Regression Discontinuity Designs with Clustered Data: Variance and Bandwidth Choice

Otávio C. Bartalotti
Iowa State University, bartalot@iastate.edu

Quentin O. Brummet
United States Census Bureau, quentin.o.brummet@census.gov

Follow this and additional works at: http://lib.dr.iastate.edu/econ_workingpapers

Part of the Econometrics Commons

Recommended Citation
http://lib.dr.iastate.edu/econ_workingpapers/4

This Article is brought to you for free and open access by the Economics at Digital Repository @ Iowa State University. It has been accepted for inclusion in Economics Working Papers (2016-- ) by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
Regression Discontinuity Designs with Clustered Data: Variance and Bandwidth Choice *

Otávio Bartalotti and Quentin Brummet†

April 22, 2016

Abstract

Regression Discontinuity 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 unreasonable in many common applications. To relax this assumption, we derive the properties of traditional estimators in a setting that incorporates 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. may lead to higher MSE due to inadequate bandwidth choice. We apply our procedure to analyze the impact of Low-Income Housing Tax Credits 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 Matias Cattaneo, 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. We also are grateful to Matthew Freedman and Emily Owens for providing us with data and code to compute QCT eligibility for all census tracts. The views expressed within are those of the authors and not necessarily those of the U.S. Census Bureau. Any errors or omissions are our own.

†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, 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. For example, a researcher might be interested to use student-level microdata to examine a policy implemented based on a school-level running variable. A framework that considers clustering at the level of the running variable allows the researcher to select bandwidths, estimate parameters and perform tests, 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. This inference approach directly contradicts with commonly used bandwidth selection procedures such as those in Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titunik (2014) that assume i.i.d. data. 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.

In this study, we derive the asymptotic bias and variance 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 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 framework, 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 the Mean Squared Error (MSE) optimal bandwidth choice procedures proposed by Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titunik (2014) (henceforth, “IK” and “CCT,” respectively) 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 perform a simulation study demonstrating that a cluster-robust bandwidth choice procedure outperforms traditional bandwidth choices in terms of 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 outcomes in this application are observed at the 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 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 RD Designs

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

\[
Y_i = Y_i(0)1\{X_i < 0\} + Y_i(1)1\{X_i \geq 0\}
\]

where \( 1\{\cdot\} \) denotes the indicator function. For convenience, define

\[
\mu(x) = E[Y_i|X_i = x]
\]

In most cases the population parameter of interest is \( \tau = E[Y(1) - Y(0)|X = \bar{x}] \) (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,\(^1\) $\tau$ is nonparametrically identified (Hahn, Todd, and Van der Klaauw, 2001) by:

$$\tau = \mu_+ - \mu_-$$

where $\mu_+ = \lim_{x \to 0^+} \mu(x)$, and $\mu_- = \lim_{x \to 0^-} \mu(x)$ (3)

In general one might also be interested in the discontinuity of a higher order derivative of the conditional expectation at the threshold such as in the “regression kink” literature (Card, Lee, and Pei, 2009). Let $\mu^{(\eta)}(x) = \frac{d^\eta \mu(x)}{dx^\eta}$ be the $\eta^{th}$ 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 (Fan and Gijbels, 1992; Hahn, Todd, and Van der Klaauw, 2001; Porter, 2003; Calonico, Cattaneo, and Farrell, 2015).

For a local polynomial of order $p$, we use the following estimator:

$$
\hat{\tau}^{(\eta)} = \hat{\mu}_+^{(\eta)} - \hat{\mu}_-^{(\eta)}
$$

$$(\hat{\beta}_+, \hat{\beta}_+^{(1)}, \ldots, \hat{\beta}_+^{(p)})' = \text{argmin}_{b_0, b_1, \ldots, b_p} \sum_{i=1}^{N} \mathbb{I}\{X_i \geq 0\} (Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2 \cdot K_h(X_i)$$

$$(\hat{\beta}_-, \hat{\beta}_-^{(1)}, \ldots, \hat{\beta}_-^{(p)})' = \text{argmin}_{b_0, b_1, \ldots, b_p} \sum_{i=1}^{N} \mathbb{I}\{X_i < 0\} (Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2 \cdot K_h(X_i)$$

where the kernel function is given by $K_h(x_{ig}) = K \left( \frac{x_{ig}}{h} \right) \frac{1}{h}$, and $\hat{\mu}^{(\eta)} = \eta \hat{\beta}^{(\eta)}$.

### 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 (Wooldridge, 2010, p. 864). This sampling scheme is assumed to generate observations that are independent across clusters. Then, for a random sample of $G$ groups of

\(^1\)The assumptions used in the derivations and results presented here closely follow IK and are discussed in Appendix A.1.
fixed size $N_g$, we observe

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

(4)

Where the subscript $ig$ refers to unit $i$ in cluster $g$. The asymptotic approximations below assume 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 asymptotic variance and MSE-optimal choice of bandwidth in RD designs under clustering, letting $\text{Var}(Y|X) = I_G \otimes \Omega(x)$, where its elements, $\Omega_{ij}$, are denoted as $\sigma_{ij}(x)$, and its limits $\lim_{x \to 0^+} \sigma_{ij} = \sigma_{ij}^+$ and $\lim_{x \to 0^-} \sigma_{ij} = \sigma_{ij}^-$ throughout the paper.

3 Main Results

3.1 Asymptotic Approximations

Given this setup, we derive the asymptotic properties of $\hat{\tau}^{(n)}$. 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 $\Gamma$ and $\Delta$ as $(p+1) \times (p+1)$ matrices with element $(i, j)$ given by $\nu_{i+j-2}$ and $\pi_{i+j-2}$, 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 $Gh \to \infty$.

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

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

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

$$\text{Var}[\hat{\tau}^{(n)} - \tau^{(n)}|X] = \left[ h^2 \left( \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_i} \sigma_{is}^+}{GN^2_h^2 h^{2\eta+1} f(0)} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_i} \sigma_{is}^-}{GN^2_h^2 h^{2\eta+1} f(0)} \right) \epsilon' \Gamma^{-1} \Delta \Gamma^{-1} \epsilon + o_p(1) \right] \{1 + o_p(1)\}$$
3. (D) If $G h^{2p+3} \to 0$, then

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

This result is the analogue to the results for i.i.d data in CCT (Lemma A.1, p. 16), and provides insight into the impact of clustering in the asymptotic behavior of the estimators usually employed in RD designs. Note that the asymptotic variance formula now encompasses the clustered nature of the data and includes both within-cluster variance and covariance terms. The formula in Part (V) makes clear that ignoring dependence in the data misrepresents the variance of the estimators and could impact significantly the choice of bandwidth by the researcher when using a data-driven algorithm based on i.i.d. data approximations.

Furthermore, the asymptotic variance formula relates closely with the typical cluster-robust standard error formulas. As pointed out by CCT and Calonico, Cattaneo, and Farrell (2015) in the i.i.d. case, the commonly used “fixed-n” cluster-robust standard errors are valid. Hence, the results above extend that insight so that these estimators can be used in RD studies utilizing a non-parametric local polynomial estimator with clustering at the running variable level. This is also 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. See Section 3.1 of Calonico, Cattaneo, and Farrell (2015) for a careful treatment of this issue.

As a secondary point, note that the form of the leading term of the asymptotic bias in Part (B) is not directly affected by the presence of clustering as described here. As pointed out by IK and CCT, the leading bias term depends on the higher order derivatives of the conditional outcome above and below the cutoff and does not depend on the form of the asymptotic variance.

Part (D) presents the asymptotic validity of the test statistic based on the “conventional” distributional assumption, $G h^{2p+3} \to 0$. As discussed in CCT, this condition is explicitly imposed to eliminate the contribution of the leading bias term by ensuring that the bandwidth is “small” enough. Violations of this condition will have important impacts on the validity of resulting tests, and its consequences are discussed in detail below in Section 3.4.

### 3.2 Optimal Bandwidth Selection

In this section we derive the optimal-MSE bandwidth choice for RD designs with clustered data. 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 \( h^\ast \) that minimizes \( MSE(h) \):

\[
MSE(h) = E[(\hat{\tau} - \tau)^2 | X]
\]

(5)

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

1. **(MSE)**

\[
MSE(h) = \frac{1}{Gh^{2(\eta+1)}} C_{2,\eta} \left[ \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_s} \sigma^{+}_{is}}{N_g^2 f(0)} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_s} \sigma^{-}_{is}}{N_g^2 f(0)} \right]
\]

\[+ h^{2(p+1-\eta)} C_{1,\eta} \left[ \mu^{(p+1)} - (-1)^{(p+1+\eta)} \mu^{(p+1)} \right]^2 + o_p \left( \frac{1}{Gh^{2(\eta+1)} + h^{2(p+1-\eta)}} \right)\]

Where \( C_{1,\eta} = \frac{\eta!}{(p+1)!} e_{\eta}^{\prime} \Gamma^{-1} \) and \( C_{2,\eta} = \eta! e_{\eta}^{\prime} \Gamma^{-1} \Delta \Gamma^{-1} e_{\eta} \).

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

\[
h_{opt} = \left[ \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_s} \sigma^{+}_{is}}{N_g^2 f(0)} + \frac{\sum_{i=1}^{N_g} \sum_{s=1}^{N_s} \sigma^{-}_{is}}{N_g^2 f(0)} \right]^{\frac{1}{2(\eta+1)-p}}
\]

(6)

where \( C_{\eta} = \frac{2^{\eta+1}}{2(p+1-\eta)} C_{2,\eta}^{\frac{1}{C_{1,\eta}}} \).

The results presented in Lemma 3.2 follow very closely those presented for the local linear estimator by IK and the more general case developed by CCT (Lemma 1, p.11) for the i.i.d. data case. Comparing Equation (6) to the infeasible bandwidth choice in IK or CCT, 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 traditional bandwidth choice.

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 \( \Omega \) takes the familiar “random effects”
structure:

\[
\Omega_g = \begin{pmatrix}
\sigma^2_c + \sigma^2_u & \sigma^2_c & \cdots & \sigma^2_c \\
\sigma^2_c & \sigma^2_c + \sigma^2_u & \cdots & \sigma^2_c \\
\vdots & \vdots & \ddots & \vdots \\
\sigma^2_c & \sigma^2_c & \cdots & \sigma^2_c + \sigma^2_u
\end{pmatrix}
\]

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

\[
h_{\text{opt}} = \left( \frac{C_{2,0}}{4C_{1,0}} \right)^{1/5} \left[ \frac{(\sigma^2_{u,+} + N_g \sigma^2_{c,+}) + (\sigma^2_{u,-} + N_g \sigma^2_{c,-})}{f(0) \left[ \mu^{(2)}_+ - \mu^{(2)}_- \right]^2} \right]^{1/5} N^{-1/5}
\]

This rewrite makes clear that the key components driving differences in the cluster-robust and conventional optimal bandwidth formulas are cluster size and within-cluster dependence. As cluster size or within-cluster dependence increase, the current approach produces bandwidths that differ from the usual formulas 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 minimizes a restricted (incomplete) MSE and the resulting bandwidth does not correctly assess the trade-off between bias and variance.

3.3 Implementation

Similar to what was proposed by IK and CCT, a cluster-robust bandwidth can be implemented using a data-driven direct plug-in (DPI) bandwidth selector that replaces the unknown quantities in Equation 7 by their consistent estimates. The key difference in the cluster-robust case discussed above is the need for a variance estimator that captures the cluster dependence. In particular, one could use the following:

\[
\hat{\sigma}^2_{c,+} \equiv \frac{1}{N_{h_1,+}} \sum_{g \mid c \leq x_g < c+h_1} \left( \sum_i \sum_s \hat{\epsilon}_{ig} \hat{\epsilon}_{sg} \right)
\]

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

Those DPI bandwidth selectors require the direct estimation of \( f(0) \), cluster covariance terms directly. As discussed by CCT, this may be cumbersome and can be avoided by estimating the
whole denominator in Equation 7 at once. CCT provides an algorithm to obtain consistent optimal bandwidth estimators in that fashion, which can be modified easily to account of off-diagonal terms in the variance-covariance matrix. Note also that both the DPI approaches in IK and CCT incorporate a “regularization” term to avoid small denominators that could arise if the estimated differences in curvature at the cutoff are close to zero. This regularization term can be modified in a straightforward manner to account for cluster dependence in the data using the variance estimator shown above.

Thankfully, the cluster-robust variance estimator and several bandwidth choice algorithms have recently been incorporated into the rdrobust package in STATA and R, making it easily available to practitioners (Calonico et al., 2016).

3.4 Extensions

The main result of the analysis above is that cluster dependence should to be taken into account not only when performing inference but also selecting bandwidths. That insight, as well as the cluster-robust variance formulas above can be transferred directly to a variety of other results in the RD literature.

First, the cluster dependent variance results presented above extend naturally to fuzzy RD designs as the usual estimates can be rewritten as a linear combination of the discontinuities observed at the outcome and treatment status at the cutoff. As discussed above, the leading term of the bias is not affected by clustering in the setting we analyze. The asymptotic variance formulas for fuzzy RD follow the descriptions in IK Section 5.1 and CCT’s Lemma 2 with the simple variance terms being replaced by the cluster-robust formulas above. One would naturally want to allow the outcome and treatment status to be potentially correlated within cluster by allowing $\text{Cov}(Y,T|X) = I_G \otimes \Omega_{YT}(x)$ to have non-zero off diagonal elements within clusters.\footnote{Note that the fuzzy RD design implementation is available in STATA through the rdrobust package (Calonico et al., 2016).}

An important caveat to the results in Section 3.2 is that, as discussed by CCT, the MSE-optimal bandwidth proposed does not satisfy the “bias condition,” $Gh^{2p+3} \rightarrow 0$, and leads to bandwidths that are potentially too large, inducing a first-order bias in the distributional approximation. A common strategy to address this issue is to “undersmooth” by choosing a bandwidth that is smaller than the value suggested by the optimal-MSE algorithm, typically by shrinking the bandwidth by an ad-hoc quantity. Recently, Calonico, Cattaneo, and Farrell (2015) have shown formally that explicit bias-correction coupled with an analytical adjustment to the standard errors “yields confidence intervals with coverage that is accurate, or better, than
the best possible undersmoothing approach” (Calonico, Cattaneo, and Farrell, 2015, p. 2). It is straightforward to allow for cluster-robust bandwidth selection in the context of bias correction, as the lead bias term is not affected by the cluster dependence, as shown in Lemma 3.1, and the standard-error correction proposed in CCT follows similarly with the addition of off-diagonal variance terms that account for cluster dependence.\(^3\)

Instead of relying on optimal-MSE bandwidths, Calonico, Cattaneo, and Farrell (2015) propose bandwidth choice methods for RD designs that focus on delivering confidence intervals with optimal coverage error rates. While the authors note that the bandwidth formulas for optimal coverage rate results are “prohibitively cumbersome” (Calonico, Cattaneo, and Farrell, 2015, p. 29), they suggest a rule-of-thumb bandwidth choice based on an adjusted MSE-optimal bandwidth, which provides the correct coverage error rate. As it is the case with other DPI bandwidth choice approaches, a straightforward modification of these procedures to account for cluster dependence can be incorporated in these settings.

Last, in a more general theoretical 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. 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. 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, \(X\), guaranteeing that cluster dependence stays intact asymptotically. 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 want to implement cluster-robust methods for inference and bandwidth choice described in this paper.

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). All

\(^3\)Cluster-robust implementation of the explicit bias-correction in RD designs is available as part of the rdrobust package in STATA (Calonico et al., 2016).
simulation results use the implementation developed by Calonico et al. (2016) for STATA. 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 \( \sigma^2_c \), and \( u_{ig} \) is idiosyncratic error term with variance \( \sigma^2_u \). 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):

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

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 \( \rho \). Note that this DGP is identical to that found in IK and CCT, with the addition of data dependence as described above.

We present results that utilize both our cluster-robust bandwidth and the traditional 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 procedure averages all observations in a given cluster and performs estimation with the aggregated data, thereby treating each cluster as a single observation. This ad hoc approach is sometimes used by researchers such as Ahn and Vigdor (2014) when facing clustering issues in RD designs. By aggregating the data to the running variable level, the researcher collapses the dependence structure in the data, exploiting the fact that clusters are independent from each other. This sidesteps the usual 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.

---

\(^4\)We use the software’s version updated in April 2016, available for download at https://sites.google.com/site/rdpackages/rdrobust. Additional simulations are available in the online appendix at the author’s website: http://https://sites.google.com/site/quentinbrummet.
The simulation results in Table 1 align with these predictions. As expected, higher levels of within-cluster dependence, \( \rho \), 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 the traditional approach even 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 \( \rho \).

Figure 1 presents these results graphically. Each panel plots the empirical MSE of each procedure for different values of \( \rho \), 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 \( \rho \) increases. In particular, the both columns show that the cluster-robust procedure performs similarly to the traditional procedure for small values of \( \rho \), and performs significantly better as \( \rho \) increases, when 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, and the usual tendency of researchers to attempt “undersmoothing,” it is useful to consider a situation where local linear estimators will struggle with bias due to extreme curvature of the conditional mean function near the cutoff. Therefore, in the second design we use a DGP studied in CCT where the mean function is altered so that typical estimators will be heavily biased:

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

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 the traditional bandwidth selection procedure. As before, accounting for clustering becomes more important as cluster size or \( \rho \) 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). 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), Baum-Snow and Marion (2009), and Freedman and McGavock (2015), projects located in QCTs are eligible for up to 30 percent larger tax credits than projects in tracts not labeled as QCTs. Importantly, during the time period studied this designation is based on the fraction of households whose income falls below 60 percent of Area Median Gross Income (AMGI). If the majority of households in a census tract have household income less than 60 percent of AMGI, the tract 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. As in Baum-Snow and Marion (2009), we restrict to census tracts in metropolitan areas, and exclude Alaska and Hawaii. 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.

\footnote{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.}
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. In other words, both procedures using the microdata perform inference with the same “cluster-robust” standard error formulas, while tract-level regressions utilize heteroskedasticity-robust standard errors. The first column presents the results applying cluster-robust bandwidth selection procedures, as described above, applied to the microdata. Next, the second column presents results using the traditional 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 and 4, the cluster-robust optimal bandwidth should be similar to the usual bandwidth choices in the absence of data dependence. The sizable differences between the bandwidth values 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 neighborhood characteristics at the QCT designation threshold. Note that 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. However, there is clear evidence of jumps in the implementation of new LIHTC units 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 procedure that imposes i.i.d. data on the microdata produces a negative 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 positive effect of QCT status on the number of LIHTC projects in a tract, as intended by policymakers. Also, when focusing on the number of LIHTC units available in the tract, the three approaches provide very different optimal bandwidths and estimates, even though they all indicate a positive effect of the tax credits as the policy was designed to achieve.

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 there is positive dependence within clusters. 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 account for clustering. In fact, on the whole both the cluster-robust and the aggregated data procedures provide similar results, giving 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.

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 algorithms in Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) to these situations. This setup encompasses a number of common applications, such as those with treatment being assigned at a higher level than the unit of observation or discrete running variables. The insights provided by this analysis can also be applied directly to fuzzy and kink RD designs, and the robust (bias-corrected) methods in Calonico, Cattaneo, and Titiunik (2014). 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


Table 1: Simulation Results – DGP 1

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>250 Clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0021</td>
<td>0.0040</td>
<td>0.0088</td>
<td>0.0149</td>
<td>0.0352</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0021</td>
<td>0.0052</td>
<td>0.0153</td>
<td>0.0593</td>
<td>0.2743</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9921</td>
<td>0.7707</td>
<td>0.5778</td>
<td>0.2523</td>
<td>0.1283</td>
</tr>
<tr>
<td>Size=25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0006</td>
<td>0.0030</td>
<td>0.0068</td>
<td>0.0138</td>
<td>0.0314</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0007</td>
<td>0.0062</td>
<td>0.0755</td>
<td>0.1127</td>
<td>4.2636</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9712</td>
<td>0.4868</td>
<td>0.0895</td>
<td>0.1224</td>
<td>0.0074</td>
</tr>
<tr>
<td>500 Clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0011</td>
<td>0.0021</td>
<td>0.0036</td>
<td>0.0062</td>
<td>0.0153</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0011</td>
<td>0.0023</td>
<td>0.0047</td>
<td>0.0111</td>
<td>0.0403</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.0047</td>
<td>0.9072</td>
<td>0.7611</td>
<td>0.5629</td>
<td>0.3805</td>
</tr>
<tr>
<td>Size=25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0004</td>
<td>0.0015</td>
<td>0.0030</td>
<td>0.0063</td>
<td>0.0157</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0004</td>
<td>0.0020</td>
<td>0.0061</td>
<td>0.0181</td>
<td>0.0765</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9523</td>
<td>0.7624</td>
<td>0.5007</td>
<td>0.3481</td>
<td>0.2051</td>
</tr>
<tr>
<td>Size=200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0001</td>
<td>0.0014</td>
<td>0.0032</td>
<td>0.0065</td>
<td>0.0164</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0001</td>
<td>0.0033</td>
<td>0.0197</td>
<td>0.0843</td>
<td>9.9816</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9591</td>
<td>0.4300</td>
<td>0.1620</td>
<td>0.0771</td>
<td>0.0016</td>
</tr>
<tr>
<td>1000 Clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0007</td>
<td>0.0012</td>
<td>0.0019</td>
<td>0.0035</td>
<td>0.0069</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0007</td>
<td>0.0012</td>
<td>0.0022</td>
<td>0.0054</td>
<td>0.0137</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9895</td>
<td>0.9641</td>
<td>0.8697</td>
<td>0.6414</td>
<td>0.5053</td>
</tr>
<tr>
<td>Size=25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0002</td>
<td>0.0009</td>
<td>0.0017</td>
<td>0.0031</td>
<td>0.0078</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0002</td>
<td>0.0010</td>
<td>0.0027</td>
<td>0.0075</td>
<td>0.0280</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9711</td>
<td>0.8983</td>
<td>0.6248</td>
<td>0.4109</td>
<td>0.2775</td>
</tr>
<tr>
<td>Size=200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0000</td>
<td>0.0008</td>
<td>0.0016</td>
<td>0.0032</td>
<td>0.0077</td>
</tr>
<tr>
<td>Traditional Bandwidth MSE</td>
<td>0.0000</td>
<td>0.0017</td>
<td>0.0054</td>
<td>0.0184</td>
<td>0.1166</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9879</td>
<td>0.4789</td>
<td>0.2892</td>
<td>0.1744</td>
<td>0.0656</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Clusters</th>
<th>( \rho )</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 250 ) Clusters</td>
<td></td>
<td>( \rho )</td>
<td>0.0021</td>
<td>0.0044</td>
<td>0.0082</td>
<td>0.0151</td>
</tr>
<tr>
<td>( \text{Size}=5 )</td>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0001</td>
<td>0.0020</td>
<td>0.0036</td>
<td>0.0067</td>
<td>0.0155</td>
</tr>
<tr>
<td></td>
<td>Traditional Bandwidth MSE</td>
<td>0.0010</td>
<td>0.0023</td>
<td>0.0047</td>
<td>0.0110</td>
<td>0.0372</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>1.0140</td>
<td>0.9023</td>
<td>0.7620</td>
<td>0.6088</td>
<td>0.4174</td>
</tr>
<tr>
<td>( \text{Size}=25 )</td>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0003</td>
<td>0.0014</td>
<td>0.0032</td>
<td>0.0064</td>
<td>0.0159</td>
</tr>
<tr>
<td></td>
<td>Traditional Bandwidth MSE</td>
<td>0.0003</td>
<td>0.0020</td>
<td>0.0062</td>
<td>0.0181</td>
<td>0.0793</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>1.0077</td>
<td>0.7131</td>
<td>0.5187</td>
<td>0.3538</td>
<td>0.2006</td>
</tr>
<tr>
<td>( \text{Size}=200 )</td>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0001</td>
<td>0.0013</td>
<td>0.0031</td>
<td>0.0061</td>
<td>0.0167</td>
</tr>
<tr>
<td></td>
<td>Traditional Bandwidth MSE</td>
<td>0.0000</td>
<td>0.0036</td>
<td>0.0408</td>
<td>0.1642</td>
<td>9.9464</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>1.0192</td>
<td>0.3719</td>
<td>0.0771</td>
<td>0.0374</td>
<td>0.0017</td>
</tr>
<tr>
<td>( \text{Size}=200 )</td>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0005</td>
<td>0.0011</td>
<td>0.0018</td>
<td>0.0034</td>
<td>0.0073</td>
</tr>
<tr>
<td></td>
<td>Traditional Bandwidth MSE</td>
<td>0.0005</td>
<td>0.0012</td>
<td>0.0021</td>
<td>0.0047</td>
<td>0.0123</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>1.0040</td>
<td>0.9346</td>
<td>0.8442</td>
<td>0.7345</td>
<td>0.6037</td>
</tr>
<tr>
<td>( \text{Size}=200 )</td>
<td>Cluster-Robust Bandwidth MSE</td>
<td>0.0001</td>
<td>0.0008</td>
<td>0.0016</td>
<td>0.0032</td>
<td>0.0076</td>
</tr>
<tr>
<td></td>
<td>Traditional Bandwidth MSE</td>
<td>0.0001</td>
<td>0.0010</td>
<td>0.0025</td>
<td>0.0066</td>
<td>0.0223</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>1.0035</td>
<td>0.7635</td>
<td>0.6269</td>
<td>0.4858</td>
<td>0.3629</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Cluster-Robust Bandwidth</th>
<th>Traditional Bandwidth</th>
<th>Tract-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[SE]</td>
<td>[SE]</td>
<td>[SE]</td>
</tr>
<tr>
<td></td>
<td>w=</td>
<td>w=</td>
<td>w=</td>
</tr>
<tr>
<td>Homeownership</td>
<td>-0.0089 [0.0130]</td>
<td>-0.0335 [0.0217]</td>
<td>-0.0042 [0.0153]</td>
</tr>
<tr>
<td></td>
<td>w=0.098</td>
<td>w=0.027</td>
<td>w=0.105</td>
</tr>
<tr>
<td>Fraction Non-White</td>
<td>0.0024 [0.0173]</td>
<td>-0.0048 [0.0386]</td>
<td>0.0042 [0.0153]</td>
</tr>
<tr>
<td></td>
<td>w=0.107</td>
<td>w=0.022</td>
<td>w=0.105</td>
</tr>
<tr>
<td>High School Diploma or Higher</td>
<td>-0.0073 [0.0078]</td>
<td>-0.0006 [0.0123]</td>
<td>-0.0105* [0.0061]</td>
</tr>
<tr>
<td></td>
<td>w=0.129</td>
<td>w=0.052</td>
<td>w=0.121</td>
</tr>
<tr>
<td>Bachelors Degree or Higher</td>
<td>0.0039 [0.0057]</td>
<td>0.0041 [0.0077]</td>
<td>0.0024 [0.0064]</td>
</tr>
<tr>
<td></td>
<td>w=0.115</td>
<td>w=0.067</td>
<td>w=0.103</td>
</tr>
<tr>
<td>Employment Rate</td>
<td>0.0081 [0.0054]</td>
<td>0.0218** [0.0093]</td>
<td>-0.0028 [0.0045]</td>
</tr>
<tr>
<td></td>
<td>w=0.082</td>
<td>w=0.027</td>
<td>w=0.121</td>
</tr>
<tr>
<td></td>
<td>w=0.099</td>
<td>w=0.018</td>
<td>w=0.103</td>
</tr>
<tr>
<td>Number of LIHTC Projects</td>
<td>0.0341 [0.0581]</td>
<td>-0.0473 [0.1020]</td>
<td>0.0662* [0.0258]</td>
</tr>
<tr>
<td></td>
<td>w=0.068</td>
<td>w=0.006</td>
<td>w=0.122</td>
</tr>
<tr>
<td>N</td>
<td>30,330,540</td>
<td>30,330,540</td>
<td>45,294</td>
</tr>
<tr>
<td>N Clusters</td>
<td>44,716</td>
<td>44,716</td>
<td>45,294</td>
</tr>
</tbody>
</table>

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 CCT bandwidths unadjusted for clustering. All point estimates are not bias corrected, and are derived 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

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

The proofs and notation presented in this appendix are based and follow as close as possible the proofs in IK and CCT regarding the asymptotic properties of the local polynomial estimators used in RD designs as well as the choice of MSE-optimal bandwidths.

Lemma A.1. Define $F_j = \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{N_g} K_h(z_{ig})z_{ij} = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{N_g} \sum_{i=1}^{N_g} K_h(z_{ig})z_{ij} = \frac{1}{G} \sum_{g=1}^{G} N_g A_{jg}$, where $A_{jg} = \frac{1}{N_g} \sum_{i=1}^{N_g} K_h(z_{ig})z_{ij}$. 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 $\nu_j$ 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 $g = 1, ..., G$:

$$E[A_{jg}] = E \left[ \frac{1}{N_g} \sum_{i=1}^{N_g} K_h(z_{ig})z_{ig}^j \right] = h^j \int_0^{\infty} K(x) x^j f(hx)dx$$

$$= h^j \int_0^{\infty} K(x) x^j f(0)dx + h^{j+1} \int_0^{\infty} K(x) x^{j+1} f(hx) - f(0) \frac{f(0)}{hx} dx$$

$$= h^j f(0) \nu_j + O(h^{j+1})$$
Then,

\[ E[F_j] = \frac{1}{G} \sum_{g=1}^{G} N_g E[A_{jg}] \]

\[ = N_g h^j f(0) v_j + O(h^{j+1}) \]

For the variance,

\[ Var[A_{jg}] = E[A_{jg}^2] - E[A_{jg}]^2 \]

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

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

\[ Var[F_j] = Var \left[ \frac{1}{G} \sum_{g=1}^{G} N_g A_{jg} \right] = \frac{1}{G^2} \sum_{g=1}^{G} N_g^2 Var[A_{jg}] = \frac{1}{G^2} \sum_{g=1}^{G} O(h^{2j-1}) = o \left( (h^j)^2 \right) \]

Then,

\[ F_j = E[F_j] + O_p(\text{Var}(F_j)^{1/2}) \]

\[ = N_g h^j (f(0) v_j + o_p(1)) \]

\[ \square \]

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 \( Q_{tj} = G^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{s=1}^{N_i} K_h^2(z_g) z_g^{t+j} \sigma_{is} (z_g) \). Then,

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

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 \( \pi^j \) defined in the text.
Proof.

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

\[ = G^{-1} \sum_{g=1}^{G} \int_0^{\infty} \frac{1}{h^2} K^2 \left( \frac{z}{h} \right) z_g^{t+j} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}(z) f(z) dz \]

\[ = h^{t+j-1} f(0) \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}(0) \int_{-\infty}^{\infty} K^2(x)x^{t+j} dx + O(h^{t+j}) \]

\[ = h^{t+j-1} f(0) \pi_{t+j} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}(0) + O(h^{t+j}) \]

To bound \( \text{Var}(Q_{tj}|X) \):

\[ \text{Var}[Q_{tj}] = E[(Q_{tj})^2] - E[Q_{tj}]^2 \]

The first term:

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

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

Note that all cross products will be of the type:

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

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

\[ E \left[ K_h^4(z_g)z_g^{2q} \sigma_{is}(z_g) \sigma_{tl}(z_g) \right] = \int_0^{\infty} K_h^4(z)z^{2q} \sigma_{is}(z) \sigma_{tl}(z) f(z) dz \]

\[ = h^{2q-3} \int_0^{\infty} K^4(x)x^{2q} \sigma_{is}(hx) \sigma_{tl}(hx) f(hx) dx \]

\[ = O(h^{2q-3}) = o\left(h^{q-1}\right)^2 \]
Then,

\[
E \left[ (Q_{tj})^2 \right] = N^{-2} \sum_{g=1}^{G} E \left[ K_h^4(z_g)z_g^{2(t+j)} \left( \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{gis} (z_g) \right)^2 \right]
\]

\[
= G^{-2} \sum_{g=1}^{G} \sum_{t=1}^{N_g} \sum_{i=1}^{N_g} \sum_{l=1}^{N_g} E \left[ K_h^4(z_g)z_g^{2(t+j)} \sigma_{il} (z_g) \sigma_{il} (z_g) \right]
\]

\[
= G^{-1} N^4 o \left( (h^{t+j})^{-1} \right) = o \left( (G^{-\frac{1}{2}} N^2 h_{t+j}^{-1})^2 \right) = o \left( (h^{t+j})^{-1} \right)
\]

and

\[
Q_{tj} = E[Q_{tj}] + O_p(Var(Q_{tj})^\frac{1}{2})
\]

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

With 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. These proofs follow closely those of IK and CCT.

**Proof.** Proof of Lemma 3.1 For analyzing the asymptotic approximation to the bias term, note that \( y_{ig} = \mu(x_{ig}) + \epsilon_{ig} \). Let \( R = \left( \begin{array}{c} \epsilon \ X \ \cdots \ X^p \end{array} \right) \) with typical row given by \( r_p(x) = \left( \begin{array}{c} 1 \ x \ \cdots \ x^p \end{array} \right) \) and \( e^\eta_0 \) be a vector of zeros except for the \((\eta+1)^{th}\) entry equal to one, e.g., \( e^\eta_0 = \left( \begin{array}{c} 1 \ 0 \ \cdots \ 0 \end{array} \right) \). Then,

\[
\dot{\mu}_{tj}^{(n)} = \eta \epsilon^{\eta}_n (R'WR)^{-1} R'WY = \eta \epsilon^{\eta}_n (R'WR)^{-1} R'W [\mu(X) + \epsilon]
\]

\[
= \eta \epsilon^{\eta}_n (R'WR)^{-1} R'W \mu(X) + \eta \epsilon^{\eta}_n (R'WR)^{-1} R'W \epsilon
\]

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 \( E(\dot{\mu}(0)|X) = \epsilon^\eta_0 (R'WR)^{-1} R'WM \), 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 \( M = (\mu(x_{11}), \mu(x_{21}), \ldots, m(x_{12}), \mu(x_{22}), \ldots, \mu(x_{NG}))' \). Then

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

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

Then,

\[
Bias(\hat{\mu}^{(\eta)}) = \eta! e_\eta' (R'W)^{-1} R'WM - \mu^{(\eta)}(0) = \eta! e_\eta' (R'W)^{-1} R'W (S + T)
\]

Note that,

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

Using the definition and results on Lemma A.1:

\[
\frac{1}{G} R'WR = \begin{pmatrix}
F_0 & F_1 & \cdots & F_p \\
F_1 & F_2 & \cdots & F_{p+1} \\
\vdots & \vdots & \ddots & \vdots \\
F_p & F_{p+1} & \cdots & F_{2p}
\end{pmatrix} = H \begin{pmatrix}
F_0^* + \eta_p(1) & F_1^* + \eta_p(1) & \cdots & F_p^* + \eta_p(1) \\
F_1^* + \eta_p(1) & F_2^* + \eta_p(1) & \cdots & F_{p+1}^* + \eta_p(1) \\
\vdots & \vdots & \ddots & \vdots \\
F_p^* + \eta_p(1) & F_{p+1}^* + \eta_p(1) & \cdots & F_{2p}^* + \eta_p(1)
\end{pmatrix}
\]

where, \( H = diag(1, h, \ldots, h^p) \). Recalling that \( F_j^* = N_g f(0) \nu_j \) and that \( \frac{1}{f(0)} \eta_p(1) = o_p(1) \) we can obtain its inverse:
\[
\left( \frac{1}{G} R'W \right)^{-1} = \frac{1}{f(0) N_g} H^{-1} \begin{pmatrix}
\nu_0 + o_p(1) & \nu_1 + o_p(1) & \cdots & \nu_p + o_p(1) \\
\nu_1 + o_p(1) & \nu_2 + o_p(1) & \cdots & \nu_{p+1} + o_p(1) \\
\vdots & \vdots & \ddots & \vdots \\
\nu_p + o_p(1) & \nu_{p+1} + o_p(1) & \cdots & \nu_{2p} + o_p(1)
\end{pmatrix}^{-1} H^{-1}
\]

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

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

Where \( \gamma_{ij} \) is a deterministic function of \( \nu \) known and computable for a given kernel and polynomial order. Examining \( \frac{1}{G} R'WT \):

\[
\left| \frac{1}{G} R'WT \right| \leq \frac{1}{G} R' \left| W \right| \left( \sup_x |\mu^{(p+2)}(x)||x_{11}^{p+2}| \right) = \sup_x |\mu^{(p+2)}(x)| \left( \begin{array}{c} F_{p+2} \\ \vdots \\ F_{2(p+1)} \end{array} \right) \leq \left( \begin{array}{c} o_p(h^{p+1}) \\ \vdots \\ o_p(h^{2p+1}) \end{array} \right)
\]

Combining the results above, we obtain

\[
e'_{\eta} (R'W)R^{-1} = o_p(h^{p+1-\eta}).
\]

For the first term, \( \frac{1}{G} R'WS \),

\[
\frac{1}{G} R'WS = \frac{\mu^{(p+1)}(0)}{(p+1)!} \left( \frac{1}{G} \sum_{i=1}^N K_h(X_i)X_i^{p+1} \right) \left( \begin{array}{c} F_{p+1} \\ \vdots \\ F_{2p+1} \end{array} \right) = \frac{\mu^{(p+1)}(0)}{(p+1)!} N_g f(0) \left( \begin{array}{c} \nu_{p+1} h^{p+1} + o_p(h^{p+1}) \\ \vdots \\ \nu_{2p+1} h^{2p+1} + o_p(h^{2p+1}) \end{array} \right)
\]
Let $\Gamma^{-1}$ be a $(p+1) \times (p+1)$ matrix with typical element $\gamma_{ij}$. Then,

$$
e^\prime_s (R'WR)^{-1} R'WS = \frac{h^{p+1-\eta}}{(p+1)!} \mu^{(p+1)}(0)e^\prime_s \Gamma^{-1} \left(\begin{array}{c}
\nu_{p+1} \\
\vdots \\
\nu_{2p+1}
\end{array}\right) + o_p(h^{p+1-\eta})$$

Hence,

$$E[\hat{\mu}^{(\eta)} - \mu^{(\eta)}|X] = \eta! \frac{h^{p+1-\eta}}{(p+1)!} \mu^{(p+1)} e^\prime_s \Gamma^{-1} \left(\begin{array}{c}
\nu_{p+1} \\
\vdots \\
\nu_{2p+1}
\end{array}\right) + o_p(h^{p+1-\eta})$$

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:

$$V(\hat{\mu}^{(\eta)}(0)|X) = \eta!^2 e^\prime_s (R'WR)^{-1} R'W \Sigma R(R'WR)^{-1} e_s$$

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

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

Where $\sigma_{ij}$ is the term in the $i$ - $th$ line and $j$ - $th$ column in $\Omega_g$.

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

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

Focusing on the case that $X$ is defined at the cluster level and, hence, $x_{ig} = x_g \forall i = 1, \cdots, N_g$,
Note that each term in matrix $A$ will be a combination of products of the terms $\nu_j$ and $\pi_j$ plus an $o_p(1)$ term, hence

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

Then,

$$G(R' WR)^{-1} R' W \Sigma R (R' WR)^{-1} = \frac{1}{h f(0)} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}(0) H^{-1} A H^{-1}$$

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

$$(G(R' WR)^{-1} R' W \Sigma R (R' WR)^{-1})_{ij} = \frac{[a_{ij} + o_p(1)]}{h^{i+j-2}} \frac{1}{h f(0)} \sum_{t=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{ts}(0)$$

Where $a_{ij}$ is a deterministic function of $\nu$ and $\pi$ known and computable for a given kernel and polynomial order.

$$G \left( e_\eta' (R' WR)^{-1} R' W \Sigma R (R' WR)^{-1} e_\eta \right) = \frac{[a_{ij} + o_p(1)]}{h^{2\eta+1} f(0)} \sum_{t=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{ts}(0)$$

It follows then that

$$\text{Var}[\hat{\mu}^{(n)}_+ - \mu^{(n)}_+ | X] = \eta^2 \frac{1}{GH^{2\eta+1} f(0)} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^+ e_\eta' \Gamma^{-1} \Delta \Gamma^{-1} e_\eta + o_p \left( \frac{1}{GH^{2\eta+1}} \right)$$

$$= \eta^2 \frac{1}{Nh^{2\eta+1} f(0)} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is}^+ e_\eta' \Gamma^{-1} \Delta \Gamma^{-1} e_\eta + o_p \left( \frac{1}{GH^{2\eta+1}} \right)$$

**Asymptotic Distribution**

We have seen that:

$$\frac{\hat{\mu}^{(n)}_+ - \mu^{(n)}_+}{\sqrt{\text{Var}[\hat{\mu}^{(n)}_+ - \mu^{(n)}_+ | X]}} = \frac{\hat{\mu}^{(n)}_+ - E[\hat{\mu}^{(n)}_+ | X] + E[\hat{\mu}^{(n)}_+ | X] - \mu^{(n)}_+}{\sqrt{\text{Var}[\hat{\mu}^{(n)}_+ - \mu^{(n)}_+ | X]}}$$

$$= \varepsilon_1 + \varepsilon_2 = \varepsilon_1 + o_p(1)$$
Then,

\[ \varepsilon_1 = \left( \text{Var}[\hat{\mu}_+^{(\eta)} - \mu_+ | X] \right)^{-\frac{1}{2}} \left( \hat{\mu}_+^{(\eta)} - E[\hat{\mu}_+^{(\eta)} | X] \right) = \left( \text{Var}[\mu_+^{(\eta)} - \mu_+ | X] \right)^{-\frac{1}{2}} \left( \eta! \epsilon'(R'WR)^{-1}R'W \epsilon \right) \]

and,

\[ \varepsilon_2 = \frac{E[\hat{\mu}_+^{(\eta)} | X] - \mu_+^{(\eta)}}{\sqrt{\text{Var}[\hat{\mu}_+^{(\eta)}] - \mu_+^{(\eta)} | X]} = O_p \left( \sqrt{Gh^{d+2p}} \right) = o_p(1) \]

Note that, since clusters have the same variable for the running variable,

\[ R'W \epsilon = \sum_{g=1}^{G} R'_g W_g \epsilon_g = \sum_{g=1}^{G} K(x_g) R'_g \epsilon_g = \sum_{g=1}^{G} K(x_g) r_p(x_g) \sum_{i=1}^{N_g} \epsilon_{ig} \]

and, \( \varepsilon_1 = \bar{\varepsilon}_1 + o_p(1) \), where \( \bar{\varepsilon}_1 = \sum_{g=1}^{G} \omega_g \epsilon_g \) and

\[ \omega_g = \left( \frac{1}{G N_g h^{2q+1} f(0)} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{is} \epsilon'_s \Gamma^{-1} \Delta \Gamma^{-1} \epsilon \right)^{-\frac{1}{2}} \frac{h^{-\eta} \epsilon'_g \Gamma^{-1} K(x_g) r_p(x_g)}{G} \]

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

Apart from the introduction of cross-sectional dependence within cluster in the definition of the variance term, the proof of Lemma 3.2 below follows from CCT.

**Proof.** Proof of Lemma 3.2

**MSE(h):**
$$E[(\hat{\tau}(\eta) - \tau(\eta))^2 | X] = \text{Var}[\hat{\mu}_+ - \mu_+ | X] + \text{Var}[\hat{\mu}_- - \mu_- | X] + \{ E[\hat{\mu}_+ - \mu_+] - E[\hat{\mu}_- - \mu_- | X] \}$$

$$= \left[ \eta!^2 \left( \frac{\sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \sigma_+^2}{GN_2^2 N_2^2 f(0)} + \frac{\sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \sigma_-^2}{GN_2^2 h^{2p+1} f(0)} \right) e_q' \Gamma^{-1} e_q + o_p \left( \frac{1}{h^{2(p+1)}} \right) \right]^2$$

$$+ \left[ \eta! h^{2p+1-\eta} \left( \frac{(p+1)\Delta - 1}{(p+1)!} \right) e_q' \Gamma^{-1} e_q + o_p (h^{2(p+1-\eta)}) \right]^2$$

$$= \frac{1}{Gh^{2\eta+1}} C_{2\eta} \left[ \sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \frac{\sigma_+^2}{N_2^2 f(0)} + \sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \frac{\sigma_-^2}{N_2^2 f(0)} \right]$$

$$+ h^{2(p+1-\eta)} C_{1\eta} \left[ \mu_+^{(p+1)} - (1)^{(p+1)} \mu_-^{(p+1)} \right]^2 + o_p \left( \frac{1}{Gh^{2\eta+1}} + h^{2(p+1-\eta)} \right)$$

Where $C_{1\eta} = \left[ \frac{\eta!}{(p+1)!} e_q' \Gamma^{-1} e_q \right]^2$ and $C_{2\eta} = \eta!^2 e_q' \Gamma^{-1} e_q$. The optimal bandwidth $h_{opt}$ solves, as in CCT,

$$h_{opt} = \arg \min \left( C_{2\eta} \left[ \sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \frac{\sigma_+^2}{N_2^2 f(0)} + \sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \frac{\sigma_-^2}{N_2^2 f(0)} \right] + h^{2(p+1-\eta)} C_{1\eta} \left[ \mu_+^{(p+1)} - (1)^{(p+1)} \mu_-^{(p+1)} \right]^2 \right)$$

$$= \left[ \frac{\sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \frac{\sigma_+^2}{N_2^2 f(0)} + \sum_{g=1}^{N_g} \sum_{s=1}^{N_s} \frac{\sigma_-^2}{N_2^2 f(0)}}{\mu_+^{(p+1)} - (1)^{(p+1)} \mu_-^{(p+1)}} \right]^{\frac{1}{p+1}}$$

where $C_{kn} = \frac{2^{p+1}}{2(p+1-\eta)} C_{2\eta}$. \(\Box\)