CARRA Working Paper Series

Working Paper *#2015-06*

Estimation and Inference in Regression Discontinuity Designs with Clustered Sampling

Otávio Bartalotti
Iowa State University

Quentin Brummet
U.S. Census Bureau

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: August 27, 2015

Estimation and Inference in Regression Discontinuity Designs with Clustered Sampling *

Otávio Bartalotti and Quentin Brummet†

August 20, 2015

Abstract

Regression Discontinuity (RD) designs have become popular in empirical studies due to their attractive properties for estimating causal effects under transparent assumptions. Nonetheless, most popular procedures assume i.i.d. data, which is not reasonable in many common applications. To relax this assumption, we derive the properties of traditional non-parametric estimators in a setting that incorporates potential clustering at the level of the running variable, and propose an accompanying optimal-MSE bandwidth selection rule. Simulation results demonstrate that falsely assuming data are i.i.d. when selecting the bandwidth may lead to the choice of bandwidths that are too small relative to the optimal-MSE bandwidth. Last, we apply our procedure using person-level microdata that exhibits clustering at the census tract level to analyze the impact of the Low-Income Housing Tax Credit program on neighborhood characteristics and low-income housing supply.

Keywords: Regression discontinuity designs, Local polynomials, Clustering, Optimal bandwidth selection

JEL: C13, C14, C21

*We are grateful to Gary Solon, Helle Bunzel, Valentin Verdier, Thomas Fujiwara, Maggie Jones and participants at the 2014 North American Summer Meetings of the Econometric Society, 2014 Midwest Econometrics Group, 2015 Econometric Society World Congress and U.S. Census Bureau for helpful comments. The views expressed within are those of the authors and not necessarily those of the U.S. Census Bureau.

†Bartalotti: Department of Economics, Iowa State University. 260 Heady Hall, Ames, IA 50011. Email: bartalot@iastate.edu. Brummet: Center for Administrative Records Research and Applications, United States Census Bureau. 4600 Silver Hill Road, Washington, DC 20233. Email: quentin.o.brummet@census.gov.

1 Introduction

Regression Discontinuity (RD) designs have become one of the leading empirical strategies in economics, public policy evaluation, and other social sciences. While these designs provide consistent estimation of causal effects under transparent assumptions, the current literature on estimation and inference in RD designs typically assumes that the observations around the cutoff are independent and identically distributed,[1] which limits the applicability of such procedures in at least two relevant empirical settings. First, researchers may wish to use microdata to implement a RD design based on a higher-level running variable. [2] An estimation and inference procedure that assumes clustering at the level of the running variable allows the researcher to estimate parameters, select bandwidths, and perform inference in a way that is compatible with the use of microdata in RD designs. Another salient example of such an application is an RD design with a discrete running variable. Following the advice of Lee and Card (2008), researchers implementing RD designs in applications with discrete running variables typically conduct inference using cluster-robust standard errors.[3] This inference procedure directly contradicts with commonly used bandwidth selection procedures that assume i.i.d. data, however.[4] Therefore researchers performing RD designs with discrete running variables are left with the choice of either using an *ad hoc* bandwidth or relying on a bandwidth selection procedure whose assumptions are clearly violated.[5]

In this study, we derive asymptotic distributions for local polynomial estimators of treatment effects in RD designs under a setup that allows for unrestricted dependence among observations within clusters defined at the running variable level. These results demonstrate that the widely used "cluster-robust" standard errors are appropriate in this setting. This finding relates to the results found in Lee and Card (2008), who suggest the use of cluster-robust standard errors to

[1] See, for example, Hahn, Todd, and Van der Klaauw (2001), Porter (2003), Ludwig and Miller (2007), Imbens and Kalyanaraman (2012), or Calonico, Cattaneo, and Titiunik (2014).

[2] For example, researchers could use student-level microdata to examine a policy implemented based on a school-level running variable.

[3] These applications are too numerous to adequately summarize here, but recent examples of studies that use a variety of discrete running variables include birth weight (Almond et al., 2010), days until unemployment cutoffs (Schmieder, Von Wachter, and Bender, 2012), prison inmate security scores (Chen and Shapiro, 2007), discrete test scores (Scott-Clayton, 2011), age (Card, Dobkin, and Maestas, 2008), and date of birth (Dobkin and Ferreira, 2010; Elder, 2010).

[4] In particular, the bandwidth selection procedure developed by Imbens and Kalyanaraman (2012) is very widely used by applied researchers. For example, a recent Google Scholar search returns over 600 articles citing Imbens and Kalyanaraman (2012), the majority of which are empirical applications.

[5] As discussed in Lee and Card (2008), non-parametric identification in the RD design is infeasible with a discrete running variable, and the clustered standard errors used by researchers are intended to correct for specification error in the conditional mean function. Nevertheless, this procedure still contrasts with a bandwidth selection procedure assuming i.i.d. data and our approximation to the data generating process provides a transparent, data-driven bandwidth selection procedure for practitioners in these cases.

account for specification errors in a specific class of models that are amenable to parametric RD designs. Our analysis demonstrates that in the context of our model, the intuitive idea of using cluster-robust standard errors holds even when using non-parametric local polynomial estimators.

In addition, we propose an optimal bandwidth selection procedure in RD designs with dependence among observations. The procedure extends Imbens and Kalyanaraman (2012) (henceforth, "IK") by allowing for clustered sampling with unrestricted dependence structure within cluster, and the resulting optimal bandwidth estimator collapses to traditional optimal bandwidth estimators when observations are i.i.d. We provide a simple implementation of the algorithm and perform a small simulation study demonstrating that our procedure outperforms traditional bandwidth choices in terms of Mean Squared Error (MSE) in many practical settings.

Finally, we demonstrate the empirical importance and usefulness of the procedure in an application analyzing the impact of Low-Income Housing Tax Credits (LIHTC) on neighborhood characteristics. The data in this application are person level, but the running variable is defined at the census tract level, generating clustering issues. The results show that accounting for this clustering in the data when choosing bandwidths can lead to practically significant changes in the interpretation of the empirical results.

The remainder of the paper is structured as follows. Section 2 presents the setup and Section 3 presents our main results. Section 4 then provides a small simulation study. Finally, Section 5 presents the application to the impacts of Low-Income Housing Tax Credits on neighborhood characteristics, and Section 6 concludes.

2 Setup

2.1 General RD Design

In the typical sharp RD setting, a researcher wishes to estimate the local causal effect of treatment at a given threshold. The running variable, X_i, determines treatment assignment. Given a known threshold, $\bar{x}$, set to zero without loss of generality, a unit receives treatment if $X_i \geq 0$ or does not receive treatment if $X_i < 0$. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for unit i given it receives treatment and in the absence of treatment, respectively. Hence, the observed sample is comprised of the running variable, X_i, and

$$Y_i = Y_i(0)\mathbb{1}\{X_i < 0\} + Y_i(1)\mathbb{1}\{X_i \geq 0\} \tag{1}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. For convenience, define

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] \tag{2}$$

In most cases the population parameter of interest is $\tau = \mathbb{E}[Y(1) - Y(0)|X = \bar{x}]$ (i.e., the average treatment effect at the threshold). Under continuity and smoothness conditions on both the conditional distribution of X_i and the first moments of $Y(0)$ and $Y(1)$ at the cutoff,[6] τ is nonparametrically identified (Hahn, Todd, and Van der Klaauw, 2001) by:

$$\begin{aligned}
&\tau = \mu_+ - \mu_- \\
&\quad \text{where } \mu_+ = lim_{x \to 0^+} \mu(x), \text{ and } \mu_- = lim_{x \to 0^-} \mu(x)
\end{aligned} \tag{3}$$

In general one might also be interested in the discontinuity of a higher order derivative of the conditional expectation at the threshold.[7] Let $\mu^{(\eta)}(x) = \frac{d^\eta \mu(x)}{dx^\eta}$ be the η^{th} derivative of the unknown regression function and define $\mu_+^{(\eta)} = lim_{x \to 0^+} \mu^{(\eta)}(x)$ and $\mu_-^{(\eta)} = lim_{x \to 0^-} \mu^{(\eta)}(x)$. The parameter of interest in those cases is given by $\tau^{(\eta)} = \mu_+^{(\eta)} - \mu_-^{(\eta)}$.

The estimation of $\tau^{(\eta)}$ in RD designs focuses on the problem of approximating $\mathbb{E}[Y(1)|X = x]$ and $\mathbb{E}[Y(0)|X = x]$ near the cutoff. Due to its desirable properties when estimating regression functions at the boundary, the most common approach fits separate kernel-weighted local polynomial regressions in neighborhoods on both sides of the threshold.[8] For a local polynomial of order p, we use the following estimator:

$$\begin{aligned}
\hat{\tau}^{(\eta)} &= \hat{\mu}_+^{(\eta)} - \hat{\mu}_-^{(\eta)} \\
(\hat{\beta}_+, \hat{\beta}_+^{(1)}, \ldots, \hat{\beta}_+^{(p)})' &= argmin_{b_0, b_1, \ldots, b_p} \sum_{i=1}^{N} \mathbb{1}\{X_i \geq 0\}(Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2 \cdot K_h(X_i) \\
(\hat{\beta}_-, \hat{\beta}_-^{(1)}, \ldots, \hat{\beta}_-^{(p)})' &= argmin_{b_0, b_1, \ldots, b_p} \sum_{i=1}^{N} \mathbb{1}\{X_i < 0\}(Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2 \cdot K_h(X_i)
\end{aligned}$$

where $K_h(x_{ig}) = K\left(\frac{x_{ig}}{h}\right)\frac{1}{h}$ and $\hat{\mu}^{(\eta)} = \eta! \hat{\beta}^{(\eta)}$.

[6] The assumptions used in the derivations and results presented here closely follow IK and are discussed in Appendix A.1.

[7] See, for example, the "regression kink" literature (Card, Lee, and Pei, 2009).

[8] See Hahn, Todd, and Van der Klaauw (2001), Porter (2003) or Fan and Gijbels (1992) for discussions of the properties of local polynomial regressions for boundary problems.

2.2 Clustering in RD Designs

Building on this traditional RD setup, we now turn to the setting where clustering exists at the level of the running variable. Consider sampling from a large number of clusters and, for each group g, we observe data on the outcome, running variable and potential covariates for N_g observations.[9] This sampling scheme is assumed to generate observations that are independent across clusters. Then, for a random sample of G groups of fixed size N_g, we observe

$$Y_{ig} = \mu(x_{ig}) + \epsilon_{ig} \tag{4}$$

Where the subscript ig refers to unit i in cluster g. The asymptotic theory developed below assumes that the number of clusters increases while cluster size is held fixed and the bandwidth shrinks (i.e., $G \to \infty$, $h \to 0$, and $Gh \to \infty$). We analyze inference and the optimal choice of bandwidth in RD designs under clustering, letting $Var(Y|X) = I_G \otimes \Omega(x)$, where its elements, Ω_{ij}, are denoted as $\sigma_{ij}(x)$, and its limits $lim_{x\to 0^+}\sigma_{ij}(x) = \sigma^+_{ij}$ and $lim_{x\to 0^-}\sigma_{ij}(x) = \sigma^-_{ij}$ throughout the paper.[10]

3 Main Results

3.1 Asymptotic Distribution

Given this setup, we derive the asymptotic properties of $\hat{\tau}^{(\eta)}$ and the validity of usual tests. Let $\nu_j = \int_0^\infty u^j K(u)du$ and $\pi_j = \int_0^\infty u^j K^2(u)du$ be deterministic functions of the kernel function chosen by the researcher. Additionally, define Γ and Δ as $(p+1)\times(p+1)$ matrices with element (i, j) given by ν_{i+j-2} and π_{i+j-2}, respectively. Assumptions for the results presented below include the standard smoothness conditions of the conditional expectation and variance of Y around the cutoff found in the RD literature and other regularity conditions, and are described in Appendix A.1. Proofs are collected in Appendix A.2.

Lemma 3.1. *Suppose assumptions 1-5 hold and $Gh \to \infty$.*

[9] This reflects the standard clustered data setup as discussed in Wooldridge (2010).

[10] An alternative question is whether asymptotic approximations with $N_g \to \infty$ and $G \to \infty$ following Hansen (2007) can provide additional insight. This is beyond the scope of this paper.

1. ***(B)** If $h \to 0$, then*

$$E[\hat{\tau}^{(\eta)}|X] = \tau^{(\eta)} + \eta! \frac{h^{p+1-\eta}}{(p+1)!} \left(\mu_{+}^{(p+1)} - (-1)^{(p+1+\eta)} \mu_{-}^{(p+1)} \right) e_{\eta} \Gamma^{-1} \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} + o_p(h^{p+1-\eta})$$

2. ***(V)** If $h \to 0$, then*

$$Var[\hat{\tau}^{(\eta)} - \tau^{(\eta)}|X] = \left[\eta!^2 \left(\frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{+}}{G N_g^2 h^{2\eta+1} f(0)} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{-}}{G N_g^2 h^{2\eta+1} f(0)} \right) e_{\eta} \Gamma^{-1} \Delta \Gamma^{-1} e_{\eta} \right] \{1 + o_p(1)\}$$

3. ***(D)** If $Gh^{2p+3} \to 0$, then*

$$\frac{\hat{\tau}_{+}^{(\eta)} - \tau^{(\eta)}}{\sqrt{Var[\hat{\tau}^{(\eta)} - \tau^{(\eta)}|X]}} \to_d N(0,1)$$

Hence, the traditional standardized t-statistic and the conventional confidence intervals are asymptotically valid. Note that this asymptotic variance formula relates closely with the typical cluster-robust standard error formulas and suggests that these estimators can be used in RD studies utilizing a non-parametric local polynomial estimator with clustering at the running variable level.[11]

In a more general setting, one could face a situation where clusters contain observations with different values of the running variable. In that case, the covariance terms in the asymptotic variance would vanish under the current normalization, and the clustering issue would disappear asymptotically.[12] This result is similar to the situation described by Bhattacharya (2005) in the context of multi-stage sampling. Intuitively, as the number of clusters increases and the bandwidth shrinks around the threshold, the proportion of units from a given cluster within the bandwidth goes to zero.[13] However, as noted in Bhattacharya (2005), in empirical applications with finite sample size and nonzero bandwidth, the vanishing clustering may not be ignorable. Therefore, even in a general clustering setup, practitioners may with to implement cluster-robust methods for inference and bandwidth choice.

[11] This is the cluster analogue of the point made by Imbens and Lemieux (2008) that usual parametric heteroskedasticity-robust standard errors can be used in traditional RD designs with i.i.d. data.

[12] Calculations demonstrating this result are available from the authors upon request.

[13] This is not an issue in the current setup because we focus our discussion on the case where clusters are defined at the level of the running variable, X and clustering does not vanish asymptotically.

3.2 Optimal Bandwidth Selection

3.2.1 Infeasible Optimal Bandwidth Choice

This section derives the optimal bandwidth choice for RD designs with clustered sampling. As pointed out by IK, the local nature of RD designs makes it desirable to define our error criteria in terms of the quality of the local approximation to the conditional expectations at the cutoff. We obtain an optimal bandwidth h^* that minimizes $MSE(h)$:

$$MSE(h) = \mathbb{E}\left[(\hat{\tau} - \tau)^2\right] \tag{5}$$

Lemma 3.2. *Suppose assumptions 1-5 in Appendix A.1 hold. Then,*

1. *(**MSE**)*

$$\begin{aligned} MSE(h) = {} & \frac{1}{Gh^{2\eta+1}} C_{2,\eta} \left(\frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{+}}{N_g^2 f(0)} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{-}}{N_g^2 f(0)} \right) \\ & + h^{2(p+1-\eta)} C_{1,\eta} \left[\mu_{+}^{(p+1)} - (-1)^{(p+1)} \mu_{-}^{(p+1)}\right]^2 + o_p\left(\frac{1}{Gh^{2\eta+1}} + h^{2(p+1-\eta)}\right) \end{aligned}$$

Where $C_{1,\eta} = \left[\frac{\eta!}{(p+1)!} e_\eta \Gamma^{-1} \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix}\right]^2$ *and* $C_{2,\eta} = \eta!^2 e_\eta \Gamma^{-1} \Delta \Gamma^{-1} e_\eta$.

2. *(**Optimal Bandwidth**) If $\mu_{+}^{(p+1)} \neq \mu_{-}^{(p+1)}$, then the optimal bandwidth that minimizes the asymptotic approximation to $MSE(h)$ is*

$$h_{opt} = \left[C_{\kappa\eta} \frac{\frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{+}}{N N_g f(0)} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{-}}{N N_g f(0)}}{\left[\mu_{+}^{(p+1)} - (-1)^{(p+1)} \mu_{-}^{(p+1)}\right]^2} \right]^{\frac{1}{2p+3}} \tag{6}$$

where $C_{\kappa\eta} = \dfrac{(p+1)!^2 (2\eta+1) e'_\eta \Gamma^{-1} \Delta \Gamma^{-1} e_\eta}{2(p+1-\eta) \left[e'_\eta \Gamma^{-1} \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} \right]^2}$.

This lemma extends the results in IK to the case in which data is clustered. Comparing Equation (6) to the infeasible bandwidth choice in IK, the numerator includes additional variance terms that allow for dependence of observations within cluster. Additionally, if the errors are indeed i.i.d., this bandwidth collapses to the IK optimal bandwidth.

For further insight, consider the case for a linear local estimator ($p = 1$) in the standard RD design ($\eta = 0$) with a constant group-level shock, c_g, and Ω takes the familiar "random effects" structure:

$$\Omega_g = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \cdots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \cdots & \sigma_c^2 + \sigma_u^2 \end{pmatrix}$$

Under this setup, Equation (6) can be written as follows:

$$h_{opt} = \frac{C_{2,0}}{4C_{1,0}}^{\frac{1}{5}} \left[\frac{(\sigma_{u,+}^2 + N_g \sigma_{c,+}^2) + (\sigma_{u,-}^2 + N_g \sigma_{c,-}^2)}{f(0) \left(\mu_+^{(2)} - \mu_-^{(2)}\right)^2} \right]^{\frac{1}{5}} N^{-1/5} \tag{7}$$

This rewrite makes clear that the key components driving differences in the cluster-robust and traditional procedures are cluster size and within-cluster dependence. As cluster size or within-cluster dependence increase, the current procedure produces bandwidths that differ from a bandwidth selection algorithm that assumes i.i.d. data. Intuitively, if there is strong within-cluster dependence each observation provides relatively less information to the researcher than if the observations were independent. This reflects the fact that when using the traditional bandwidth choice algorithm in the presence of clustering, the researcher is minimizing a restricted (incomplete) MSE and the resulting bandwidth does not correctly assess the trade-off between bias and variance.

3.2.2 Feasible Optimal Bandwidth Choice

A natural feasible bandwidth selector based on the optimal bandwidth described in Lemma 3.2 replaces all unknown parameters by estimates obtained from the data:[14]

$$\hat{h}_{opt} = \frac{C_2}{4C_1}^{\frac{1}{5}} \left[\frac{\frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \hat{\sigma}_{is+}}{\hat{f}_x(0) N N_g} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \hat{\sigma}_{is-}}{\hat{f}_x(0) N N_g}}{\left(\hat{\mu}_+^{(2)} - \hat{\mu}_-^{(2)}\right)^2} \right]^{\frac{1}{5}} \tag{8}$$

The denominator in Equation (8) could be close to zero in finite samples due to the lack of curvature in the regression using the polynomial of order $(p + 1)$ fitted to the data. Even if

[14] Throughout this section we use as example the case of the local linear estimator ($p = 1$). The extension for general p is straightforward and follows from Equation (6).

the true value of the bias term is not zero, the precision with which we estimate the second derivatives $\mu_+^{(2)}$ and $\mu_-^{(2)}$ is likely to be low. Hence, following IK and Calonico, Cattaneo, and Titiunik (2014) we introduce a regularization term that accounts for this lack of precision in the estimation of $\beta_+^{(2)}$ and $\beta_-^{(2)}$:

$$3 \cdot \quad V(\hat{\beta}_+^{(2)}) + V(\hat{\beta}_-^{(2)})$$

IK propose an approximation for the variance of $\hat{\beta}_+^{(2)}$ and $\hat{\beta}_-^{(2)}$ based on the specific case of homoskedasticity and a uniform kernel, which would be incompatible with the clustered sampling analysis implemented in this paper. We propose to set $\hat{V}(\hat{\beta}_+^{(2)})$ and $\hat{V}(\hat{\beta}_+^{(2)})$ equal to the "clustered variances" for $\hat{\beta}_+^{(2)}$ and $\hat{\beta}_-^{(2)}$ in the local quadratic regression using a pilot bandwidth and the implementation of the optimal bandwidth selector uses $\hat{\mu}_+^{(2)}$, $\hat{\mu}_-^{(2)}$, $\hat{V}(\hat{\beta}_+^{(2)})$ and $\hat{V}(\hat{\beta}_+^{(2)})$ from an initial local quadratic regression around the cutoff using a pilot bandwidth. Therefore, the optimal bandwidth can be implemented by

$$\hat{h}_{opt} = \quad \frac{C_2}{4C_1} \quad ^{\frac{1}{5}} \left[\frac{\frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \hat{\sigma}_{is+}}{\hat{f}_x(0) N N_g} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \hat{\sigma}_{is-}}{\hat{f}_x(0) N N_g}}{\hat{\mu}_+^{(2)} - \hat{\mu}_-^{(2)} \quad ^2 + \hat{r}_+ + \hat{r}_-} \right]^{\frac{1}{5}} \tag{9}$$

Where $\hat{r}_+ = 3\hat{V}(\hat{\beta}_+^{(2)})$ and $\hat{r}_- = 3\hat{V}(\hat{\beta}_-^{(2)})$.

In summary, the proposed implementation of the plug-in estimator given in Equation (9) follows these steps:

1. Choose a pilot bandwidth using the Silverman Rule,
   $h_1 = 2.576 \cdot S_x N^{-\frac{1}{5}}$ for a triangular kernel, where S_x is the sample variance of the running variable.

2. Let $N_{h,+}$ and $N_{h,-}$ be the number of observations within a bandwidth h above and below the threshold, respectively. Estimate $f(0)$:
   $\hat{f}(0) = \frac{N_{h_1,-} + N_{h_1,+}}{2h_1 N}$

3. Estimate the variance term using the following estimator:

$$\hat{\sigma}_{c,+}^2 \equiv \frac{1}{N_{h_1,+}} \sum_{g|c \leq x_g < c+h_1} \sum_{i} \sum_{s} \hat{}_{ig} \hat{}_{sg}$$

   Where $\hat{}_{ig} = y_{ig} - \bar{y}$. Then, use the corresponding estimator on the other side of the cutoff.

4. Estimate the curvature $\mu_+^{(2)}$ and $\mu_-^{(2)}$ by a local quadratic fit using a second pilot bandwidth,

h_2.[15]

5. Obtain the estimated regularization terms, which are locally approximated by

$$\hat{r}_+ = \frac{2160 \cdot \hat{\sigma}^2_{c,+}}{h^4_{2,+} N_{h_2,+}}$$

where $\hat{\sigma}^2_{c,+}$ is defined above. As before, use the corresponding estimator on the other side of the cutoff.

6. Plug the estimated quantities into Equation (9), obtaining the estimated optimal bandwidth.

4 Simulations

To illustrate the practical importance of adequately accounting for clustering when performing RD designs, we present a simulation study based on two data generating processes (DGPs).[16] For clarity, the setup follows a random effects structure:

$$Y_{ig} = m(x) + c_g + u_{ig}$$

Here, $m(x)$ is mean function, c_g is group-level shock with variance σ^2_c, and u_{ig} is idiosyncratic error term with variance σ^2_u. Simulations are run for various values of within cluster dependence, $\rho \equiv \frac{\sigma^2_c}{\sigma^2_c + \sigma^2_u}$. Throughout this section estimation is performed using a local linear estimator, the preferred method in most applications. In the first design, let $m(x)$ take the following form, which mimics the data in Lee (2008):

$$m_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 \;\; if \;\; x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 \;\; if \;\; x \geq 0 \end{cases}$$

Both u and c are normally distributed, the variance of u is set to 0.1295^2 and the variance of c is adjusted to obtain the desired value of ρ.[17]

We present results that utilize both our cluster-robust bandwidth and the traditional IK bandwidth that assumes i.i.d. data. Additionally, we perform simulations with data aggregated to the running variable level using the traditional bandwidth choice. This *ad hoc* approach is

[15] We follow IK in choosing this bandwidth to be optimal for minimizing MSE.

[16] Results from three additional simulations are available in Appendix B.

[17] This DGP is identical to that found in IK and Calonico, Cattaneo, and Titiunik (2014), with the addition of data dependence as described above.

sometimes used by researchers facing clustering issues in RD designs.[18] By aggregating the data to the running variable level, the researcher collapses the dependence structure and sidesteps the cluster issues, but ignores within-cluster variation in the data.

Based on the results presented in Section 3, we expect accounting for clustering to become more important as cluster size or within-cluster dependence increases. In addition, note that our procedure requires the estimation of a more complex variance formula that includes off-diagonal terms in the variance-covariance matrix. Therefore, our cluster-robust procedure may perform worse in practice when there is no within-cluster dependence when compared to a procedure that truthfully assumes i.i.d. data.

The simulation results in Table 1 align with these predictions. As expected, higher levels of within-cluster dependence, ρ, lead to situations where the cluster-robust procedure dominates procedures using traditional bandwidth selection algorithms in terms of empirical MSE. Moreover, as the size of clusters increases the current procedure far outperforms traditional bandwidth choices using the microdata. For small cluster sizes, our procedure performs similarly to IK in the case where $\rho = 0$ and the data is in fact i.i.d. However, for large cluster sizes the cluster-robust procedure can perform poorly for $\rho = 0$, reflecting the added difficulty of estimating the variance terms in bandwidth selection. Nonetheless, improved performance by the cluster-robust procedure can be observed for relatively small values of ρ.

Figure 1 presents these results graphically. Each panel plots the empirical MSE of each procedure for different values of ρ, where panels are separated by cluster size and number of clusters. Note first that the procedure using aggregated data overlaps almost entirely with the cluster-robust procedure, as both procedures perform very similarly for this DGP. These plots also make clear that there is a divergence between the cluster-robust procedure and the traditional procedure as ρ increases. In particular, the first column shows that with 250 clusters the cluster-robust procedure performs very similar to both the traditional procedure for small values of ρ, and performs significantly better as ρ increases. With 1000 clusters, the cluster-robust procedure performs slightly worse than the traditional procedure for small values of ρ, but accounting for clustering becomes more important with larger dependence.

One concern with the cluster-robust procedure proposed is that it often yields larger bandwidths. Given the well known trade-off between bias and variance that is inherent in RD designs,[19] it is useful to consider a situation where local linear estimators will struggle with

[18]See, for example, Ahn and Vigdor (2014).

[19]As pointed out in Section 3.2.1, the traditional approach might misrepresent the bias-variance trade-off embedded in the MSE by imposing no within-cluster dependence on the data.

bias due to extreme curvature of the conditional mean function near the cutoff. Therefore, in the second design we use a DGP studied in Calonico, Cattaneo, and Titiunik (2014) where the mean function is altered so that typical estimators will be heavily biased:

$$m_2(x) = \begin{cases} 0.48 + 1.27x - 0.5 \cdot 7.18x^2 + 0.7 \cdot 20.21x^3 + 1.1 \cdot 21.54x^4 + 1.5 \cdot 7.33x^5 \;\; if \;\; x < 0 \\ 0.52 + 0.84x - 0.1 \cdot 3.00x^2 - 0.3 \cdot 7.99x^3 - 0.1 \cdot 9.01x^4 + 3.56x^5 \;\; if \;\; x \geq 0 \end{cases}$$

This provides a natural setting to check whether our new procedure is able to accommodate conditional mean functions with extreme local curvature around the cutoff.

Table 2 and Figure 2 present the results of this simulation. Here, we can see that the cluster-robust procedure in general performs as well or better than traditional bandwidth selection procedure. In addition, unlike the first DGP, the cluster-robust procedure consistently outperforms the *ad hoc* procedure using the aggregated data set. Therefore, this case provides one example of a situation where the cluster-robust procedure produces improvements in MSE relative to a procedure that aggregates the data to the running variable level. As before, accounting for clustering becomes more important as cluster size or ρ increase. These results provide evidence that our procedure produces improvements in MSE in situations with data dependence even when there is extreme curvature of the conditional mean function at the cutoff.

5 Application: LIHTC and Neighborhood Characteristics

We now demonstrate the usefulness of these new methods using an empirical application that examines the effect of low-income housing subsidies on housing development and neighborhood characteristics. In particular, we focus the effects of the LIHTC, a program that has provided funding for roughly one third of all new units in multifamily housing built in the U.S. over the past thirty years (Khadduri, Climaco, and Burnett, 2012).[20] We exploit a discontinuity in program eligibility rules designating whether a particular census tract becomes a Qualified Census Tract (QCT). As discussed in Hollar and Usowski (2007) and Baum-Snow and Marion (2009), projects located in QCTs are eligible for up to 30 percent larger tax credits than projects in tracts not labeled as QCTs. Importantly, this designation is based on the fraction of households whose income falls below 60 percent of Area Median Gross Income (AMGI).[21] If the majority of households in a census tract have household income less than 60 percent of AMGI, the tract

[20] See Hollar and Usowski (2007) or Freedman and McGavock (2015) for overviews of the LIHTC program.

[21] The QCT designation methodology has changed since the period studied in the current analysis, but this does not influence the results presented here.

becomes eligible to receive QCT status. Therefore, the percent of households below 60 percent AMGI forms our running variable and the cutoff is 50 percent. By comparing only individuals that lived in tracts with a similar percentage of households below 60 percent of AMGI, we exploit random variation in QCT designation near the cutoff to identify the impact of the tax credits on housing development and neighborhood outcomes.

We perform this application using restricted access individual-level data from Census 2000 long form microdata.[22] We restrict to census tracts in metropolitan areas, and exclude Alaska and Hawaii.[23] Table 3 displays descriptive statistics for this data set. The number of LIHTC units and projects variables refer to the number of these units in the census tract. Clearly, QCT tracts contain much more disadvantaged populations than non-QCT tracts, a fact that is obvious due to the construction of the QCT status. In addition, note that QCT tracts have much larger numbers of LIHTC units and projects than non-QCT tracts. However, these descriptive differences between QCT and non-QCT tracts are not necessarily caused by LIHTC development or QCT designation, motivating the use of an RD design.

Table 4 displays results of three estimation procedures applied to the data. All estimates represent the results of local linear regressions using a triangular kernel, with standard errors that are robust to clustering at the tract level.[24] The first column presents the results of our bandwidth selection procedure applied to the microdata. Next, the second column presents results using the traditional IK bandwidth selection algorithm that does not account for clustering at the tract level. Finally, the last column presents results from applying this same procedure to data that has been aggregated to the tract level. These estimates are intended to replicate what a researcher would do when only aggregate data is available and the clustering issue is sidestepped.

The results show that accounting for potential dependence in outcomes within a census tract can substantially change the benchmark minimum-MSE bandwidth. As argued in Sections 3.2.1 and 4, the cluster-robust optimal bandwidth should be similar to the usual IK bandwidth in the absence of data dependence. The sizable differences between the bandwidth values obtained suggests that the usual algorithms potentially misrepresent the MSE bias/variance trade-off by failing to capture the dependence in the data.

In terms of the point estimates, the results show little evidence of a discontinuity in neigh-

[22] Since QCT classification and eligibility to extra tax credits was based on 1990 census tracts, location in 2000 is converted to tract location in 1990 using U.S. Census Bureau tract relationship files available at `https://www.census.gov/geo/maps-data/data/relationship.html`.

[23] These restrictions are similar to previous work by Baum-Snow and Marion (2009).

[24] Note that both procedures using the microdata perform inference with the same "cluster-robust" standard error formulas. Tract-level regressions utilize heteroskedasticity-robust standard errors.

borhood characteristics at the QCT threshold.[25] However, there is clear evidence of jumps in the implementation of new LIHTC units and projects at the boundary, indicating that the QCT policy is indeed producing increases in LIHTC construction. This is one area where the cluster-robust procedure leads to different empirical results than the traditional IK bandwidth selection. In particular, the IK procedure on the microdata produces a small and statistically insignificant estimate of the effect of QCT status on the number of LIHTC projects in the tract, whereas both the aggregated data and the current procedure produce estimates that suggest that there is a strong, statistically significant positive effect of QCT status on the number of LIHTC projects in a tract, as intended by policymakers.

Turning to standard error estimates, we see that applying the cluster-robust bandwidth choice procedure to the microdata produces estimates that are more precise than those obtained using a traditional bandwidth selection algorithm. This result is unsurprising, as accounting for the clustering will typically lead to larger bandwidth choices. When comparing the cluster-robust and aggregated data procedures, there is no clear relationship between the magnitude of the standard error estimates. Again, this reinforces the idea that both the cluster-robust and the aggregated data procedure are different approaches of accounting for clustering. In fact, on the whole both the cluster-robust and the aggregated data procedures provide similar results, and give a different empirical perspective than simply applying the IK bandwidth selection algorithm to the microdata.

6 Conclusion

Even though many recent RD analyses perform inference using cluster-robust standard error estimates, the justification for these methods is typically *ad hoc*. Moreover, current bandwidth selection procedures do not account for potential dependence among observations, creating a conflict in the assumptions between the bandwidth selection algorithm and inference procedures in RD studies.

In this study, we derive the asymptotic properties of local polynomial estimators in RD designs with data clustered at the running variable level and demonstrate a procedure which extends the popular minimum-MSE bandwidth selection algorithm by Imbens and Kalyanaraman (2012) to these situations. This procedure can be applied in a number of common applications, such as those with treatment being assigned at a higher level than the unit of observation or dis-

[25]This analysis differs from Baum-Snow and Marion (2009) in that it considers levels of neighborhood characteristics in 2000 instead of changes in characteristics from 1990 to 2000. Therefore, the two analyses are not directly comparable.

crete running variables. Simulation results indicate that in some practically important settings failing to account for dependence among observations leads to non-trivial increases in MSE due to bandwidth choices that are too small. We also present a simple application that demonstrates the practical importance of the cluster-robust optimal bandwidth choice algorithm by analyzing the impact of LIHTCs on neighborhood characteristics.

References

Ahn, Thomas and Jacob Vigdor. 2014. "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina." NBER Working Paper No. 20511.

Almond, Douglas, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. 2010. "Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns." *Quarterly Journal of Economics* 125 (2):591–634.

Baum-Snow, Nathaniel and Justin Marion. 2009. "The Effects of Low Income Housing Tax Credit Developments on Neighborhoods." *Journal of Public Economics* 93 (5):654–666.

Bhattacharya, Debopam. 2005. "Asymptotic Inference from Multi-stage Samples." *Journal of Econometrics* 126 (1):145–171.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–2326.

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

Card, David E, David Lee, and Zhuan Pei. 2009. "Quasi-Experimental Identification and Estimation in the Regression Kink Design." Working Paper, Princeton University.

Chen, M. Keith and Jesse M. Shapiro. 2007. "Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach." *American Law and Economics Review* 9 (1):1–29.

Dobkin, Carlos and Fernando Ferreira. 2010. "Do School Entry Laws Affect Educational Attainment and Labor Market Outcomes?" *Economics of Education Review* 29 (1):40–54.

Elder, Todd E. 2010. "The Importance of Relative Standards in ADHD diagnoses: Evidence Based on Exact Birth Dates." *Journal of Health Economics* 29 (5):641–656.

Fan, Jianqing and Irene Gijbels. 1992. "Variable Bandwidth and Local Linear Regression Smoothers." *Annals of Statistics* 20 (4):1669–2195.

Freedman, Matthew and Tamara McGavock. 2015. "Low-Income Housing Development, Poverty Concentration, and Neighborhood Inequality." *Journal of Policy Analysis and Management* Forthcoming.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1):201–209.

Hansen, Christian B. 2007. "Asymptotic Properties of a Robust Covariance Estimator when T is Large." *Journal of Econometrics* 141 (2):597–620.

Hollar, Michael and Kurt Usowski. 2007. "Low-Income Housing Tax Credit Qualified Census Tracts." *Cityscape* 9 (3):153–159.

Imbens, Guido W. and Karthik Kalyanaraman. 2012. "Optimal Bandwitdth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3):933–959.

Imbens, Guido W and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2):615–635.

Khadduri, Jill, Carissa Climaco, and Kimberly Burnett. 2012. "What Happens to Low-Income Housing Tax Credit Properties at Year 15 and Beyond?" U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2):675–697.

Lee, David S. and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142 (2):655–674.

Ludwig, Jens and Douglas 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.

Porter, Jack. 2003. "Estimation in the Regression Discontinuity Model." Unpublished Manuscript, Department of Economics, University of Wisconsin – Madison:5–19.

Schmieder, Johannes F., Till Von Wachter, and Stefan Bender. 2012. "The Effects of Extended Unemployment Insurance over the Business Cycle: Evidence from Regression Discontinuity Estimates over 20 Years." *Quarterly Journal of Economics* 127 (2):701–752.

Scott-Clayton, Judith. 2011. "On Money and Motivation: A Quasi-Experimental Analysis of Financial Incentives for College Achievement." *Journal of Human Resources* 46 (3):614–646.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, M.A.: MIT Press, 2$^{\text{nd}}$ ed.

Table 1: Simulation Results – DGP 1

		ρ				
		0	0.2	0.4	0.6	0.8
250 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0022	0.0030	0.0040	0.0068	0.0138
	Traditional Bandwidth MSE	0.0021	0.0030	0.0045	0.0082	0.0189
	Ratio	1.0088	0.9763	0.8757	0.8288	0.7311
Size=25	Cluster-Robust Bandwidth MSE	0.0016	0.0025	0.0038	0.0061	0.0134
	Traditional Bandwidth MSE	0.0013	0.0025	0.0051	0.0097	0.0255
	Ratio	1.1872	0.9854	0.7461	0.6242	0.5242
Size=200	Cluster-Robust Bandwidth MSE	0.0009	0.0024	0.0037	0.0063	0.0138
	Traditional Bandwidth MSE	0.0002	0.0029	0.0087	0.1255	0.0637
	Ratio	3.5405	0.8064	0.4266	0.0502	0.2164
500 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0018	0.0022	0.0030	0.0040	0.0074
	Traditional Bandwidth MSE	0.0018	0.0021	0.0031	0.0046	0.0095
	Ratio	1.0206	1.0327	0.9518	0.8777	0.7826
Size=25	Cluster-Robust Bandwidth MSE	0.0011	0.0020	0.0026	0.0035	0.0075
	Traditional Bandwidth MSE	0.0008	0.0016	0.0028	0.0051	0.0134
	Ratio	1.3538	1.2289	0.9080	0.6834	0.5567
Size=200	Cluster-Robust Bandwidth MSE	0.0004	0.0019	0.0026	0.0037	0.0073
	Traditional Bandwidth MSE	0.0001	0.0017	0.0039	0.0085	0.0233
	Ratio	3.5180	1.1666	0.6550	0.4316	0.3129
1000 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0014	0.0018	0.0022	0.0027	0.0046
	Traditional Bandwidth MSE	0.0014	0.0016	0.0021	0.0028	0.0055
	Ratio	1.0445	1.1304	1.0598	0.9699	0.8295
Size=25	Cluster-Robust Bandwidth MSE	0.0005	0.0016	0.0021	0.0028	0.0042
	Traditional Bandwidth MSE	0.0004	0.0010	0.0018	0.0030	0.0069
	Ratio	1.4220	1.5429	1.1647	0.9290	0.6134
Size=200	Cluster-Robust Bandwidth MSE	0.0001	0.0015	0.0020	0.0027	0.0043
	Traditional Bandwidth MSE	0.0000	0.0009	0.0021	0.0046	0.0112
	Ratio	2.9571	1.5970	0.9376	0.5821	0.3841

Numbers in cells refer to MSE from a particular procedure. Ratio refers to MSE from cluster-robust procedure divided by MSE from traditional procedure.

Table 2: Simulation Results – DGP 2

		ρ				
		0	0.2	0.4	0.6	0.8
250 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0030	0.0050	0.0077	0.0125	0.0269
	Traditional Bandwidth MSE	0.0019	0.0038	0.0064	0.0108	0.0244
	Ratio	1.5324	1.3337	1.2047	1.1578	1.1034
Size=25	Cluster-Robust Bandwidth MSE	0.0009	0.0029	0.0056	0.0106	0.0257
	Traditional Bandwidth MSE	0.0005	0.0028	0.0063	0.0122	0.0296
	Ratio	1.7908	1.0489	0.8915	0.8682	0.8692
Size=200	Cluster-Robust Bandwidth MSE	0.0002	0.0022	0.0053	0.0111	0.0240
	Traditional Bandwidth MSE	0.0001	0.0038	0.0111	0.0250	0.0702
	Ratio	2.3749	0.5795	0.4795	0.4429	0.3424
500 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0013	0.0022	0.0036	0.0064	0.0135
	Traditional Bandwidth MSE	0.0010	0.0019	0.0033	0.0060	0.0128
	Ratio	1.2848	1.1674	1.0963	1.0582	1.0577
Size=25	Cluster-Robust Bandwidth MSE	0.0004	0.0014	0.0028	0.0053	0.0134
	Traditional Bandwidth MSE	0.0002	0.0015	0.0034	0.0065	0.0158
	Ratio	1.4689	0.9332	0.8504	0.8134	0.8464
Size=200	Cluster-Robust Bandwidth MSE	0.0001	0.0011	0.0026	0.0054	0.0122
	Traditional Bandwidth MSE	0.0000	0.0020	0.0052	0.0102	0.0262
	Ratio	1.9594	0.5765	0.5014	0.5291	0.4671
1000 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0006	0.0011	0.0018	0.0031	0.0072
	Traditional Bandwidth MSE	0.0005	0.0010	0.0018	0.0031	0.0071
	Ratio	1.1352	1.0615	1.0135	1.0096	1.0150
Size=25	Cluster-Robust Bandwidth MSE	0.0002	0.0007	0.0015	0.0027	0.0064
	Traditional Bandwidth MSE	0.0001	0.0008	0.0019	0.0037	0.0085
	Ratio	1.2308	0.8559	0.7856	0.7411	0.7574
Size=200	Cluster-Robust Bandwidth MSE	0.0000	0.0006	0.0013	0.0028	0.0067
	Traditional Bandwidth MSE	0.0000	0.0011	0.0025	0.0056	0.0132
	Ratio	1.6494	0.5820	0.5170	0.4901	0.5064

Numbers in cells refer to MSE from a particular procedure. Ratio refers to MSE from cluster-robust procedure divided by MSE from traditional procedure.

Table 3: Descriptive Statistics

	QCT	Non-QCT
Homeownership	0.3316	0.6984
	(0.4708)	(0.4590)
Fraction Non-White	0.7565	0.2778
	(0.4292)	(0.4479)
High School Diploma or Higher	0.5744	0.8367
	(0.4944)	(0.3696)
Bachelors Degree or Higher	0.1110	0.2819
	(0.3142)	(0.4499)
Employment Population Ratio	0.4808	0.6363
	(0.4996)	(0.4811)
Number of LIHTC Projects	0.2714	0.1096
	(0.7147)	(0.5675)
Number of LIHTC Units	16.8094	8.8745
	(55.7813)	(43.7748)
Running Variable	0.1155	-0.2514
	(0.0913)	(0.1097)
N	3,063,042	27,879,680
N Clusters	6,778	37,938

Source: Microdata from the long form of the 2000 decennial census. Cells contains sample means. Standard deviations are in parentheses.

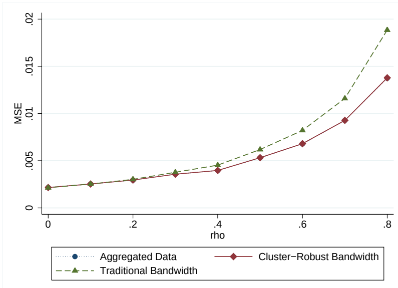
Table 4: Local Linear Estimates of the Effect of QCT Status

Dependent Variable	Cluster-Robust Bandwidth	Traditional Bandwidth	Tract-Level
Homeownership	-0.0054	-0.0098	-0.0044
	[0.0085]	[0.0145]	[0.0063]
	w=0.246	w=0.074	w=0.240
Fraction Non-White	0.0054	-0.0080	0.0051
	[0.0168]	[0.0374]	[0.0149]
	w=0.114	w=0.023	w=0.109
High School Diploma or Higher	-0.0075	-0.0001	-0.0102**
	[0.0074]	[0.0112]	[0.0049]
	w=0.142	w=0.061	w=0.197
Bachelors Degree or Higher	0.0055	0.0040	0.0021
	[0.0040]	[0.0055]	[0.0047]
	w=0.231	w=0.121	w=0.203
Employment Rate	0.0046	0.0065	-0.0024
	[0.0032]	[0.0052]	[0.0039]
	w=0.289	w=0.088	w=0.151
Number of LIHTC Units	7.279***	10.945**	4.949***
	[2.074]	[5.285]	[1.370]
	w=0.224	w=0.029	w=0.281
Number of LIHTC Projects	0.0731***	0.0297	0.0753***
	[0.0237]	[0.0578]	[0.0183]
	w=0.342	w=0.058	w=0.258
N	30,330,540	30,330,540	45,294
N Clusters	44,716	44,716	45,294

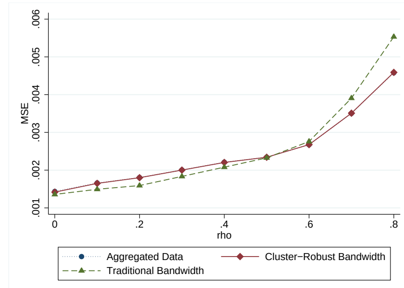
Source: Microdata and tract-level data from the long form of the 2000 decennial census. Standard errors in brackets are adjusted for clustering at the tract level. "w" refers to bandwidth, where tract-level regressions use the standard IK bandwidth. All estimates are from local linear regressions using a triangular kernel. ** indicates significance at the .05 level, *** indicates significance at the .01 level.

Figure 1: Simulation Results – Data Generating Process 1

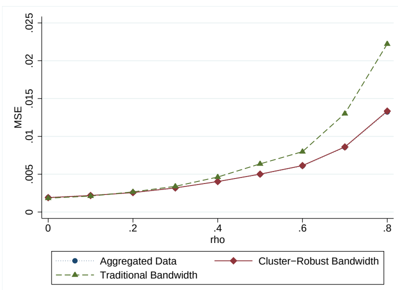
(a) Size = 5, Number of Clusters = 250



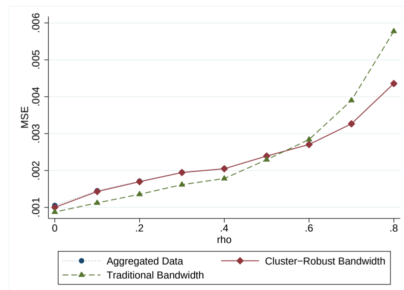
(b) Size = 5, Number of Clusters = 1000



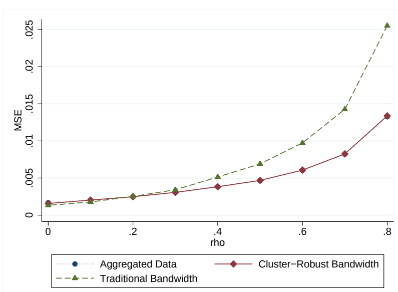
(c) Size = 10, Number of Clusters = 250



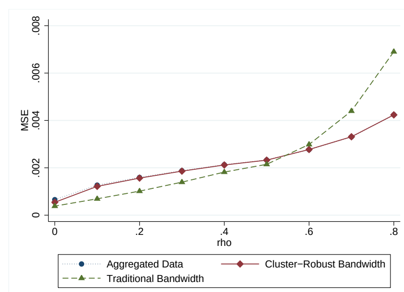
(d) Size = 10, Number of Clusters = 1000



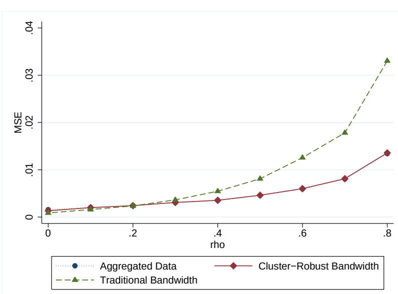
(e) Size = 25, Number of Clusters = 250



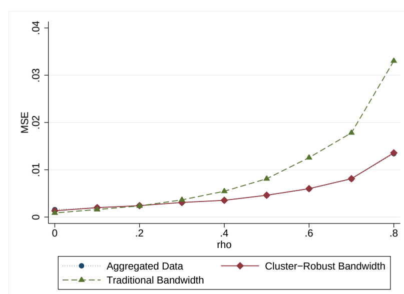
(f) Size = 25, Number of Clusters = 1000



(g) Size = 50, Number of Clusters = 250



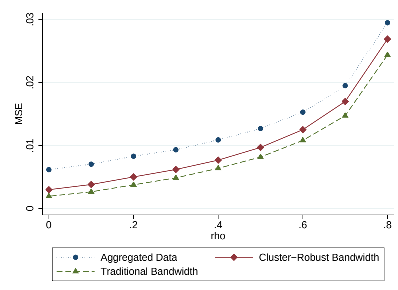
(h) Size = 50, Number of Clusters = 1000



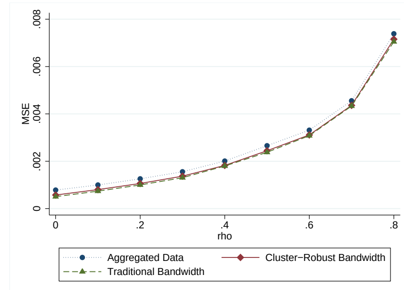
Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.

Figure 2: Simulation Results – Data Generating Process 2 (High Bias)

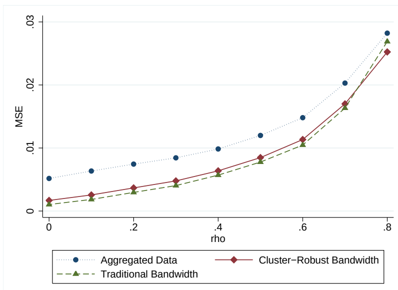
(a) Size = 5, Number of Clusters = 250



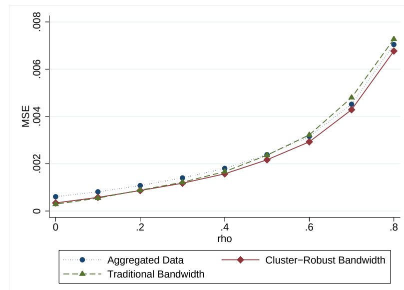
(b) Size = 5, Number of Clusters = 1000



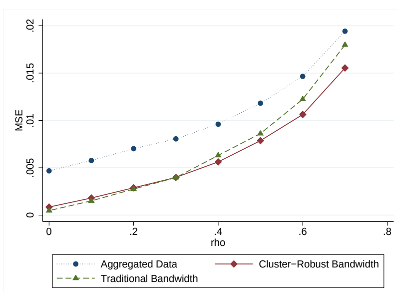
(c) Size = 10, Number of Clusters = 250



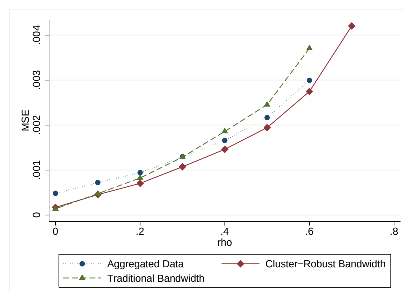
(d) Size = 10, Number of Clusters = 1000



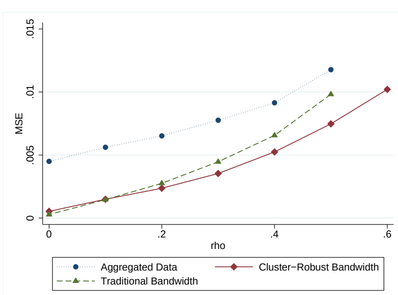
(e) Size = 25, Number of Clusters = 250



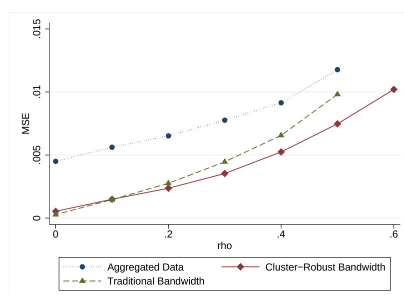
(f) Size = 25, Number of Clusters = 1000



(g) Size = 50, Number of Clusters = 250



(h) Size = 50, Number of Clusters = 1000



Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.

A Assumptions and Proofs

A.1 Assumptions

We use the following standard assumptions in the RD literature. For some $\kappa_0 > 0$, the following holds in the neighborhood $(-\kappa_0, \kappa_0)$ around the threshold $\bar{x} = 0$.

1. We have G independent and identically distributed clusters, with data (Y_g, X_g) , where Y_g and X_g are $1 \times N_g$ vectors for $g = 1, ..., G$ and for any given cluster $X_g = (x_g, x_g, ..., x_g)$.
2. $m(x) = E[Y|X]$ is at least $p + 2$ times continuously differentiable.
3. The density of the forcing variable X, denoted $f(X)$, is continuous and bounded away from zero.
4. The conditional variance $\Omega(x) = Var(Y|X) = I_G \otimes \Omega(x)$ is bounded and right and left continuous at $\bar{x}$. The right and left limit at the threshold exist and are positive definite.
5. The kernel $K(\cdot)$ is non-negative, bounded, differs from zero on a compact interval $[0, \kappa]$, and is continuous on $(0, \kappa)$ for some $\kappa > 0$.

A.2 Proofs

Lemma A.1. *Define* $F_j = \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{N_g} K_h(Z_{ig}) Z_{ig}^j = \frac{1}{G} \sum_{g=1}^{G} N_g \frac{1}{N_g} \sum_{i=1}^{N_g} K_h(Z_{ig}) Z_{ig}^j = \frac{1}{G} \sum_{g=1}^{G} N_g A_{jg}$*, where* $A_{jg} = \frac{1}{N_g} \sum_{i=1}^{N_g} K_h(Z_{ig}) Z_{ig}^j$*. If* N_g *is equal for all* G *clusters, then* $F_j = \frac{1}{G} \sum_{g=1}^{G} A_{jg}$*. Under Assumptions 1-5, (i) for non-negative integer* j

$$F_j = N_g h^j f(0) \nu_j + o_p(h^j) = N_g h^j (F_j^* + o_p(1))$$

with ν_j *defined in the main text and* $F_j^* \equiv f(0)\nu_j$ *and (ii) if* $j \geq 1$, $F_j = o_p(h^{j-1})$.

Proof. Focusing at A_{jg} for each cluster $g = 1, ..., G$:

$$\begin{aligned}
E\left[A_{jg}\right] &= E\left[\frac{1}{N_g} \sum_{i=1}^{N_g} K_h(Z_{ig}) Z_{ig}^j\right] = h^j \int_0^\infty K\left(x\right) x^j f(hx) dx \\
&= h^j \int_0^\infty K\left(x\right) x^j f(0) dx + h^{j+1} \int_0^\infty K\left(x\right) x^{j+1} \frac{f(hx) - f(0)}{hx} dx \\
&= h^j f(0) v_j + O(h^{j+1})
\end{aligned}$$

Then,

$$\begin{aligned}E[F_j] &= \frac{1}{G}\sum_{g=1}^{G} N_g E[A_{jg}] \\ &= N_g h^j f(0)v_j + O(h^{j+1})\end{aligned}$$

For the variance,

$$\begin{aligned}Var\left[A_{jg}\right] &= E\left(A_{jg}^2\right) - E\left[A_{jg}\right]^2 \\ &\leq \frac{1}{N_g^2}E\left[\sum_{i=1}^{N_g} K_h(Z_{ig})Z_{ig}^{2j}\right] = \frac{1}{N_g}E\left(K_h^2(Z_{ig})Z_{ig}^{2j}\right) \\ &= \frac{1}{N_g}h^{2j-1}\int_0^{\infty} K^2(x)x^{2j}f(xh)dx = O\left(\frac{h^{2j-1}}{N_g}\right) = O\left(h^{2j-1}\right)\end{aligned}$$

By noting that A_{jg} are independent across clusters.

$$\begin{aligned}Var\left[F_g\right] = Var\left(\frac{1}{G}\sum_{g=1}^{G} N_g A_{jg}\right) &= \frac{1}{G^2}\sum_{g=1}^{G} N_g^2 Var\left[A_{jg}\right] \\ = \frac{1}{G^2}\sum_{g=1}^{G} N_g^2 O\left(\frac{h^{2j-1}}{N_g}\right) &= \frac{1}{G^2}\sum_{g=1}^{G} O\left(h^{2j-1}\right) = O\left(\frac{h^{2j-1}}{G}\right) = o\left(\left(h^j\right)^2\right)\end{aligned}$$

Then,

$$\begin{aligned}F_j &= E[F_j] + O_p(Var(F_j)^{1/2}) \\ &= N_g h^j f(0)v_j + O(h^{j+1}) + o_p(h^j) \\ &= N_g h^j\left(f(0)v_j + o_p(1)\right)\end{aligned}$$

□

As discussed in the main text, we focus our attentions to the case in which cluster determination is based on the value of the running variable or, conversely, the running variable is defined at the group level, so $X_{ig} = X_g$. With this in mind we can show the following result.

Lemma A.2. *Define* $Q_{tj} = G^{-1}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K_h^2(z_g)z_g^{t+j}\sigma_{is}\left(z_g\right)$. *Then,*

$$Q_{tj} = h^{t+j-1}\left(f(0)\pi_{t+j}\frac{\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}\left(0\right)}{G} + o_p\left(1\right)\right)$$

If N_g is the same for all clusters and $\Omega_g = \Omega$ for all g,

$$Q_{tj} = h^{t+j-1}\left[f(0)\pi_{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(0) + o_p(1)\right]$$

with π^j defined in the text.

Proof.

$$\begin{aligned}
E\left[Q_{tj}\right] &= E\left[G^{-1}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}K_h^2(z_g)z_g^{t+j}\sigma_{is}(z_g)\right] \\
&= G^{-1}\sum_{g=1}^{G}\int_0^{\infty}\frac{1}{h^2}K^2\left(\frac{z}{h}\right)z^{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(z)f(z)dz \\
&= \int_0^{\infty}h^{t+j-1}K^2(x)x^{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(hx)f(hx)dx \\
&= h^{t+j-1}f(0)\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(0)\int_{-\infty}^{\infty}K^2(x)x^{t+j}dx + O(h^{t+j}) \\
&= h^{t+j-1}f(0)\pi_{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(0) + O(h^{t+j})
\end{aligned}$$

Now, to bound $Var(Q_{tj}|X)$:

$$Var\left[Q_{tj}\right] = E\left[(Q_{tj})^2\right] - E\left[Q_{tj}\right]^2$$

The first term:

$$\begin{aligned}
E\left[(Q_{tj})^2\right] &= E\left[\left(G^{-1}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}K_h^2(z_g)z_g^{t+j}\sigma_{is}(z_g)\right)^2\right] \\
&= G^{-2}\sum_{g=1}^{G}E\left[\left(K_h^2(z_g)z_g^{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(z_g)\right)^2\right] \\
&= G^{-2}\sum_{g=1}^{G}E\left[K_h^4(z_g)z_g^{2(t+j)}\left(\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}(z_g)\right)^2\right]
\end{aligned}$$

Note that all cross products will be of the type:

$$E\left[K_h^4(z_g)z_g^{2q}\sigma_{gis}\left(z_g\right)\sigma_{gtl}\left(z_g\right)\right]$$

Where q ranges from 0 to 2p, with p being the order of the polynomial used. Then,

$$\begin{aligned}
E\left[K_h^4(z_g)z_g^{2q}\sigma_{is}\left(z_g\right)\sigma_{tl}\left(z_g\right)\right] &= \int_0^\infty K_h^4(z)z^{2q}\sigma_{is}\left(z\right)\sigma_{tl}\left(z\right)f(z)dz \\
&= \int_0^\infty \frac{h^{2q}}{h^4}K^4(x)x^{2q}\sigma_{is}\left(hx\right)\sigma_{gtl}\left(hx\right)f(hx)hdx \\
&= h^{2q-3}\int_0^\infty K^4(x)x^{2q}\sigma_{is}\left(hx\right)\sigma_{tl}\left(hx\right)f(hx)dx \\
&= O\left(h^{2q-3}\right) = O\left(\left(\frac{h^{q-1}}{h^{\frac{1}{2}}}\right)^2\right) = o\left(\left(\frac{h^{q-1}}{h}\right)^2\right) = o\left(\left(h^{q-1}\right)^2\right)
\end{aligned}$$

Then,

$$\begin{aligned}
E\left[(Q_{tj})^2\right] &= N^{-2}\sum_{g=1}^{G}E\left[K_h^4(z_g)z_g^{2(t+j)}\left(\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{gis}\left(z_g\right)\right)^2\right] \\
&= G^{-2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sum_{t=1}^{N_g}\sum_{l=1}^{N_g}E\left[K_h^4(z_g)z_g^{2(t+j)}\sigma_{is}\left(z_g\right)\sigma_{tl}\left(z_g\right)\right] \\
&= G^{-2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sum_{t=1}^{N_g}\sum_{l=1}^{N_g}o\left(\left(h^{(t+j)-1}\right)^2\right) \\
&= G^{-1}N_g^4 o\left(\left(h^{(t+j)-1}\right)^2\right) \\
&= o\left(\left(G^{-\frac{1}{2}}N_g^2h^{t+j-1}\right)^2\right) = o\left(\left(G^{-\frac{1}{2}}h^{t+j-1}\right)^2\right) = o\left(\left(h^{t+j-1}\right)^2\right)
\end{aligned}$$

and

$$\begin{aligned}
Q_{tj} &= E[Q_{tj}] + O_p(Var(Q_j)^{\frac{1}{2}}) \\
&= h^{t+j-1}f(0)\pi_{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}\left(0\right) + o_p\left(h^{t+j-1}\right) \\
&= h^{t+j-1}\left[f(0)\pi_{t+j}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}\left(0\right) + o_p\left(1\right)\right]
\end{aligned}$$

□

With the results from the two lemmas above we can analyze the asymptotic distribution presented in 3.1 as well as the approximation to $MSE(h)$ in Lemma 3.2 and the subsequent optimal bandwidth formula.

Proof. Proof of Lemma 3.1 For analyzing the asymptotic approximation to the bias term, note that $y_{ig} = \mu(x_{ig}) + {}_{ig}$. Let $R = \begin{matrix} \iota & X & \cdots & X^p \end{matrix}$ with typical row given by $r_p(x) = \begin{matrix} 1 & x & \cdots & x^p \end{matrix}$ and e_η be a vector of zeros except for the $(\eta+1)^{th}$ entry equal to one, e.g., $e_0 = \begin{matrix} 1 & 0 & \cdots & 0 \end{matrix}$. Then,

$$\hat{\mu}_+^{(\eta)} = \eta! e_\eta (R\ WR)^{-1} R\ WY = \eta! e_\eta (R\ WR)^{-1} R\ W\left[\mu(X) + \ \right] \tag{10}$$

$$= \eta! e_\eta (R\ WR)^{-1} R\ W\mu(X) + \eta! e_\eta (R\ WR)^{-1} R\ W \tag{11}$$

We separate the analysis of the asymptotic properties of the estimator in three parts, the bias due to the potential local misspecification in the neighborhood of the cutoff, the estimator's variance, and its distribution which will be inherited from the second term in the equation above.

Bias

Let $E(\hat{\mu}(0)|X) = e_0(R\ WR)^{-1}R\ WM$, where M is defined below. Taking a Taylor expansion of $m(\cdot)$ around 0:

$$\mu(x_{ig}) = \mu(0) + \mu^{(1)}(0)x_{ig} + \frac{1}{2}\cdot\mu^{(2)}(0)x_{ig}^2 + \cdots + \frac{1}{(p+1)!}\cdot\mu^{(p+1)}(0)x_{ig}^{p+1} + T_{ig}$$

Where $|T_{ig}| \leq sup_x|\mu^{(p+2)}(x)x_{ig}^{p+2}|$.

Let $M = (\mu(x_{11}), \mu(x_{21}), ..., m(x_{12}), \mu(x_{22}), ..., \mu(x_{N_G G}))$. Then

$$M = R\begin{pmatrix} \mu(0) \\ \mu^{(1)}(0) \\ \vdots \\ \frac{\mu^{(p)}(0)}{p!} \end{pmatrix} + S + T$$

Where $S_{ig} = \frac{1}{(p+1)!}\mu^{(p+1)}(0)x_{ig}^{p+1}$.

Then,

$$Bias(\hat{\mu}^{(\eta)}) = \eta! e_\eta (R\ WR)^{-1} R\ WM - \mu^{(\eta)}(0) = \eta! e_\eta (R\ WR)^{-1} R\ W(S+T)$$

Note that,

$$
R\,WR = \begin{pmatrix}
\sum_{g=1}^{G}\sum_{i=1}^{N_g} K_h(x_{ig}) & \sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}K_h(x_{ig}) & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}^{p}K_h(x_{ig}) \\
\sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}K_h(x_{ig}) & \sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}^{2}K_h(x_{ig}) & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}^{p+1}K_h(x_{ig}) \\
\vdots & & \ddots & \\
\sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}^{p}K_h(x_{ig}) & \sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}^{p+1}K_h(x_{ig}) & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{N_g} x_{ig}^{2p}K_h(x)_{ig}
\end{pmatrix}
$$

Using the definition and results on Lemma A.1:

$$
\frac{1}{G}R\,WR = \begin{pmatrix}
F_0 & F_1 & \cdots & F_p \\
F_1 & F_2 & \cdots & F_{p+1} \\
\vdots & & \ddots & \\
F_p & F_{p+1} & \cdots & F_{2p}
\end{pmatrix}
$$

$$
= \begin{pmatrix}
F_0^* + o_p\left(1\right) & h\left[F_1^* + o_p\left(1\right)\right] & \cdots & h^p \;\; F_p^* + o_p\left(1\right) \\
h\left[F_1^* + o_p\left(1\right)\right] & h^2\left[F_2^* + o_p\left(1\right)\right] & \cdots & h^{p+1} \;\; F_{p+1}^* + o_p\left(1\right) \\
\vdots & & \ddots & \\
h^p \;\; F_p^* + o_p\left(1\right) & h^{p+1} \;\; F_{p+1}^* + o_p\left(1\right) & \cdots & h^{2p} \;\; F_{2p}^* + o_p\left(1\right)
\end{pmatrix}
$$

$$
= \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & h & \cdots & 0 \\
\vdots & & \ddots & \\
0 & 0 & \cdots & h^p
\end{pmatrix}
\begin{pmatrix}
F_0^* + o_p\left(1\right) & F_1^* + o_p\left(1\right) & \cdots & F_p^* + o_p\left(1\right) \\
F_1^* + o_p\left(1\right) & F_2^* + o_p\left(1\right) & \cdots & F_{p+1}^* + o_p\left(1\right) \\
\vdots & & \ddots & \\
F_p^* + o_p\left(1\right) & F_{p+1}^* + o_p\left(1\right) & \cdots & F_{2p}^* + o_p\left(1\right)
\end{pmatrix}
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & h & \cdots & 0 \\
\vdots & & \ddots & \\
0 & 0 & \cdots & h^p
\end{pmatrix}
$$

Recalling that $F_j^* \equiv N_g f(0)\nu_j$ and that $\frac{1}{f(0)}o_p(1) = o_p(1)$:

$$
= f(0)N_g \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & h & \cdots & 0 \\
\vdots & & \ddots & \\
0 & 0 & \cdots & h^p
\end{pmatrix}
\begin{pmatrix}
\nu_0 + o_p\left(1\right) & \nu_1 + o_p\left(1\right) & \cdots & \nu_p + o_p\left(1\right) \\
\nu_1 + o_p\left(1\right) & \nu_2 + o_p\left(1\right) & \cdots & \nu_{p+1} + o_p\left(1\right) \\
\vdots & & \ddots & \\
\nu_p + o_p\left(1\right) & \nu_{p+1} + o_p\left(1\right) & \cdots & \nu_{2p} + o_p\left(1\right)
\end{pmatrix}
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & h & \cdots & 0 \\
\vdots & & \ddots & \\
0 & 0 & \cdots & h^p
\end{pmatrix}
$$

Then,

$$\left(\frac{1}{G}R\,WR\right)^{-1} = \frac{1}{f(0)N_g}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & h^{-1} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & h^{-p} \end{pmatrix}\begin{pmatrix} \nu_0 + o_p(1) & \nu_1 + o_p(1) & \cdots & \nu_p + o_p(1) \\ \nu_1 + o_p(1) & \nu_2 + o_p(1) & \cdots & \nu_{p+1} + o_p(1) \\ \vdots & & \ddots & \\ \nu_p + o_p(1) & \nu_{p+1} + o_p(1) & \cdots & \nu_{2p} + o_p(1) \end{pmatrix}^{-1}.$$

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & h^{-1} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & h^{-p} \end{pmatrix}$$

Each term of the matrix in the middle above will be a a combination of products of the terms ν_j plus an $o_p(1)$ term.

$$\left(\frac{1}{G}R\,WR\right)^{-1}_{ij} = \frac{1}{h^{i+j-2}N_g f(0)}\left[\gamma_{ij} + o_p(1)\right] = O_p\left(\frac{1}{h^{i+j-2}}\right)$$

Where γ_{ij} is a deterministic function of ν known and computable for a given kernel and polynomial order. Examining $|\frac{1}{N}R\,WT|$:

$$\begin{aligned} \left|\frac{1}{G}R\,WT\right| &\leq \frac{1}{G}R\,|W|\begin{pmatrix} sup_x|\mu^{(p+2)}(x)||x_{11}^{p+2}| \\ \vdots \\ sup_x|\mu^{(p+2)}(x)||x_{N_G G}^{p+2}| \end{pmatrix} \\ &= sup_x|\mu^{(p+2)}(x)|\begin{pmatrix} \frac{1}{G}\sum_{i=1}^{N} K(x_i)x_i^{p+2} \\ \vdots \\ \frac{1}{G}\sum_{i=1}^{N} K(x_i)x_i^{2(p+1)} \end{pmatrix} \\ &= sup_x|\mu^{(p+2)}(x)|\begin{pmatrix} F_{p+2} \\ \vdots \\ F_{2(p+1)} \end{pmatrix} \leq \begin{pmatrix} o_p(h^{p+1}) \\ \vdots \\ o_p(h^{2p+1}) \end{pmatrix} \end{aligned}$$

Combining the results above, we obtain

$$e_\eta(R\,WR)^{-1}R\,WT = o_p(h^{p+1-\eta}).$$

For the first term, $\frac{1}{G}R\,WS$,

$$
\frac{1}{G} R\, WS = \frac{1}{(p+1)!}\mu^{(p+1)}(0)\begin{pmatrix} \frac{1}{G} \quad {}_{i=1}^{N} K_h(X_i)X_i^{p+1} \\ \vdots \\ \frac{1}{G} \quad {}_{i=1}^{N} K_h(X_i)X_i^{2p+1} \end{pmatrix}
$$

$$
= \frac{1}{(p+1)!}\mu^{(p+1)}(0)\begin{pmatrix} F_{p+1} \\ \vdots \\ F_{2p+1} \end{pmatrix} = \frac{1}{(p+1)!}\mu^{(p+1)}(0)N_g f(0)\begin{pmatrix} \nu_{p+1}h^{p+1} + o_p(h^{p+1}) \\ \vdots \\ \nu_{2p+1}h^{2p+1} + o_p(h^{2p+1}) \end{pmatrix}
$$

Let Γ^{-1} be a $(p+1)\times(p+1)$ matrix with typical element γ_{ij}. Then,

$$
\begin{aligned}
e_\eta(R\, WR)^{-1}R\, WS + o_p(h^{p+1-\eta}) =& \frac{1}{(p+1)!}\mu^{(p+1)}(0)N_g f(0) \\
& \quad \frac{1}{h^\eta N_g f(0)} \quad \gamma_{(\eta+1)1} + o_p(1) \quad \cdots \quad \frac{1}{h^{\eta+p} N_g f(0)} \quad \gamma_{(\eta+1)(p+1)} + o_p(1) \quad \cdot \\
& \begin{pmatrix} \nu_{p+1}h^{p+1} + o_p(h^{p+1}) \\ \vdots \\ \nu_{2p+1}h^{2p+1} + o_p(h^{2p+1}) \end{pmatrix} + o_p(h^{p+1-\eta}) \\
=& \frac{1}{(p+1)!}\mu^{(p+1)}(0) \\
& \quad \frac{1}{h^\eta} \quad \gamma_{(\eta+1)1} + o_p(1) \quad \frac{1}{h^{\eta+1}} \quad \gamma_{(\eta+1)2} + o_p(1) \quad \cdots \quad \frac{1}{h^{\eta+p}} \quad \gamma_{(\eta+1)(p+1)} + o_p(1) \\
& \begin{pmatrix} \nu_{p+1}h^{p+1} + o_p(h^{p+1}) \\ \vdots \\ \nu_{2p+1}h^{2p+1} + o_p(h^{2p+1}) \end{pmatrix} + o_p(h^{p+1-\eta}) \\
=& \frac{h^{p+1-\eta}}{(p+1)!}\mu^{(p+1)}(0) \quad \gamma_{(\eta+1)1} \quad \gamma_{(\eta+1)2} \quad \cdots \quad \gamma_{(\eta+1)(p+1)} \quad \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} \\
& + o_p(h^{p+1-\eta})
\end{aligned}
$$

Hence,

$$
E[\hat{\mu}_+^{(\eta)} - \mu_+^{(\eta)}|X] = \eta!\frac{h^{p+1-\eta}}{(p+1)!}\mu_+^{(p+1)} e_\eta \Gamma^{-1}\begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} + o_p(h^{p+1-\eta})
$$

And similarly to the estimates obtained below the threshold $E[\hat{\mu}_{-}^{(\eta)} - \mu_{-}^{(\eta)}|X]$.

Asymptotic Variance

For the variance component, note that the conditional variance can be written as follows:

$$V(\hat{\mu}^{(\eta)}(0)|X) = \eta!^2 e_\eta (R\, WR)^{-1} R\, W\Sigma W R (R\, WR)^{-1} e_\eta$$

Defining Σ as the block diagonal matrix with blocks given by Ω_g, the variance-covariance matrix for the error term in cluster g, for $g = 1, \cdots, G$ the middle term is given by:

$$R\, W\Sigma W R = \begin{pmatrix} \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})\sigma_{gis} & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})x_{sg}^{p}\sigma_{gis} \\ \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})x_{ig}\sigma_{gis} & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})x_{ig}x_{sg}^{p}\sigma_{gis} \\ \vdots & \ddots & \vdots \\ \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})x_{ig}^{p}\sigma_{gis} & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})x_{ig}^{p}x_{sg}^{p}\sigma_{gis} \end{pmatrix}$$

Where σ_{tj} is the term in the $i-th$ line and $j-th$ column in Ω_g.

$$\frac{1}{G} R\, W\Sigma W R = \begin{pmatrix} Q_{00} & Q_{01} & \cdots & Q_{0p} \\ Q_{10} & Q_{11} & \cdots & Q_{1p} \\ \vdots & & \ddots & \\ Q_{p0} & Q_{p1} & \cdots & Q_{pp} \end{pmatrix}$$

Where $Q_{tj} = G^{-1} \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{ig})K(x_{sg})x_{ig}^{t}x_{sg}^{j}\sigma_{gis}$.

Focusing on the case that X is defined at the cluster level and, hence, $x_{ig} = x_g \forall i = 1, ..., N_g$, substitute from the Lemma A.2, $Q_{tj} = h^{t+j-1} \; f(0)\pi_{t+j} \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}(0) + o_p(1)$

$$G^{-1} R\, W\Sigma W R = f(0) \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}(0) \begin{pmatrix} h^{-1}(\pi_0 + o_p(1)) & \pi_1 + o_p(1) & \cdots & h^{p-1}(\pi_p + o_p(1)) \\ \pi_1 + o_p(1) & h^{1}(\pi_2 + o_p(1)) & \cdots & h^{p}(\pi_{p+1} + o_p(1)) \\ \vdots & & \ddots & \\ h^{p-1}(\pi_p + o_p(1)) & h^{p}(\pi_{p+1} + o_p(1)) & \cdots & h^{2p-1}(\pi_{2p} + o_p(1)) \end{pmatrix}$$

$$= h^{-1} f(0) \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}(0)\, H \begin{pmatrix} \pi_0 + o_p(1) & \pi_1 + o_p(1) & \cdots & \pi_p + o_p(1) \\ \pi_1 + o_p(1) & \pi_2 + o_p(1) & \cdots & \pi_{p+1} + o_p(1) \\ \vdots & & \ddots & \\ \pi_p + o_p(1) & \pi_{p+1} + o_p(1) & \cdots & \pi_{2p} + o_p(1) \end{pmatrix} H$$

where, $H = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & h^1 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & h^p \end{pmatrix}$. Then,

$$G(R\,WR)^{-1}R\,W\Sigma WR(R\,WR)^{-1} =$$

$$= \frac{1}{hf(0)} \frac{\overset{N_g}{\underset{i=1}{}}\ \overset{N_g}{\underset{s=1}{}} \sigma_{is}\,(0)}{N_g^2} H^{-1} \begin{pmatrix} \nu_0 + o_p\,(1) & \nu_1 + o_p\,(1) & \cdots & \nu_p + o_p\,(1) \\ \nu_1 + o_p\,(1) & \nu_2 + o_p\,(1) & \cdots & \nu_{p+1} + o_p\,(1) \\ \vdots & & \ddots & \\ \nu_p + o_p\,(1) & \nu_{p+1} + o_p\,(1) & \cdots & \nu_{2p} + o_p\,(1) \end{pmatrix}^{-1} \cdot$$

$$\begin{pmatrix} \pi_0 + o_p\,(1) & \pi_1 + o_p\,(1) & \cdots & \pi_p + o_p\,(1) \\ \pi_1 + o_p\,(1) & \pi_2 + o_p\,(1) & \cdots & \pi_{p+1} + o_p\,(1) \\ \vdots & & \ddots & \\ \pi_p + o_p\,(1) & \pi_{p+1} + o_p\,(1) & \cdots & \pi_{2p} + o_p\,(1) \end{pmatrix} \begin{pmatrix} \nu_0 + o_p\,(1) & \nu_1 + o_p\,(1) & \cdots & \nu_p + o_p\,(1) \\ \nu_1 + o_p\,(1) & \nu_2 + o_p\,(1) & \cdots & \nu_{p+1} + o_p\,(1) \\ \vdots & & \ddots & \\ \nu_p + o_p\,(1) & \nu_{p+1} + o_p\,(1) & \cdots & \nu_{2p} + o_p\,(1) \end{pmatrix}^{-1} H^{-1}$$

$$= \frac{1}{hf(0)} \frac{\overset{N_g}{\underset{i=1}{}}\ \overset{N_g}{\underset{s=1}{}} \sigma_{is}\,(0)}{N_g^2} H^{-1}AH^{-1}$$

Note that each term in matrix A will be a combination of products of the terms ν_j and π_j plus an $o_p(1)$ term, hence

$$G(R\,WR)^{-1}R\,W\Sigma WR(R\,WR)^{-1}\ {}_{ij} = \frac{[a_{ij} + o_p(1)]}{h^{i+j-2}} \frac{1}{hf(0)} \frac{\overset{N_g}{\underset{t=1}{}}\ \overset{N_g}{\underset{s=1}{}} \sigma_{ts}\,(0)}{N_g^2}$$

Where a_{ij} is a deterministic function of ν and π known and computable for a given kernel and polynomial order.

$$G\ \ e_\eta(R\,WR)^{-1}R\,W\Sigma WR(R\,WR)^{-1}e_\eta\ \ = \frac{a_{(\eta+1)(\eta+1)} + o_p(1)}{h^{2\eta+1}f(0)} \frac{\overset{N_g}{\underset{i=1}{}}\ \overset{N_g}{\underset{s=1}{}} \sigma_{is}\,(0)}{N_g^2}$$

$$\begin{aligned} Var[\hat{\mu}_+^{(\eta)} - \mu_+^{(\eta)}|X] &= \eta!^2 \frac{1}{Gh^{2\eta+1}f(0)} \frac{\overset{N_g}{\underset{i=1}{}}\ \overset{N_g}{\underset{s=1}{}} \sigma_{is}^+}{N_g^2} e_\eta \Gamma^{-1}\Delta\Gamma^{-1}e_\eta + o_p\ \ \frac{1}{Gh^{2\eta+1}} \\ &= \eta!^2 \frac{1}{Nh^{2\eta+1}f(0)} \frac{\overset{N_g}{\underset{i=1}{}}\ \overset{N_g}{\underset{s=1}{}} \sigma_{is}^+}{N_g} e_\eta \Gamma^{-1}\Delta\Gamma^{-1}e_\eta + o_p\ \ \frac{1}{Gh^{2\eta+1}} \end{aligned}$$

Asymptotic Distribution

We have seen that:

$$\frac{\hat{\mu}_{+}^{(\eta)} - \mu_{+}^{(\eta)}}{\sqrt{Var[\hat{\mu}_{+}^{(\eta)} - \mu_{+}^{(\eta)}|X]}} = \frac{\hat{\mu}_{+}^{(\eta)} - E\left(\hat{\mu}_{+}^{(\eta)}|X\right) + E\left(\hat{\mu}_{+}^{(\eta)}|X\right) - \mu_{+}^{(\eta)}}{\sqrt{Var[\hat{\mu}_{+}^{(\eta)} - \mu_{+}^{(\eta)}|X]}} \tag{12}$$

$$= \varepsilon_1 + \varepsilon_2 = \varepsilon_1 + o_p(1) \tag{13}$$

Then,

$$\varepsilon_1 = \left(Var[\hat{\mu}_{+}^{(\eta)} - \mu_{+}^{(\eta)}|X]\right)^{-\frac{1}{2}} \left(\hat{\mu}_{+}^{(\eta)} - E\left(\hat{\mu}_{+}^{(\eta)}|X\right)\right) \tag{14}$$

$$= \left(Var[\hat{\mu}_{+}^{(\eta)} - \mu_{+}^{(\eta)}|X]\right)^{-\frac{1}{2}} \left(\frac{\eta! e_\eta (R WR)^{-1} R W}{G}\right) \tag{15}$$

and,

$$\varepsilon_2 = \frac{E\left(\hat{\mu}_{+}^{(\eta)}|X\right) - \mu_{+}^{(\eta)}}{\sqrt{Var[\hat{\mu}_{+}^{(\eta)} - \mu_{+}^{(\eta)}|X]}} \tag{16}$$

$$= O_p\left(\sqrt{Gh^{2\eta+1}}\right) O_p\left(h^{p+1-\eta}\right) = O_p\left(\sqrt{Gh^{3+2p}}\right) = o_p(1) \tag{17}$$

Note that,

$$R W = \sum_{g=1}^{G} R_g W_g \quad g = \sum_{g=1}^{G} K(x_g) R_g \quad g \tag{18}$$

$$= \sum_{g=1}^{G} K(x_g) \sum_{i=1}^{N_g} r_p(x_{ig}) \quad ig = \sum_{g=1}^{G} K(x_g) r_p(x_g) \sum_{i=1}^{N_g} \quad ig \tag{19}$$

and, $\varepsilon_1 = \tilde{\varepsilon}_1 + o_p(1)$, where

$$\tilde{\varepsilon}_1 = \sum_{g=1}^{G} \omega_g \quad g \tag{20}$$

$$\omega_g = \left(\frac{1}{GN_g h^{2\eta+1} f(0)} \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^{+}}{N_g} e_\eta \Gamma^{-1} \Delta \Gamma^{-1} e_\eta\right)^{-\frac{1}{2}} \frac{h^{-\eta} e_\eta \Gamma^{-1} K(x_g) r_p(x_g)}{G} \tag{21}$$

Since the vector of disturbances is independent across clusters and the clusters are randomly sampled we have that $E[\tilde{\varepsilon}_1] = 0$ and $V[\tilde{\varepsilon}_1] \to 1$. Hence, it will follow a central limit theorem converging to a $N(0,1)$. And similar results holds for $\hat{\mu}_{-}^{(\eta)}$.

□

Proof. Proof of Lemma 3.2

MSE(h):

$$
\begin{aligned}
E[(\hat{\tau}^{(\eta)} - \tau^{(\eta)})^2|X] &= E[(\hat{\mu}_+^{(\eta)} - \hat{\mu}_-^{(\eta)} - (\mu_+^{(\eta)} - \mu_-^{(\eta)}))^2|X] \\
&= Var[\hat{\mu}_+^{(\eta)} - \mu_+^{(\eta)}|X] + Var[\hat{\mu}_-^{(\eta)} - \mu_-^{(\eta)}|X] + \{E[\hat{\mu}_+^{(\eta)} - \mu_+^{(\eta)}|X] - E[\hat{\mu}_-^{(\eta)} - \mu_-^{(\eta)}|X]\}^2 \\
&= \eta!^2 \left(\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}^+}{GN_g^2 h^{2\eta+1} f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}^-}{GN_g^2 h^{2\eta+1} f(0)} \right) e_\eta \Gamma^{-1} \Delta \Gamma^{-1} e_\eta + o_p\left(\frac{1}{Gh^{2\eta+1}}\right) \\
&+ \left[\eta! \frac{h^{p+1-\eta}}{(p+1)!} \left(\mu_+^{(p+1)} - (-1)^{(p+1)} \mu_-^{(p+1)} \right) e_\eta \Gamma^{-1} \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} + o_p(h^{p+1-\eta}) \right]^2 \\
&= \eta!^2 \frac{1}{Gh^{2\eta+1}} \left(\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}^+}{N_g^2 f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}^-}{N_g^2 f(0)} \right) e_\eta \Gamma^{-1} \Delta \Gamma^{-1} e_\eta \\
&+ \eta!^2 \frac{h^{2(p+1-\eta)}}{(p+1)!^2} \left[\left(\mu_+^{(p+1)} - (-1)^{(p+1)} \mu_-^{(p+1)} \right) e_\eta \Gamma^{-1} \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} \right]^2 \\
&+ o_p\left(\frac{1}{Gh^{2\eta+1}} + h^{2(p+1-\eta)} \right) \\
&= \frac{1}{Gh^{2\eta+1}} C_{2\eta} \left(\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}^+}{N_g^2 f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{is}^-}{N_g^2 f(0)} \right) \\
&+ h^{2(p+1-\eta)} C_{1\eta} \left(\mu_+^{(p+1)} - (-1)^{(p+1)} \mu_-^{(p+1)} \right)^2 + o_p\left(\frac{1}{Gh^{2\eta+1}} + h^{2(p+1-\eta)} \right)
\end{aligned}
$$

Where $C_{1\eta} = \left[\frac{\eta!}{(p+1)!} e_\eta \Gamma^{-1} \begin{pmatrix} \nu_{p+1} \\ \vdots \\ \nu_{2p+1} \end{pmatrix} \right]^2$ and $C_{2\eta} = \eta!^2 e_\eta \Gamma^{-1} \Delta \Gamma^{-1} e_\eta$. The optimal bandwidth solves

$$
\begin{aligned}
h_{opt} &= arg\min \left(C_{2\eta} \left(\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{+}}{Gh^{2\eta+1}N_g^2 f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{-}}{Gh^{2\eta+1}N_g^2 f(0)} \right) + h^{2(p+1-\eta)} C_{1\eta} \left(\mu_{+}^{(p+1)} - (-1)^{(p+1)}\mu_{-}^{(p+1)} \right)^2 \right) \\
&= \left[\frac{C_{2\eta}(2\eta+1)}{2(p+1-\eta)C_{1\eta}} \frac{\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{+}}{GN_g^2 f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{-}}{GN_g^2 f(0)}}{\left(\mu_{+}^{(p+1)} - (-1)^{(p+1)}\mu_{-}^{(p+1)}\right)^2} \right]^{\frac{1}{2p+3}} \\
&= \left[C_{\kappa\eta} \frac{\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{+}}{GN_g^2 f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{-}}{GN_g^2 f(0)}}{\left(\mu_{+}^{(p+1)} - (-1)^{(p+1)}\mu_{-}^{(p+1)}\right)^2} \right]^{\frac{1}{2p+3}} \\
&= \left[C_{\kappa\eta} \frac{\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{+}}{NN_g f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{is}^{-}}{NN_g f(0)}}{\left(\mu_{+}^{(p+1)} - (-1)^{(p+1)}\mu_{-}^{(p+1)}\right)^2} \right]^{\frac{1}{2p+3}}
\end{aligned}
$$

where $C_{\kappa\eta} = \frac{(p+1)!^2(2\eta+1)e'_\eta\Gamma^{-1}\Delta\Gamma^{-1}e_\eta}{2(p+1-\eta)\left[e'_\eta\Gamma^{-1}\begin{pmatrix}\nu_{p+1}\\ \vdots \\ \nu_{2p+1}\end{pmatrix}\right]^2}$.

$\square$

B Supplemental Simulations

In addition, we consider two additional data generating processes derived from those studied in IK.[26] The cluster dependence setup is the same as before, but here we consider alternative conditional mean functions proposed in IK:

$$m_3(x) = \begin{cases} 3x^2 \;\; if \;\; x < 0 \\ 4x^2 \;\; if \;\; x \geq 0 \end{cases}$$

$$m_4(x) = \begin{cases} 0.42 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 \;\; if \;\; x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 \;\; if \;\; x \geq 0 \end{cases}$$

DGP 3 is of interest as the quadratic data generating process implies that the regularization term will be more important. In addition, DGP 4 shows a case similar to that in the first simulation, but with a constant average treatment effect.

Tables B.1-B.2 and Figures B.1-B.2 present results from the quadratic and constant average treatment effect data generating processes, respectively. All graphs show that the new procedure often performs better than the traditional IK bandwidth, particularly in settings where cluster size or ρ are large.

Last, Table B.3 and Figure B.3 presents simulation results from a linear DGP where the local linear model is correctly specified:

$$m_5(x) = \begin{cases} 0.48 + 1.27x \;\; if \;\; x < 0 \\ 0.52 + 0.84x \;\; if \;\; x \geq 0 \end{cases}$$

These results show that the cluster-robust procedure performs well in this setting as well.

[26]These simulations are simulation designs 2 and 3 in IK.

Table B.1: Simulation Results – Quadratic DGP

		ρ				
		0	0.2	0.4	0.6	0.8
250 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0021	0.0025	0.0036	0.0057	0.0131
	Traditional Bandwidth MSE	0.0022	0.0029	0.0046	0.0074	0.0189
	Ratio	0.9537	0.8637	0.7831	0.7631	0.6922
Size=25	Cluster-Robust Bandwidth MSE	0.0019	0.0023	0.0032	0.0053	0.0128
	Traditional Bandwidth MSE	0.0019	0.0030	0.0053	0.0099	0.0264
	Ratio	0.9932	0.7870	0.6026	0.5303	0.4853
Size=200	Cluster-Robust Bandwidth MSE	0.0018	0.0022	0.0033	0.0057	0.0126
	Traditional Bandwidth MSE	0.0017	0.0037	0.0088	0.0228	0.0644
	Ratio	1.0181	0.6078	0.3748	0.2499	0.1955
500 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0019	0.0022	0.0024	0.0034	0.0066
	Traditional Bandwidth MSE	0.0020	0.0024	0.0030	0.0046	0.0094
	Ratio	0.9864	0.9133	0.8261	0.7327	0.7031
Size=25	Cluster-Robust Bandwidth MSE	0.0018	0.0020	0.0024	0.0031	0.0066
	Traditional Bandwidth MSE	0.0018	0.0023	0.0036	0.0058	0.0142
	Ratio	0.9991	0.8653	0.6691	0.5280	0.4664
Size=200	Cluster-Robust Bandwidth MSE	0.0017	0.0020	0.0023	0.0032	0.0063
	Traditional Bandwidth MSE	0.0017	0.0027	0.0048	0.0096	0.0245
	Ratio	1.0091	0.7426	0.4933	0.3364	0.2563
1000 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0019	0.0020	0.0021	0.0024	0.0037
	Traditional Bandwidth MSE	0.0019	0.0021	0.0023	0.0030	0.0053
	Ratio	0.9962	0.9615	0.8880	0.7926	0.7024
Size=25	Cluster-Robust Bandwidth MSE	0.0018	0.0019	0.0020	0.0022	0.0035
	Traditional Bandwidth MSE	0.0018	0.0020	0.0025	0.0035	0.0074
	Ratio	1.0015	0.9559	0.7898	0.6374	0.4731
Size=200	Cluster-Robust Bandwidth MSE	0.0017	0.0019	0.0020	0.0024	0.0037
	Traditional Bandwidth MSE	0.0017	0.0022	0.0031	0.0054	0.0122
	Ratio	1.0052	0.8804	0.6543	0.4406	0.3068

Numbers in cells refer to MSE from a particular procedure. Ratio refers to MSE from cluster-robust procedure divided by MSE from traditional procedure.

Table B.2: Simulation Results – Constant Average Treatment Effect DGP

		ρ				
		0	0.2	0.4	0.6	0.8
250 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0034	0.0048	0.0072	0.0112	0.0224
	Traditional Bandwidth MSE	0.0041	0.0058	0.0092	0.0146	0.0299
	Ratio	0.8295	0.8231	0.7825	0.7635	0.7503
Size=25	Cluster-Robust Bandwidth MSE	0.0030	0.0046	0.0072	0.0107	0.0214
	Traditional Bandwidth MSE	0.0036	0.0065	0.0118	0.0204	0.0419
	Ratio	0.8274	0.7056	0.6157	0.5259	0.5091
Size=200	Cluster-Robust Bandwidth MSE	0.0031	0.0046	0.0071	0.0111	0.0222
	Traditional Bandwidth MSE	0.0035	0.0108	0.0346	0.0620	0.1980
	Ratio	0.8820	0.4280	0.2058	0.1793	0.1122
500 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0032	0.0038	0.0053	0.0071	0.0127
	Traditional Bandwidth MSE	0.0036	0.0044	0.0065	0.0092	0.0169
	Ratio	0.8963	0.8542	0.8241	0.7780	0.7544
Size=25	Cluster-Robust Bandwidth MSE	0.0031	0.0039	0.0048	0.0068	0.0128
	Traditional Bandwidth MSE	0.0035	0.0050	0.0074	0.0118	0.0236
	Ratio	0.9012	0.7837	0.6506	0.5730	0.5429
Size=200	Cluster-Robust Bandwidth MSE	0.0033	0.0041	0.0051	0.0066	0.0129
	Traditional Bandwidth MSE	0.0035	0.0064	0.0118	0.0203	0.0437
	Ratio	0.9287	0.6348	0.4267	0.3280	0.2951
1000 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0034	0.0036	0.0042	0.0050	0.0083
	Traditional Bandwidth MSE	0.0036	0.0040	0.0049	0.0062	0.0108
	Ratio	0.9434	0.8983	0.8556	0.8026	0.7691
Size=25	Cluster-Robust Bandwidth MSE	0.0033	0.0036	0.0042	0.0053	0.0080
	Traditional Bandwidth MSE	0.0034	0.0044	0.0058	0.0085	0.0151
	Ratio	0.9476	0.8200	0.7298	0.6170	0.5270
Size=200	Cluster-Robust Bandwidth MSE	0.0034	0.0036	0.0042	0.0052	0.0079
	Traditional Bandwidth MSE	0.0035	0.0049	0.0072	0.0120	0.0241
	Ratio	0.9563	0.7403	0.5763	0.4306	0.3278

Numbers in cells refer to MSE from a particular procedure. Ratio refers to MSE from cluster-robust procedure divided by MSE from traditional procedure.

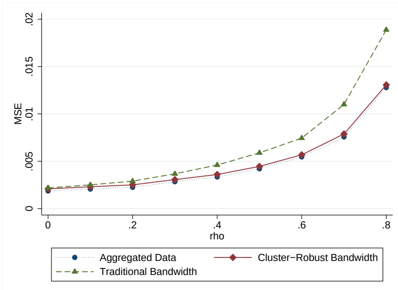
Table B.3: Simulation Results – Linear DGP

		ρ				
		0	0.2	0.4	0.6	0.8
250 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0005	0.0013	0.0025	0.0052	0.0113
	Traditional Bandwidth MSE	0.0006	0.0016	0.0032	0.0070	0.0164
	Ratio	0.8819	0.8283	0.7890	0.7410	0.6905
Size=25	Cluster-Robust Bandwidth MSE	0.0001	0.0009	0.0020	0.0043	0.0120
	Traditional Bandwidth MSE	0.0001	0.0013	0.0036	0.0085	0.0268
	Ratio	0.7369	0.6434	0.5460	0.5048	0.4464
Size=200	Cluster-Robust Bandwidth MSE	0.0000	0.0008	0.0020	0.0044	0.0113
	Traditional Bandwidth MSE	0.0000	0.0023	0.0068	0.0162	33.1257
	Ratio	0.5737	0.3397	0.2965	0.2714	0.0003
500 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0003	0.0006	0.0012	0.0024	0.0055
	Traditional Bandwidth MSE	0.0003	0.0008	0.0015	0.0032	0.0076
	Ratio	0.8821	0.8333	0.7850	0.7580	0.7169
Size=25	Cluster-Robust Bandwidth MSE	0.0001	0.0004	0.0010	0.0023	0.0055
	Traditional Bandwidth MSE	0.0001	0.0007	0.0019	0.0045	0.0123
	Ratio	0.7534	0.6076	0.5363	0.5181	0.4487
Size=200	Cluster-Robust Bandwidth MSE	0.0000	0.0004	0.0010	0.0022	0.0052
	Traditional Bandwidth MSE	0.0000	0.0012	0.0032	0.0081	0.0223
	Ratio	0.6053	0.3526	0.3033	0.2653	0.2337
1000 Clusters						
Size=5	Cluster-Robust Bandwidth MSE	0.0002	0.0003	0.0006	0.0012	0.0028
	Traditional Bandwidth MSE	0.0002	0.0004	0.0008	0.0016	0.0039
	Ratio	0.9102	0.8317	0.7866	0.7428	0.7169
Size=25	Cluster-Robust Bandwidth MSE	0.0000	0.0002	0.0005	0.0011	0.0028
	Traditional Bandwidth MSE	0.0000	0.0004	0.0010	0.0022	0.0063
	Ratio	0.7783	0.6181	0.5242	0.4893	0.4517
Size=200	Cluster-Robust Bandwidth MSE	0.0000	0.0002	0.0005	0.0011	0.0029
	Traditional Bandwidth MSE	0.0000	0.0006	0.0016	0.0040	0.0111
	Ratio	0.6284	0.3801	0.3229	0.2659	0.2605

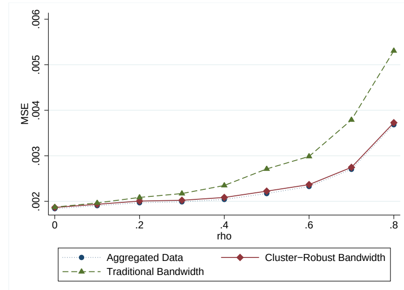
Numbers in cells refer to MSE from a particular procedure. Ratio refers to MSE from cluster-robust procedure divided by MSE from traditional procedure.

Figure B.1: Simulation Results – Quadratic DGP

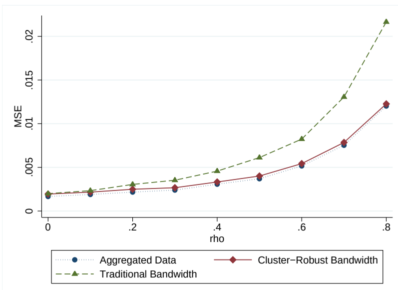
(a) Size = 5, Number of Clusters = 250



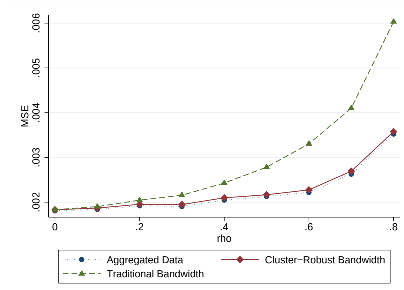
(b) Size = 5, Number of Clusters = 1000



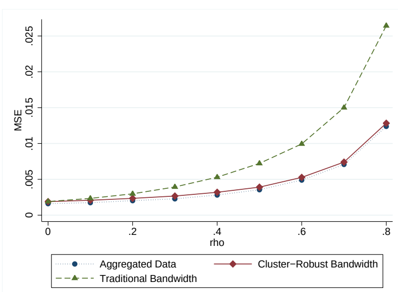
(c) Size = 10, Number of Clusters = 250



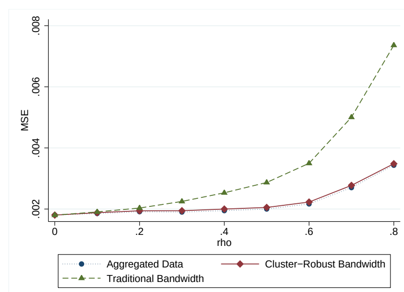
(d) Size = 10, Number of Clusters = 1000



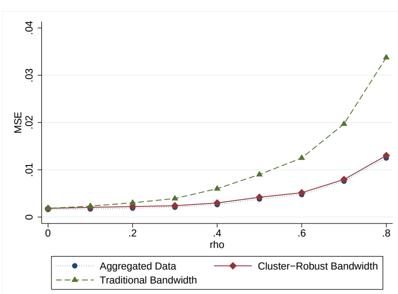
(e) Size = 25, Number of Clusters = 250



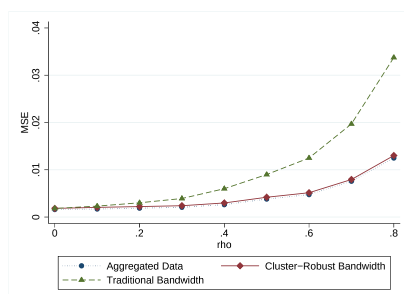
(f) Size = 25, Number of Clusters = 1000



(g) Size = 50, Number of Clusters = 250



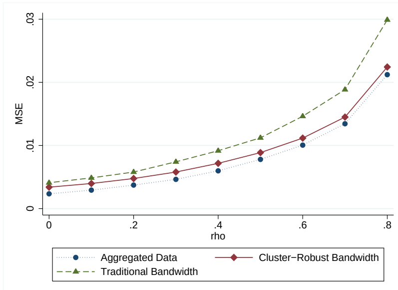
(h) Size = 50, Number of Clusters = 1000



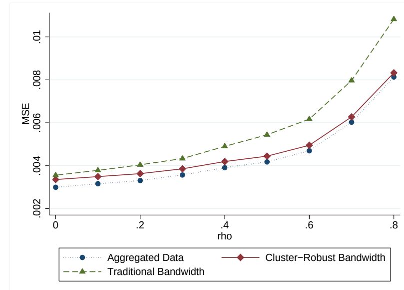
Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.

Figure B.2: Simulation Results – Constant Average Treatment Effect DGP

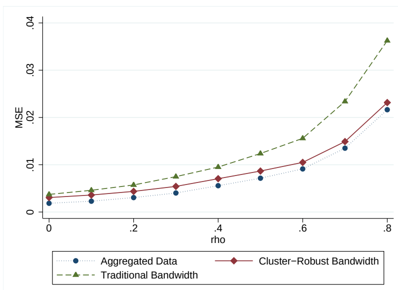
(a) Size = 5, Number of Clusters = 250



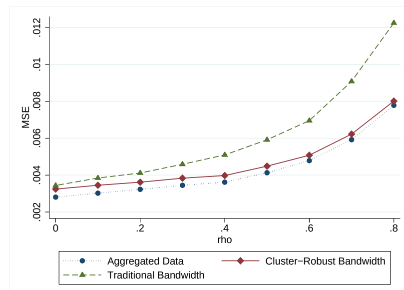
(b) Size = 5, Number of Clusters = 1000



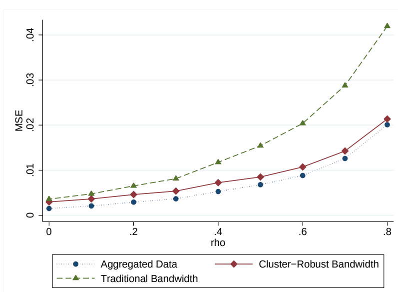
(c) Size = 10, Number of Clusters = 250



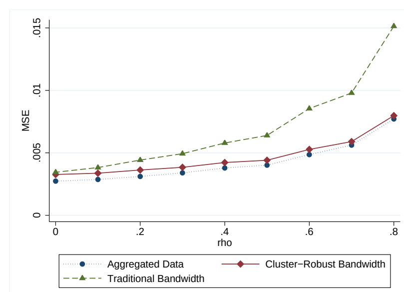
(d) Size = 10, Number of Clusters = 1000



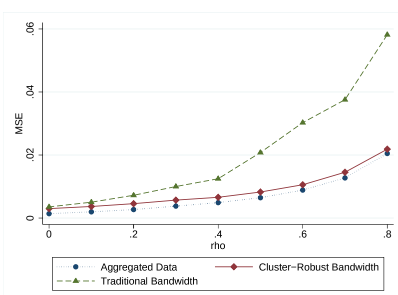
(e) Size = 25, Number of Clusters = 250



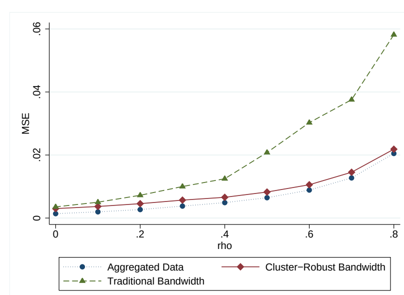
(f) Size = 25, Number of Clusters = 1000



(g) Size = 50, Number of Clusters = 250



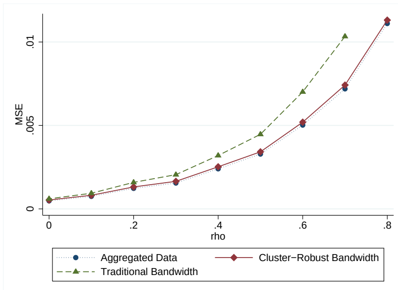
(h) Size = 50, Number of Clusters = 1000



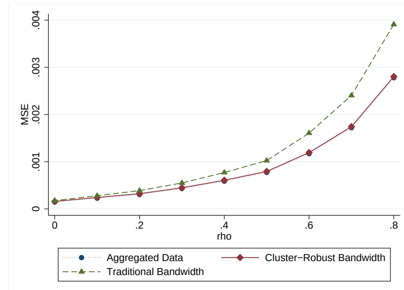
Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.

Figure B.3: Simulation Results – Linear DGP

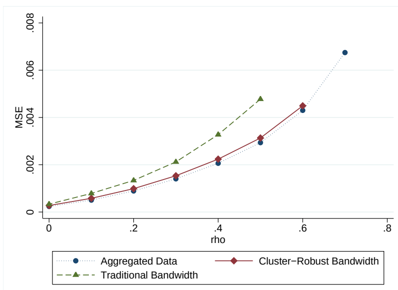
(a) Size = 5, Number of Clusters = 250



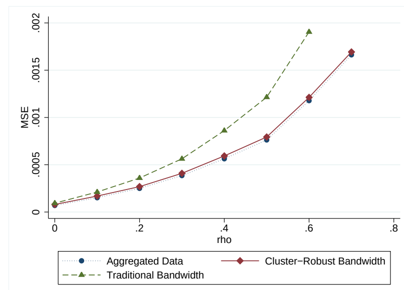
(b) Size = 5, Number of Clusters = 1000



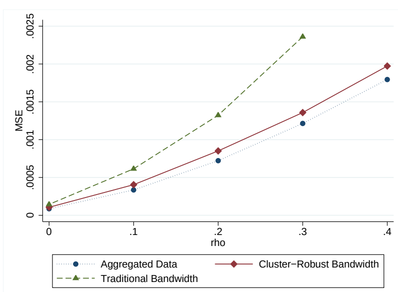
(c) Size = 10, Number of Clusters = 250



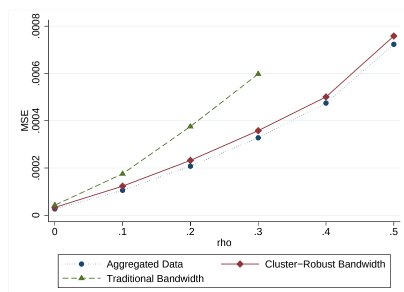
(d) Size = 10, Number of Clusters = 1000



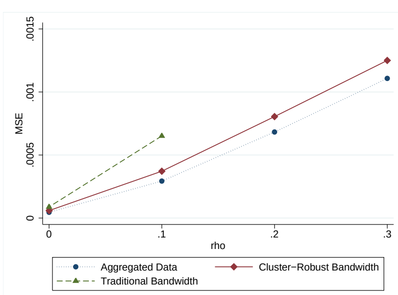
(e) Size = 25, Number of Clusters = 250



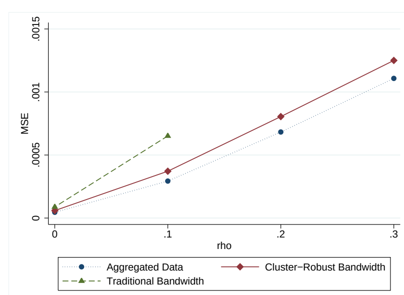
(f) Size = 25, Number of Clusters = 1000



(g) Size = 50, Number of Clusters = 250



(h) Size = 50, Number of Clusters = 1000



Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.