mathbb{P}}(1),
\end{aligned}$$

which gives $\hat{\mathscr{V}}_{\mathtt{ES},+} \to_{\mathbb{P}} \mathscr{V}_{\mathtt{ES},+}$.

For power series estimators, Newey (1997b, Theorem 4) gives $\sup_{x \in [\bar{x}, x_u]} |\hat{\mu}^{(1)}_{+,k_n}(x) - \mu^{(1)}_+(x)|^2 = O_{\mathbb{P}}(k_n^7/n + k_n^{-2S+8}) = o_{\mathbb{P}}(1)$. Using this uniform consistency result, and Lemma C.3, we have

$$\begin{aligned}
\hat{\mathcal{B}}_{\mathtt{ES},+} &= \frac{(x_u - \bar{x})^2}{12} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)}) \left(\hat{\mu}^{(1)}_{+,k_n}(\bar{X}_{+,(i)}) \right)^2 w(\bar{X}_{+,(i)}) \\
&= \frac{(x_u - \bar{x})^2}{12} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)}) \left(\mu^{(1)}_{+}(\bar{X}_{+,(i)}) \right)^2 w(\bar{X}_{+,(i)}) + o_{\mathbb{P}}(1) \\
&= \frac{(x_u - \bar{x})^2}{12} \int_{\bar{x}}^{x_u} \left(\mu^{(1)}_{+}(x) \right)^2 w(x) dx + o_{\mathbb{P}}(1),
\end{aligned}$$

which gives $\hat{\mathscr{B}}_{\mathtt{ES},+} \to_{\mathbb{P}} \mathscr{B}_{\mathtt{ES},+}$.

Putting the above together, $\hat{J}_{\mathtt{ES},+}/J_{\mathtt{ES},+} \to_{\mathbb{P}} 1$. ■

Proof of Remark III.4. For power series estimators, Newey (1997b, Theorem 4) gives $\sup_{x \in [\bar{x}, x_u]} |\hat{\mu}_{+,k_n,p}(x) - \mathbb{E}[Y(1)^p | X_i = x]|^2 = O_{\mathbb{P}}(k_n^3/n + k_n^{-2S+2}) = o_{\mathbb{P}}(1)$ for $p = 1, 2$, under the assumptions imposed, which implies $\sup_{x \in [\bar{x}, x_u]} |\hat{\sigma}^2_+(x) - \sigma^2_+|^2 =$

$O_{\mathbb{P}}(k_n^3/n + k_n^{-2S+2}) = o_{\mathbb{P}}(1)$. Using this result, Lemma C.3 and $N_+/n \to_{\mathbb{P}} P_+$,

$$\begin{aligned}
\check{\mathscr{V}}_{\mathtt{ES},+} &= \frac{1}{x_u - \bar{x}} \frac{n}{2} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^2 \hat{\sigma}_+^2(\bar{X}_{+,(i)}) w(\bar{X}_{+,(i)}) \\
&= \frac{1}{x_u - \bar{x}} \frac{N_+}{2P_+} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})^2 \sigma_+^2(\bar{X}_{+,(i)}) w(\bar{X}_{+,(i)}) + o_{\mathbb{P}}(1) \\
&= \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \frac{\sigma_+^2(x)}{f_+(x)} w(x) dx + o_{\mathbb{P}}(1),
\end{aligned}$$

which gives $\check{\mathscr{V}}_{\mathtt{ES},+} \to_{\mathbb{P}} \mathscr{V}_{\mathtt{ES},+}$

Combining this with Theorem III.3, it follows that $\check{J}_{\mathtt{ES},+}/J_{\mathtt{ES},+} \to_{\mathbb{P}} 1$. ■

Proof of Remark III.5. The results follow immediately from Lemma C.3, $N_+/n \to_{\mathbb{P}} P_+$, uniform consistency of power series estimators, and proceeding as in the proofs of Theorem III.3 and Remark III.4. ■

Proof of Theorem III.6. Recall that $p_{+,j} = \hat{F}_+^{-1}(j/J_{+,n})$ and $q_{+,j} = F_+^{-1}(j/J_{+,n})$. If $J_{+,n} < N_+$, then $\mathbb{1}(N_{+,j} > 0) = 1$, but now the partitioning scheme $\mathcal{P}_{\mathtt{QS},+,n}$ is random.

As in the proof of Theorem III.2, note that $\mathbb{E}[(\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) - \mu_+(x))^2 | \mathbf{X}_n] = \mathbb{V}[\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) | \mathbf{X}_n] + (\mathbb{E}[\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) | \mathbf{X}_n] - \mu_+(x))^2$. For the variance part, letting

$\bar{q}_{+,j} = (q_{+,j} + q_{+,j-1})/2$, we have

$$\begin{aligned}
&\int_{\bar{x}}^{x_u} \mathbb{V}[\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n})|\mathbf{X}_n] w(x) dx \\
&\quad = \sum_{j=1}^{J_{+,n}} \frac{1}{N_{+,j}^2} \left(\int_{\bar{x}}^{x_u} \mathbb{1}_{P_{+,j}}(x) w(x) dx \right) \sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i) \sigma_+^2(X_i) \\
&\quad = \frac{J_{+,n}}{N_+} \sum_{j=1}^{J_{+,n}} (p_{+,j} - p_{+,j-1}) \sigma_+^2(\bar{p}_{+,j}) w(\bar{p}_{+,j}) \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{J_{+,n}}{N_+} \sum_{j=1}^{J_{+,n}} (q_{+,j} - q_{+,j-1}) \sigma_+^2(\bar{q}_{+,j}) w(\bar{q}_{+,j}) \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{J_{+,n}}{n} \frac{1}{P_+} \int_{\bar{x}}^{x_u} \sigma_+^2(x) w(x) dx \{1 + o_{\mathbb{P}}(1)\} = \frac{J_{+,n}}{n} \mathscr{V}_{\mathtt{QS},+} \{1 + o_{\mathbb{P}}(1)\},
\end{aligned}$$

using Lemma C.2 and properties of the Riemann integral.

For the bias part, using the previous results and proceeding as in the proof of Theorem III.2, we have

$$\begin{aligned}
&\int_{\bar{x}}^{x_u} (\mathbb{E}[\hat{\mu}_{\mathtt{QS},+}(x; J_n)|\mathbf{X}_n] - \mu_+(x))^2 w(x) dx \\
&\quad = \frac{1}{12} \sum_{j=1}^{J_{+,n}} (p_{+,j} - p_{+,j-1})^3 \left(\mu_+^{(1)}(\bar{p}_{+,j}) \right)^2 w(\bar{p}_{+,j}) \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{1}{12} \sum_{j=1}^{J_{+,n}} (q_{+,j} - q_{+,j-1})^3 \left(\mu_+^{(1)}(\bar{q}_{+,j}) \right)^2 w(\bar{q}_{+,j}) \{1 + o_{\mathbb{P}}(1)\} \\
&\quad = \frac{1}{J_{+,n}^2} \frac{P_+^2}{12} \int_{\bar{x}}^{x_u} \left(\frac{\mu_+^{(1)}(x)}{f(x)} \right)^2 w(x) dx \{1 + o_{\mathbb{P}}(1)\} = J_{+,n}^{-2} \mathscr{B}_{\mathtt{QS},+} \{1 + o_{\mathbb{P}}(1)\},
\end{aligned}$$

because, for quantile-spaced partitions, expanding $F_+^{-1}(u)$ around $\bar{u} = F_+(\bar{q}_{+,j}) \in [(j-1)/J_{+,n}, j/J_{+,n}]$),

$$q_{+,j} - q_{+,j-1} = F_+^{-1}\left(\frac{j}{J_{+,n}}\right) - F_+^{-1}\left(\frac{j-1}{J_{+,n}}\right) = \frac{1}{f_+(\bar{q}_{+,j})} \frac{1}{J_{+,n}} \{1 + o_{\mathbb{P}}(1)\},$$

uniformly in $j = 1, 2, \cdots, J_{+,n}$, where $f_+(x) = \partial F_+(x)/\partial x = f(x)\mathbb{1}(x \geq \bar{x})/P_+$.

Putting the results together we verify the result for $\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})$. ■

Proofs of Theorem III.7 and Remark III.8. Proceeding as in the proofs of Theorem III.3 and Remark III.4, the results are established using Lemma C.3, $N_+/n \to_{\mathbb{P}} P_+$ and uniform consistency of power series estimators, as appropriate depending on the case. ■

Table C.1: Simulations Results for Model 1

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
6	1.558	2	4.621	6	1.558	2	4.621
7	1.295	3	2.246	7	1.295	3	2.246
8	1.146	4	1.474	8	1.146	4	1.474
9	1.061	5	1.166	9	1.061	5	1.166
10	1.017	6	1.040	10	1.017	6	1.040
11	1.000	7	1.000	11	1.000	7	1.000
12	1.001	8	1.006	12	1.001	8	1.006
13	1.014	9	1.039	13	1.014	9	1.039
14	1.037	10	1.088	14	1.037	10	1.088
15	1.067	11	1.149	15	1.067	11	1.149
16	1.102	12	1.216	16	1.102	12	1.216
$\hat{J}_{\mathtt{ES},-,n}$	1.005	$\hat{J}_{\mathtt{ES},+,n}$	1.020	$\hat{J}_{\mathtt{QS},-,n}$	1.030	$\hat{J}_{\mathtt{QS},+,n}$	1.029
$\check{J}_{\mathtt{ES},-,n}$	0.998	$\check{J}_{\mathtt{ES},+,n}$	1.016	$\check{J}_{\mathtt{QS},-,n}$	1.027	$\check{J}_{\mathtt{QS},+,n}$	1.024
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.000
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.005	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.020	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.022	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.028

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=11$	$\hat{J}_{\mathtt{ES},-,n}$	7	10	11	11.39	12	17	1.42
	$\check{J}_{\mathtt{ES},-,n}$	7	11	11	11.4	12	16	1.25
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=11$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	7	10	11	11.15	12	16	1.30
$J_{\mathtt{ES},+,n}=7$	$\hat{J}_{\mathtt{ES},+,n}$	5	6	7	7.052	8	12	0.93
	$\check{J}_{\mathtt{ES},+,n}$	5	7	7	7.034	7	11	0.78
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	5	6	7	7.006	7	11	0.84
$J_{\mathtt{QS},-,n}=11$	$\hat{J}_{\mathtt{QS},-,n}$	6	10	11	10.84	12	16	1.37
	$\check{J}_{\mathtt{QS},-,n}$	6	10	11	10.66	11	15	1.20
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=11$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	7	10	11	11.32	12	16	1.32
$J_{\mathtt{QS},+,n}=7$	$\hat{J}_{\mathtt{QS},+,n}$	4	6	7	7.096	8	12	1.04
	$\check{J}_{\mathtt{QS},+,n}$	5	7	7	7.128	8	11	0.92
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	5	6	7	6.939	7	11	0.85

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.2: Simulations Results for Model 2

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
5	1.820	1	15.305	6	1.538	3	3.158
6	1.421	2	3.995	7	1.283	4	1.946
7	1.206	3	1.979	8	1.137	5	1.425
8	1.088	4	1.336	9	1.056	6	1.175
9	1.026	5	1.090	10	1.015	7	1.053
10	1.000	6	1.000	11	1.000	8	1.000
11	0.997	7	0.984	12	1.003	9	0.987
12	1.009	8	1.008	13	1.018	10	0.998
13	1.033	9	1.054	14	1.042	11	1.025
14	1.064	10	1.115	15	1.072	12	1.064
15	1.102	11	1.185	16	1.108	13	1.110
$\hat{J}_{\mathtt{ES},-,n}$	1.042	$\hat{J}_{\mathtt{ES},+,n}$	1.014	$\hat{J}_{\mathtt{QS},-,n}$	1.055	$\hat{J}_{\mathtt{QS},+,n}$	1.053
$\check{J}_{\mathtt{ES},-,n}$	1.028	$\check{J}_{\mathtt{ES},+,n}$	1.007	$\check{J}_{\mathtt{QS},-,n}$	1.036	$\check{J}_{\mathtt{QS},+,n}$	1.019
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.096	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	0.994	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.038	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.055
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.194	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.017	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.030	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.053

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=10$	$\hat{J}_{\mathtt{ES},-,n}$	7	9	10	10.15	11	16	1.21
	$\check{J}_{\mathtt{ES},-,n}$	7	9	10	9.834	10	13	0.85
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=14$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	10	14	15	14.82	16	22	1.59
$J_{\mathtt{ES},+,n}=6$	$\hat{J}_{\mathtt{ES},+,n}$	4	6	6	6.536	7	11	0.93
	$\check{J}_{\mathtt{ES},+,n}$	5	6	6	6.526	7	10	0.79
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	4	6	7	6.873	7	12	1.10
$J_{\mathtt{QS},-,n}=11$	$\hat{J}_{\mathtt{QS},-,n}$	6	9	10	9.761	11	17	1.45
	$\check{J}_{\mathtt{QS},-,n}$	7	10	10	10.44	11	17	1.26
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=10$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	7	10	10	10.35	11	15	1.05
$J_{\mathtt{QS},+,n}=8$	$\hat{J}_{\mathtt{QS},+,n}$	5	7	7	7.588	8	13	1.13
	$\check{J}_{\mathtt{QS},+,n}$	6	7	8	8.14	9	13	1.02
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	5	7	7	7.486	8	12	1.00

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.3: Simulations Results for Model 3

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
8	1.324	3	2.907	8	1.324	3	2.907
9	1.180	4	1.816	9	1.180	4	1.816
10	1.091	5	1.352	10	1.091	5	1.352
11	1.039	6	1.136	11	1.039	6	1.136
12	1.011	7	1.037	12	1.011	7	1.037
13	1.000	8	1.000	13	1.000	8	1.000
14	1.002	9	0.999	14	1.002	9	0.999
15	1.013	10	1.021	15	1.013	10	1.021
16	1.031	11	1.057	16	1.031	11	1.057
17	1.055	12	1.103	17	1.055	12	1.103
18	1.083	13	1.157	18	1.083	13	1.157
$\hat{J}_{\mathtt{ES},-,n}$	1.007	$\hat{J}_{\mathtt{ES},+,n}$	1.013	$\hat{J}_{\mathtt{QS},-,n}$	1.006	$\hat{J}_{\mathtt{QS},+,n}$	1.022
$\check{J}_{\mathtt{ES},-,n}$	1.006	$\check{J}_{\mathtt{ES},+,n}$	1.003	$\check{J}_{\mathtt{QS},-,n}$	1.003	$\check{J}_{\mathtt{QS},+,n}$	1.010
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.000
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.019	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.006	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	0.988	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.016

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=13$	$\hat{J}_{\mathtt{ES},-,n}$	9	13	14	13.6	14	19	1.31
	$\check{J}_{\mathtt{ES},-,n}$	10	13	14	13.53	14	17	1.00
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=13$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	11	13	14	14.27	15	19	1.22
$J_{\mathtt{ES},+,n}=8$	$\hat{J}_{\mathtt{ES},+,n}$	5	7	8	7.753	8	12	0.98
	$\check{J}_{\mathtt{ES},+,n}$	6	7	8	7.719	8	11	0.79
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=8$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	6	7	8	7.864	8	12	0.96
$J_{\mathtt{QS},-,n}=13$	$\hat{J}_{\mathtt{QS},-,n}$	8	10	11	10.9	12	15	1.05
	$\check{J}_{\mathtt{QS},-,n}$	9	10	11	10.73	11	14	0.80
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=13$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	9	11	12	12.14	13	16	1.00
$J_{\mathtt{QS},+,n}=8$	$\hat{J}_{\mathtt{QS},+,n}$	5	7	8	8.07	9	13	0.91
	$\check{J}_{\mathtt{QS},+,n}$	6	8	8	8.218	9	11	0.69
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=8$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	6	8	8	8.129	9	12	0.80

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.4: Simulations Results for Model 4

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
7	1.387	2	5.131	8	1.310	5	1.771
8	1.207	3	2.460	9	1.170	6	1.391
9	1.101	4	1.580	10	1.085	7	1.187
10	1.040	5	1.219	11	1.035	8	1.077
11	1.010	6	1.061	12	1.009	9	1.022
12	1.000	7	1.000	13	1.000	10	1.000
13	1.004	8	0.990	14	1.003	11	1.001
14	1.019	9	1.010	15	1.016	12	1.016
15	1.042	10	1.048	16	1.035	13	1.043
16	1.070	11	1.099	17	1.060	14	1.077
17	1.103	12	1.158	18	1.089	15	1.117
$\hat{J}_{\mathtt{ES},-,n}$	1.024	$\hat{J}_{\mathtt{ES},+,n}$	1.012	$\hat{J}_{\mathtt{QS},-,n}$	1.035	$\hat{J}_{\mathtt{QS},+,n}$	1.040
$\check{J}_{\mathtt{ES},-,n}$	1.011	$\check{J}_{\mathtt{ES},+,n}$	1.007	$\check{J}_{\mathtt{QS},-,n}$	1.010	$\check{J}_{\mathtt{QS},+,n}$	1.011
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.062	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	0.971	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.009
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.141	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.001	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.007	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.044

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=12$	$\hat{J}_{\mathtt{ES},-,n}$	9	12	12	12.41	13	17	1.27
	$\check{J}_{\mathtt{ES},-,n}$	10	12	12	11.99	12	15	0.77
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=16$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	13	17	18	17.54	19	24	1.48
$J_{\mathtt{ES},+,n}=7$	$\hat{J}_{\mathtt{ES},+,n}$	5	7	7	7.555	8	12	1.07
	$\check{J}_{\mathtt{ES},+,n}$	5	7	7	7.5	8	13	0.91
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=8$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	5	7	8	7.705	8	12	0.97
$J_{\mathtt{QS},-,n}=13$	$\hat{J}_{\mathtt{QS},-,n}$	8	10	11	10.92	12	16	1.10
	$\check{J}_{\mathtt{QS},-,n}$	9	11	11	11.25	12	14	0.69
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=13$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	9	11	12	11.72	12	15	0.87
$J_{\mathtt{QS},+,n}=10$	$\hat{J}_{\mathtt{QS},+,n}$	5	8	9	8.933	10	14	1.25
	$\check{J}_{\mathtt{QS},+,n}$	7	9	9	9.547	10	15	1.09
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=9$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	6	8	9	8.663	9	14	1.05

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.5: Simulations Results for Model 5

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
6	1.558	2	4.621	6	1.558	2	4.621
7	1.295	3	2.246	7	1.295	3	2.246
8	1.146	4	1.474	8	1.146	4	1.474
9	1.061	5	1.166	9	1.061	5	1.166
10	1.017	6	1.040	10	1.017	6	1.040
11	1.000	7	1.000	11	1.000	7	1.000
12	1.001	8	1.006	12	1.001	8	1.006
13	1.014	9	1.039	13	1.014	9	1.039
14	1.037	10	1.088	14	1.037	10	1.088
15	1.067	11	1.149	15	1.067	11	1.149
16	1.102	12	1.216	16	1.102	12	1.216
$\hat{J}_{\mathtt{ES},-,n}$	1.012	$\hat{J}_{\mathtt{ES},+,n}$	1.029	$\hat{J}_{\mathtt{QS},-,n}$	1.030	$\hat{J}_{\mathtt{QS},+,n}$	1.028
$\check{J}_{\mathtt{ES},-,n}$	1.009	$\check{J}_{\mathtt{ES},+,n}$	1.022	$\check{J}_{\mathtt{QS},-,n}$	1.033	$\check{J}_{\mathtt{QS},+,n}$	1.019
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.000
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.008	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.024	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.022	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.025

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=11$	$\hat{J}_{\mathtt{ES},-,n}$	7	11	12	12	13	19	1.51
	$\check{J}_{\mathtt{ES},-,n}$	8	11	12	11.74	13	17	1.22
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=11$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	7	11	12	11.53	12	17	1.35
$J_{\mathtt{ES},+,n}=7$	$\hat{J}_{\mathtt{ES},+,n}$	4	6	7	7.099	8	11	1.00
	$\check{J}_{\mathtt{ES},+,n}$	5	7	7	7.041	7	11	0.82
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	4	6	7	6.99	8	11	0.91
$J_{\mathtt{QS},-,n}=11$	$\hat{J}_{\mathtt{QS},-,n}$	6	9	10	10.15	11	15	1.22
	$\check{J}_{\mathtt{QS},-,n}$	7	9	10	9.649	10	14	0.96
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=11$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	7	10	11	10.88	12	16	1.21
$J_{\mathtt{QS},+,n}=7$	$\hat{J}_{\mathtt{QS},+,n}$	4	6	7	6.781	7	11	0.90
	$\check{J}_{\mathtt{QS},+,n}$	4	6	7	6.672	7	10	0.70
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	4	6	7	6.782	7	10	0.80

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.6: Simulations Results for Model 6

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
5	1.820	1	15.305	6	1.538	3	3.158
6	1.421	2	3.995	7	1.283	4	1.946
7	1.206	3	1.979	8	1.137	5	1.425
8	1.088	4	1.336	9	1.056	6	1.175
9	1.026	5	1.090	10	1.015	7	1.053
10	1.000	6	1.000	11	1.000	8	1.000
11	0.997	7	0.984	12	1.003	9	0.987
12	1.009	8	1.008	13	1.018	10	0.998
13	1.033	9	1.054	14	1.042	11	1.025
14	1.064	10	1.115	15	1.072	12	1.064
15	1.102	11	1.185	16	1.108	13	1.110
$\hat{J}_{\mathtt{ES},-,n}$	1.010	$\hat{J}_{\mathtt{ES},+,n}$	1.020	$\hat{J}_{\mathtt{QS},-,n}$	1.076	$\hat{J}_{\mathtt{QS},+,n}$	1.066
$\check{J}_{\mathtt{ES},-,n}$	1.008	$\check{J}_{\mathtt{ES},+,n}$	1.013	$\check{J}_{\mathtt{QS},-,n}$	1.045	$\check{J}_{\mathtt{QS},+,n}$	1.046
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.084	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	0.988	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.051	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.046
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.112	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.011	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.044	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.043

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=10$	$\hat{J}_{\mathtt{ES},-,n}$	5	9	10	10.26	11	15	1.38
	$\check{J}_{\mathtt{ES},-,n}$	7	9	10	10.08	11	14	1.01
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=14$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	8	14	15	14.84	16	21	1.81
$J_{\mathtt{ES},+,n}=6$	$\hat{J}_{\mathtt{ES},+,n}$	3	6	7	6.612	7	11	1.09
	$\check{J}_{\mathtt{ES},+,n}$	4	6	6	6.508	7	11	0.88
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	4	6	7	6.946	8	13	1.11
$J_{\mathtt{QS},-,n}=11$	$\hat{J}_{\mathtt{QS},-,n}$	5	9	9	9.563	10	17	1.48
	$\check{J}_{\mathtt{QS},-,n}$	7	9	10	10.21	11	19	1.31
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=10$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	6	9	10	10.17	11	16	1.16
$J_{\mathtt{QS},+,n}=8$	$\hat{J}_{\mathtt{QS},+,n}$	4	7	8	8.447	10	19	2.19
	$\check{J}_{\mathtt{QS},+,n}$	5	7	9	9.031	10	19	2.08
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=7$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	5	7	7	7.723	8	14	1.36

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.7: Simulations Results for Model 7

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
8	1.324	3	2.907	8	1.324	3	2.907
9	1.180	4	1.816	9	1.180	4	1.816
10	1.091	5	1.352	10	1.091	5	1.352
11	1.039	6	1.136	11	1.039	6	1.136
12	1.011	7	1.037	12	1.011	7	1.037
13	1.000	8	1.000	13	1.000	8	1.000
14	1.002	9	0.999	14	1.002	9	0.999
15	1.013	10	1.021	15	1.013	10	1.021
16	1.031	11	1.057	16	1.031	11	1.057
17	1.055	12	1.103	17	1.055	12	1.103
18	1.083	13	1.157	18	1.083	13	1.157
$\hat{J}_{\mathtt{ES},-,n}$	0.995	$\hat{J}_{\mathtt{ES},+,n}$	1.021	$\hat{J}_{\mathtt{QS},-,n}$	1.063	$\hat{J}_{\mathtt{QS},+,n}$	0.987
$\check{J}_{\mathtt{ES},-,n}$	0.992	$\check{J}_{\mathtt{ES},+,n}$	1.012	$\check{J}_{\mathtt{QS},-,n}$	1.061	$\check{J}_{\mathtt{QS},+,n}$	0.973
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.000
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.006	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.014	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.023	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	0.990

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=13$	$\hat{J}_{\mathtt{ES},-,n}$	7	12	13	12.89	14	20	1.71
	$\check{J}_{\mathtt{ES},-,n}$	9	12	12	12.2	13	15	0.92
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=13$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	9	13	14	14.35	15	20	1.45
$J_{\mathtt{ES},+,n}=8$	$\hat{J}_{\mathtt{ES},+,n}$	4	7	8	8.106	9	12	1.12
	$\check{J}_{\mathtt{ES},+,n}$	6	7	8	7.963	8	11	0.81
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=8$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	5	7	8	8.199	9	13	1.07
$J_{\mathtt{QS},-,n}=13$	$\hat{J}_{\mathtt{QS},-,n}$	6	10	11	11.37	12	18	1.53
	$\check{J}_{\mathtt{QS},-,n}$	8	11	11	11.11	12	15	0.91
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=13$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	8	12	13	12.59	13	17	1.22
$J_{\mathtt{QS},+,n}=8$	$\hat{J}_{\mathtt{QS},+,n}$	4	8	9	8.822	10	13	1.11
	$\check{J}_{\mathtt{QS},+,n}$	6	8	9	8.913	9	11	0.74
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=8$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	5	8	8	8.149	9	11	0.80

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.8: Simulations Results for Model 8

Panel A: Integrated MSE for different partition sizes

$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{ES},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{ES},+,n})}$	$J_{-,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},-}(J_{-,n})}{\mathsf{IMSE}_{\mathtt{QS},-}(J_{\mathtt{QS},-,n})}$	$J_{+,n}$	$\frac{\mathsf{IMSE}_{\mathtt{QS},+}(J_{+,n})}{\mathsf{IMSE}_{\mathtt{QS},+}(J_{\mathtt{QS},+,n})}$
7	1.387	2	5.131	8	1.310	5	1.771
8	1.207	3	2.460	9	1.170	6	1.391
9	1.101	4	1.580	10	1.085	7	1.187
10	1.040	5	1.219	11	1.035	8	1.077
11	1.010	6	1.061	12	1.009	9	1.022
12	1.000	7	1.000	13	1.000	10	1.000
13	1.004	8	0.990	14	1.003	11	1.001
14	1.019	9	1.010	15	1.016	12	1.016
15	1.042	10	1.048	16	1.035	13	1.043
16	1.070	11	1.099	17	1.060	14	1.077
17	1.103	12	1.158	18	1.089	15	1.117
$\hat{J}_{\mathtt{ES},-,n}$	1.026	$\hat{J}_{\mathtt{ES},+,n}$	1.011	$\hat{J}_{\mathtt{QS},-,n}$	1.093	$\hat{J}_{\mathtt{QS},+,n}$	1.050
$\check{J}_{\mathtt{ES},-,n}$	1.007	$\check{J}_{\mathtt{ES},+,n}$	1.000	$\check{J}_{\mathtt{QS},-,n}$	1.041	$\check{J}_{\mathtt{QS},+,n}$	1.008
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.081	$J^{\mathtt{dw}}_{\mathtt{ES},+,n}$	0.981	$J^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.000	$J^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.025
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	1.131	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	1.001	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	1.040	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	1.038

Panel B: Summary Statistics for the Estimated Partition Sizes

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
$J_{\mathtt{ES},-,n}=12$	$\hat{J}_{\mathtt{ES},-,n}$	7	11	12	12.44	13	18	1.56
	$\check{J}_{\mathtt{ES},-,n}$	9	12	12	12.39	13	16	1.03
$J^{\mathtt{dw}}_{\mathtt{ES},-,n}=16$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},-,n}$	12	16	18	17.63	19	24	1.75
$J_{\mathtt{ES},+,n}=7$	$\hat{J}_{\mathtt{ES},+,n}$	4	7	7	7.393	8	14	1.12
	$\check{J}_{\mathtt{ES},+,n}$	5	7	7	7.341	8	11	0.84
$J^{\mathtt{dw}}_{\mathtt{ES},+,n}=8$	$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},+,n}$	5	7	8	7.806	8	13	1.07
$J_{\mathtt{QS},-,n}=13$	$\hat{J}_{\mathtt{QS},-,n}$	6	10	11	10.61	11	16	1.34
	$\check{J}_{\mathtt{QS},-,n}$	9	11	11	11.24	12	15	0.87
$J^{\mathtt{dw}}_{\mathtt{QS},-,n}=13$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},-,n}$	8	11	11	11.46	12	16	1.03
$J_{\mathtt{QS},+,n}=10$	$\hat{J}_{\mathtt{QS},+,n}$	4	8	9	8.737	10	16	1.29
	$\check{J}_{\mathtt{QS},+,n}$	6	9	9	9.471	10	14	1.01
$J^{\mathtt{dw}}_{\mathtt{QS},+,n}=9$	$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},+,n}$	5	8	9	8.79	9	14	1.07

Notes:

(i) Population quantities:
$J_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.2),
$J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = optimal partition size for evenly-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.5),
$J_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=1$ (Theorem III.6),
$J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = optimal partition size for quantile-spaced (ES) RD-plot with $w(x)=f(x)$ (Remark III.9).

(ii) Estimators:
$\hat{J}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J_{\mathtt{ES},\cdot,n}$ (Theorem III.3),
$\check{J}_{\mathtt{ES},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{ES},\cdot,n}$ (Remark III.4),
$\hat{J}^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{ES},\cdot,n}$ (Remark III.5),
$\hat{J}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J_{\mathtt{QS},\cdot,n}$ (Theorem III.7),
$\check{J}_{\mathtt{QS},\cdot,n}$ = polynomial regression estimator of $J_{\mathtt{QS},\cdot,n}$ (Remark III.8),
$\hat{J}^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ = spacings estimator of $J^{\mathtt{dw}}_{\mathtt{QS},\cdot,n}$ (Remark III.9).

Table C.9: Comparison of Partitioning Schemes

	$\frac{B_{\mathtt{ES},-}}{B_{\mathtt{QS},-}}$	$\frac{V_{\mathtt{ES},-}}{V_{\mathtt{QS},-}}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{QS},-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{QS},-,n})}$	$\frac{B_{\mathtt{ES},+}}{B_{\mathtt{QS},+}}$	$\frac{V_{\mathtt{ES},+}}{V_{\mathtt{QS},+}}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{QS},+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{QS},+,n})}$
Model 1	1.000	1.000	1.000	1.000	1.000	1.000
Model 2	1.031	1.234	1.166	0.541	1.234	0.940
Model 3	1.000	1.000	1.000	1.000	1.000	1.000
Model 4	1.031	1.321	1.216	0.541	1.321	0.990
Model 5	1.000	1.000	1.000	1.000	1.000	1.000
Model 6	1.031	1.234	1.166	0.541	1.234	0.940
Model 7	1.000	1.000	1.000	1.000	1.000	1.000
Model 8	1.031	1.321	1.216	0.541	1.321	0.990

(a) IMSE with uniform weighting ($w(x) = 1$).

	$\frac{B_{\mathtt{ES},-}}{B_{\mathtt{QS},-}}$	$\frac{V_{\mathtt{ES},-}}{V_{\mathtt{QS},-}}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{QS},-,n})}{\mathsf{IMSE}_{\mathtt{ES},-}(J_{\mathtt{QS},-,n})}$	$\frac{B_{\mathtt{ES},+}}{B_{\mathtt{QS},+}}$	$\frac{V_{\mathtt{ES},+}}{V_{\mathtt{QS},+}}$	$\frac{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{QS},+,n})}{\mathsf{IMSE}_{\mathtt{ES},+}(J_{\mathtt{QS},+,n})}$
Model 1	1.000	1.000	1.000	1.000	1.000	1.000
Model 2	2.290	1.000	1.309	0.784	1.000	0.908
Model 3	1.000	1.000	1.000	1.000	1.000	1.000
Model 4	2.290	1.137	1.438	0.784	1.137	1.004
Model 5	1.000	1.000	1.000	1.000	1.000	1.000
Model 6	2.290	1.000	1.309	0.784	1.000	0.908
Model 7	1.000	1.000	1.000	1.000	1.000	1.000
Model 8	2.290	1.137	1.438	0.784	1.137	1.004

(b) IMSE with design/density weighting ($w(x) = f(x)$).

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235–267.

——— (2010): "Estimation of the Conditional Variance in Paired Experiments," *Annales d'Économie et de Statistique*, 91(1), 175–187.

Altonji, J., and R. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73(4), 1053–1102.

Angrist, J. D., and V. Lavy (1999): "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114(2), 533–575.

Aras, G., S. R. Jammalamadaka, and X. Zhou (1989): "Limit Distribution of Spacings Statistics when the Sample Size is Random," *Statistics and Probability Letters*, 8(5), 451–456.

Atteberry, A., S. Loeb, and J. Wyckoff (2013): "Do first impressions matter? Improvement in early career teacher effectiveness," Discussion paper, National Bureau of Economic Research.

Baryshnikov, Y., M. D. Penrose, and J. E. Yurich (2009): "Gaussian Limits For Generalized Spacings," *Annals of Applied Probability*, 19(1), 158–185.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2013): "On the Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, forthcoming.

Boucher, V., Y. Bramoullé, H. Djebbari, and B. Fortin (2012): "Do peers affect student achievement? Evidence from Canada using group size variation," *Journal of Applied Econometrics*.

Bramoulle, Y., H. Djebbari, and B. Fortin (2009): "Identification of peer effects through social networks," *Journal of Econometrics*, 150(1), 41–55.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2014): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Estimation," working paper, University of Michigan.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014a): "Optimal Data-Driven Regression Discontinuity Plots," working paper, University of Michigan.

——— (2014b): "Robust Data-Driven Inference in the Regression-Discontinuity Design," revision requested by *Stata Journal*.

——— (2014c): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," working paper, University of Michigan.

——— (2014d): "Supplement to Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," supplemental material, University of Michigan.

——— (2014e): "`rdrobust`: An R Package for Robust Inference in Regression-Discontinuity Designs," in preparation for *Journal of Statistical Software*.

Card, D., and A. B. Krueger (1992): "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100(1), 1–40.

Card, D., and A. B. Krueger (1996): "School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina," *Journal of Economic Perspectives*, 10(4), 31–50.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2012): "Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in a Generalized Regression Kink Design," Working paper, University of California at Berkeley.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010): "Robust Data-Driven Inference for Density-Weighted Average Derivatives," *Journal of the American Statistical Association*, 105(491), 1070–1083.

——— (2013a): "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*.

——— (2013b): "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," *Econometric Theory*, 1, 1–25.

Cattaneo, M. D., and M. H. Farrell (2013a): "Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators," *Journal of Econometrics*, 174(2), 127–143.

——— (2013b): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174(2), 127–143.

Cattaneo, M. D., B. Frandsen, and R. Titiunik (2014): "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," revision requested by *Journal of Causal Inference*.

Chen, X. (2007a): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," vol. 6, Part B of *Handbook of Econometrics*, pp. 5549 – 5632. Elsevier.

——— (2007b): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, ed. by J. J. Heckman, and E. Leamer, pp. 5549–5632. Elsevier Science B.V., New York.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On Automatic Boundary Corrections," *Annals of Statistics*, 25(4), 1691–1708.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on counterfactual distributions," *Econometrica*, 81(2), 535–580.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 126(4), 1593–1660.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2007): "How and Why do Teacher Credentials Matter for Student Achievement?," NBER Working Papers 12828, National Bureau of Economic Research, Inc.

——— (2010): "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects," *Journal of Human Resources*, 45(3).

Cook, T. D. (2008): ""Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics," *Journal of Econometrics*, 142(2), 636–654.

Cunningham, B., and D. W. Andrews (1988): "The relationship of achievement and peer status to teacher attitudes toward young children," *Early Child Development and Care*, 30(1-4), 85–95.

David, H. A., and H. N. Nagaraja (1998): "Concomitants of Order Statistics," in *Handbook of Statistics*, ed. by N. Balakrishnan, and C. R. Rao, vol. 16, pp. 487–513. Elsevier Science B.V.

——— (2003): *Order Statistics*. Wiley, New Jersey.

Dee, T. S. (2004): "Teachers, Race, and Student Achievement in a Randomized Experiment," *The Review of Economics and Statistics*, 86(1), 195–210.

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

Djebbari, H., and J. Smith (2008): "Heterogeneous impacts in PROGRESA," *Journal of Econometrics*, 145(1-2), 64–80.

Dong, Y. (2012): "Jumpy or Kinky? Regression Discontinuity Without The Discontinuity," working paper, University of California at Irvine.

Dong, Y., and A. Lewbel (2012): "Regression Discontinuity Marginal Threshold Treatment Effects," working paper, Boston College.

Durlauf, S. N. (2004): "Neighborhood effects," 4, 2173–2242.

Dynarski, S., J. M. Hyman, and D. W. Schanzenbach (2011): "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion," (17533).

Evdokimov, K. (2010): "Identification and estimation of a nonparametric panel data model with unobserved heterogeneity," *Department of Economics, Princeton University*.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. Chapman Hall/CRC, New York.

Fan, J., and T. H. Yim (2004): "A crossvalidation method for estimating conditional densities," *Biometrika*, 91(4), 819–834.

Finn, J., S. Gerber, and J. Boyd-Zaharias (2005): "Small classes in the early grades, academic achievement, and graduating from high school.," *Journal of Educational Psychology*, 97(2), 214.

Foresi, S., and F. Paracchi (1992): "The Conditional Distribution of Excess Returns: An Empirical Analysis," Discussion paper.

Fortin, N., T. Lemieux, and S. Firpo (2011): "Decomposition methods in economics," *Handbook of Labor Economics*, 4, 1–102.

Frandsen, B., M. Frölich, and B. Melly (2012): "Quantile treatments effects in the regression discontinuity design," *Journal of Econometrics*, 168(2), 382–395.

Ghosh, K., and R. Jammalamadaka (2001): "A General Estimation Method Using Spacing," *Journal of Statistical Planning and Inference*, 93(1), 71–82.

Glewwe, P. (1997): "Estimating the impact of peer group effects on socioeconomic outcomes: Does the distribution of peer group characteristics matter?," *Economics of Education Review*, 16(1), 39–43.

Graham, B. S., G. W. Imbens, and G. Ridder (2010): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," NBER Working Papers 16499, National Bureau of Economic Research, Inc.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

Hall, P., J. Racine, and Q. Li (2004): "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24(3), 1141–77.

Hanushek, E. A. (2003): "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113(485), F64–F98.

Harris, D. N., and T. R. Sass (2011): "Teacher training, teacher quality and student achievement," *Journal of Public Economics*, 95(78), 798 – 812.

Heckman, J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4), 487–535.

Hoderlein, S., and H. White (2012): "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Journal of Econometrics*.

Horowitz, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics, Volume V*, ed. by J. Heckman, and E. Leamer, pp. 3159–3228. Elsevier Science B.V., New York.

Hoxby, C. M. (2000): "The Effects Of Class Size On Student Achievement: New Evidence From Population Variation," *The Quarterly Journal of Economics*, 115(4), 1239–1285.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 74. Elsevier.

Imbens, G., and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142(2), 615–635.

Imbens, G. W., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79(3), 933–959.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Jackson, E., and M. E. Page (2013): "Estimating the distributional effects of education reforms: A look at Project STAR," *Economics of Education Review*, 32, 92–103.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008): "What does certification tell us about teacher effectiveness? Evidence from New York City," *Economics of Education Review*, 27(6), 615–631.

Krueger, A. B. (1999): "Experimental Estimates Of Education Production Functions," *The Quarterly Journal of Economics*, 114(2), 497–532.

Krueger, A. B., and D. M. Whitmore (2001): "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal*, 111(468), 1–28.

Lavy, V., and A. Schlosser (2011): "Mechanisms and Impacts of Gender Peer Effects at School," *American Economic Journal: Applied Economics*, 3(2), 1–33.

Lee, D. S. (2008): "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, 142(2), 675–697.

Lee, D. S., and D. Card (2008): "Regression discontinuity inference with specification error," *Journal of Econometrics*, 142(2), 655–674.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(2), 281–355.

Lewbel, A., and S. Schennach (2007): "A Simple Ordered Data Estimator for Inverse Density Weighted Expectations," *Journal of Econometrics*, 136(1), 189–211.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set," *Annals of Statistics*, 15(3), 958–975.

Ludwig, J., and D. L. Miller (2007): "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122(1), 159–208.

Manski, C. F. (1993): "Identification of Endogenous Social Effects The Reflection Problem," *Review of Economic Studies*, 60(3), "531–42".

Marmer, V., D. Feir, and T. Lemieux (2012): "Weak Identification in Fuzzy Regression Discontinuity Designs," working paper, Univeristy of British Columbia.

Mason, D. M. (1984): "A Strong Limit Theorem for the Oscillation Modulus of the Uniform Empirical Quantile Process," *Stochastic Processes and its Applications*, 17(1), 127–136.

Matzkin, R. (2007): "Nonparametric identification," *Handbook of Econometrics*, 6, 5307–5368.

Matzkin, R. L. (2013): "Nonparametric Identification in Structural Economic Models," *Annual Review of Economics*, 5(1), 457–486.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142(2), 698–714.

Mueller, S. (2013): "Teacher experience and the class size effect Experimental evidence," *Journal of Public Economics*, 98(C), 44–52.

Newey, W. K. (1997a): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), 147–168.

——— (1997b): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and T. M. Stoker (1993): "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61(5), 1199–223.

Nye, B., S. Konstantopoulos, and L. V. Hedges (2004): "How Large Are Teacher Effects?," *Educational Evaluation and Policy Analysis*, 26(3), 237–257.

Otsu, T., and K.-L. Xu (2013): "Empirical Likelihood for Regression Discontinuity Design," working paper, LSE.

Porter, J. (2003): "Estimation in the Regression Discontinuity Model," working paper, University of Wisconsin.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005): "Teachers, Schools, and Academic Achievement," *Econometrica*, 73(2), 417–458.

Rockoff, J. E. (2004): "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94(2), 247–252.

Rothe, C. (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, 155(1), 56–70.

——— (2012): "Partial distributional policy effects," *Econometrica*, 80(5), 2269–2301.

Sacerdote, B. (2011): *Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?*vol. 3 of *Handbook of the Economics of Education*, chap. 4, pp. 249–277. Elsevier.

Schanzenbach, D. W. (2006): "What have researchers learned from Project STAR," *Brookings papers on education policy*, 2006(1), 205–228.

Shorack, G., and J. Wellner (2009): *Empirical Process with Applications to Statistics*. Siam.

Su, L., S. Hoderlein, and H. White (2010): "Testing Monotonicity in Unobservables with Panel Data," Discussion paper, Working paper, Singapore Management University. 6.

Thistlethwaite, D. L., and D. T. Campbell (1960): "Regression-discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," *Journal of Educational Psychology*, 51(6), 309–317.

van der Klaauw, W. (2008): "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2), 219–245.

Wand, M., and M. Jones (1995): *Kernel Smoothing*. Chapman Hall/CRC, Florida.

Weyl, H. (1939): *The Classical Groups: Their Invariants and Representations*, [Princeton mathematical series 1]. Princeton University Press.

Whitmore, D. (2005): "Resource and Peer Impacts on Girls Academic Achievement: Evidence from a Randomized Experiment," *American Economic Review*, 95(2), 199–203.