8-2016

Working Paper Number 16001

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

Regression Discontinuity Designs with Clustered Data: Variance and Bandwidth Choice

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 fill this gap, we derive the properties of traditional local polynomial estimators in a fixed-G setting that allows for cluster dependence in the error term. Simulation results demonstrate that accounting for clustering in the data while selecting bandwidths may lead to lower MSE while maintaining proper coverage. We then apply our cluster-robust procedure to an application examining 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

Disciplines
Econometrics

Regression Discontinuity Designs with Clustered Data *

Otávio Bartalotti and Quentin Brummet†

August 14, 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 fill this gap, we derive the properties of traditional local polynomial estimators in a fixed-G setting that allows for cluster dependence in the error term. Simulation results demonstrate that accounting for clustering in the data while selecting bandwidths may lead to lower MSE while maintaining proper coverage. We then apply our cluster-robust procedure to an application examining 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 especially grateful to Matias Cattaneo for suggestions that significantly improved the paper. We also thank 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, Advances in Econometrics RD Conference, 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 due to their ability to provide consistent estimation of causal effects under transparent assumptions. Nonetheless, the current literature on bandwidth selection and inference in RD designs typically assumes that the observations around the cutoff are independent. While many researchers explicitly consider cluster dependence when conducting inference, commonly used bandwidth selection procedures such as those in Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) explicitly assume i.i.d. data. Therefore, researchers performing RD designs with clustered data are left with the choice of either using an *ad hoc* bandwidth or relying on a bandwidth selection procedure whose assumptions are clearly violated.

We fill this gap in the RD literature by analyzing the properties of local polynomial estimators of treatment effects in RD designs with clustering and deriving cluster-robust optimal bandwidth selectors. We derive these formulas under two setups. In our first general analysis, we use fixed-G calculations to derive the properties of the estimators in a setup that allows for unrestricted dependence among observations within clusters. These results demonstrate that the widely used "cluster-robust" standard error formulas are appropriate in this setting. This finding relates to the results in Imbens and Lemieux (2008) and Calonico, Cattaneo, and Farrell (2016) (henceforth, "CCF") for the usual i.i.d. case. It is also connected to Lee and Card (2008), who suggest the use of cluster-robust standard errors to account for specification errors in RD designs with discrete running variables. Our analysis demonstrates that in the context of our framework, the intuitive idea of using cluster-robust standard errors is valid even when using non-parametric local polynomial estimators. In addition, we extend the Mean Squared Error (MSE)-optimal bandwidth choice procedures proposed by Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) (henceforth, "IK" and "CCT," respectively) to allow for unrestricted dependence structure within cluster, where the resulting optimal bandwidth collapses to traditional optimal bandwidth selectors when observations are independent.

Additionally, we consider the special case in which clustering is defined at the running variable level, for which asymptotic approximations can be obtained. The implementation of this optimal bandwidth selector is identical to the analysis above, and this analysis provides further insight into the effects of clustering on inference and bandwidth selection in RD designs. This is particularly relevant for two scenarios that are common in the applied literature. 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 in using student-level microdata to examine a policy implemented based on a school-level running variable. A framework that considers clustering 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 is that of RD designs with a discrete running variable. A cluster-robust bandwidth choice procedure allows researchers using discrete running variables to select bandwidths in a manner that is compatible with the cluster-robust inference procedure suggested by Lee and Card (2008).

We then present a simulation study which demonstrates that a cluster-robust bandwidth choice procedure outperforms traditional bandwidth choices in terms of MSE in many practical settings without any noticeable decline in coverage. We then illustrate 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 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 notation. Section 3 then presents our main results related to MSE and optimal bandwidth selection and discusses the special case in which the clusters are defined at the running variable level. Section 4 shows the distributional approximations of the statistics used for performing inference, and Section 6 then provides a small simulation study. Finally, Section 7 presents the application to the impacts of Low-Income Housing Tax Credits on neighborhood characteristics, and Section 8 concludes.

2 Setup and Notation

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, has density given by $f(X)$ and 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

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

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

$$\tau = m_+ - m_-$$
$$\text{where } m_+ = lim_{x\to 0^+} m(x), \text{ and } m_- = lim_{x\to 0^-} m(x)$$

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 (**?**). Let $m^{(\eta)}(x) = \frac{d^\eta m(x)}{dx^\eta}$ be the η^{th} derivative of the unknown regression function and define $m_+^{(\eta)} = lim_{x\to 0^+} m^{(\eta)}(x)$ and $m_-^{(\eta)} = lim_{x\to 0^-} m^{(\eta)}(x)$. The parameter of interest in these cases is given by $\tau^{(\eta)} = m_+^{(\eta)} - m_-^{(\eta)}$.

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, 2016). For a local polynomial of order p, we use the following estimator:

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

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

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

Building on this traditional RD setup, we now turn to the setting where there is cluster dependence in the data. 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 size N_g, we observe

$$Y_{ig} = m(x_{ig}) + \epsilon_{ig} \tag{3}$$

Where the subscript ig refers to unit i in cluster g.

The notation used in this paper follows closely the definitions established in CCF. Let $x_{h,ig} = \frac{x_{ig}-\bar{x}}{h}$ where $\bar{x}$ is the treatment cutoff which is set to zero without loss of generality. Also, let $r_p(u) = (1, u, u^2, \ldots, u^p)'$ and $R_p = [r_p(x_{h,11}), r_p(x_{h,21}), \cdots, r_p(x_{h,N_GG})]'$ and e'_η be a vector of zeros except for the $(\eta+1)^{th}$ entry equal to one, e.g., $e'_0 = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$. Define the weighting matrix $W_{+,p,g} = diag(h^{-1}\mathbb{1}[x_{1g} > 0]K(x_{h,1g}), h^{-1}\mathbb{1}[x_{2g} > 0]K(x_{h,2g}), \ldots, h^{-1}\mathbb{1}[x_{N_gg} > 0]K(x_{h,N_gg}))$ and $W_{-,p,g} = diag(h^{-1}\mathbb{1}[x_{1g} \leq 0]K(x_{h,1g}), h^{-1}\mathbb{1}[x_{2g} \leq 0]K(x_{h,2g}), \ldots, h^{-1}\mathbb{1}[x_{N_gg} \leq 0]K(x_{h,N_gg}))$. Then, let $W_{+,p}$ and $W_{-,p}$ be the block diagonal matrices with g-th block $W_{+,p,g}$ and $W_{-,p,g}$ for $g = 1, \ldots, G$, respectively. Also, define $H_p = diag(1, h^{-1}, \ldots, h^{-p})$. Finally, let $\Gamma_{+,p} = R'_pW_{+,p}R_p$, $\Gamma_{-,p} = R'_pW_{-,p}R_p$, $\Lambda_{+,p} = R'_pW_{+,p}\left[x^{p+1}_{h,11}, \ldots, x^{p+1}_{h,N_GG}\right]'$ and $\Lambda_{-,p} = R'_pW_{-,p}\left[x^{p+1}_{h,11}, \ldots, x^{p+1}_{h,N_GG}\right]'$. Also define $\Psi_{+,p} = R'_pW_{+,p}\Sigma W_{+,p}R_p$ and $\Psi_{-,p} = R'_pW_{-,p}\Sigma W_{-,p}R_p$, where Σ is the block diagonal matrix with typical block given by $\Omega_g(x) = Var(Y_g|X)$. It is useful to to establish notation for the (i,s) elements of the variance matrix Ω_g as σ_{gis} and its limits around the threshold $\sigma_{+,is} = lim_{x\to 0^+}\sigma_{gis}$ and $\sigma_{-,is} = lim_{x\to 0^-}\sigma_{gis}$ for describing the cluster at the running variable case. Note that this allows Ω_g to vary over X, and we do not restrict cluster size to be identical across clusters.

More details are provided in the appendix. Some useful calculations can be obtained by using the fixed-G expectation over the support of X to some of the objects defined below. Hence, adopting the notation on CCF, let $\tilde{\Lambda}_{+,p} = \mathbb{E}[\Lambda_{+,p}]$, and similarly for $\Lambda_{-,p}$, $\Gamma_{+,p}$, $\Gamma_{-,p}$, $\Psi_{+,p}$, and $\Psi_{-,p}$ (See Lemma A.1 and Lemma A.2).

In Section 3 we present two sets of results. First, to address the issue of variance estimation and bandwidth selection in the general clustering case, we use fixed-G expectations calculated conditionally on the observed values of the running variable, following Calonico, Cattaneo, and Farrell (2016). This provides a valid MSE function and accompanying optimal bandwidth rule that take account of cluster dependence in the data.

Second, we focus on the special case in which the clusters are defined at the running variable level, implying that all units in a cluster have the same value, X_g, for the running variable. In this case, cluster dependence does not vanish asymptotically even as the bandwidth shrinks. Our asymptotic approximations based on $G \to \infty$, $h \to 0$, and $Gh \to \infty$ provide insight on the sources of improvements achieved by recognizing the presence of cluster dependence.

3 Main Results

To capture the impact of cluster dependence on the behavior of the estimator, $\hat{\tau}^{(\eta)}$, we first utilize fixed-G calculations for variance and MSE obtained conditional on the observed values of the running variable following the approach in CCF.

Theorem 3.1 presents the MSE approximation for $\hat{\tau}^{(\eta)}$ and its corresponding MSE-optimal bandwidth for the general clustering case, as well as the conditional bias and variance formulas and their limiting expressions. Assumptions and proofs are provided in the appendix.

Theorem 3.1. *Under assumptions 1-6, the conditional Mean Squared Error for $\hat{\tau}^{(\eta)}$ has fixed-G approximation given by,*

$$E[(\hat{\tau}^{(\eta)} - \tau^{(\eta)})^2 | x_{11}, \ldots, x_{N_G G}] = \frac{1}{Gh^{2\eta+1}} [\mathbf{V_n}] + h^{2(p+1-\eta)} \left[\mathbf{B_n}^2 + o_p(1)\right] \tag{4}$$

$$= \frac{1}{Gh^{2\eta+1}} \left[\tilde{\mathbf{V}} + o_p(1)\right] + h^{2(p+1-\eta)} \left[\tilde{\mathbf{B}}^2 + o_p(1)\right]. \tag{5}$$

Where,

$$\mathbf{V_n} = \eta!^2 e_\eta' \left[\Gamma_{+,p}^{-1}\Psi_{+,p}\Gamma_{+,p}^{-1} + \Gamma_{-,p}^{-1}\Psi_{-,p}\Gamma_{-,p}^{-1}\right] e_\eta \tag{6}$$

$$\mathbf{B_n} = \frac{\eta!}{(p+1)!} e_\eta' \left[m_+^{(p+1)}\Gamma_{+,p}^{-1}\Lambda_{+,p} - m_-^{(p+1)}\Gamma_{-,p}^{-1}\Lambda_{-,p}\right] \tag{7}$$

$$\tilde{\mathbf{V}} = \eta!^2 e_\eta' \left[\tilde{\Gamma}_{+,p}^{-1}\tilde{\Psi}_{+,p}\tilde{\Gamma}_{+,p}^{-1} + \tilde{\Gamma}_{-,p}^{-1}\tilde{\Psi}_{-,p}\tilde{\Gamma}_{-,p}^{-1}\right] e_\eta \tag{8}$$

$$\tilde{\mathbf{B}} = \frac{\eta!}{(p+1)!} e_\eta' \left[m_+^{(p+1)}\tilde{\Gamma}_{+,p}^{-1}\tilde{\Lambda}_{+,p} - (-1)^{p+1+\eta} m_-^{(p+1)}\tilde{\Gamma}_{-,p}^{-1}\tilde{\Lambda}_{-,p}\right] \tag{9}$$

Additionally, if $\mathbf{B_n} \neq 0$, the approximated AMSE-optimal bandwidth is

$$h^{opt} = \left[\frac{2\eta+1}{2(p+1-\eta)} \frac{\tilde{\mathbf{V}}}{\tilde{\mathbf{B}}}\right]^{\frac{1}{2p+3}} G^{-\frac{1}{2p+3}} \tag{10}$$

The optimal bandwidth formula presented follows very closely the format of similar results 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. Similarly, the rate at which the optimal bandwidth shrinks is similar to the one presented in those papers, with the number of clusters replacing the full sample size ($h^{opt} \propto G^{-\frac{1}{2p+3}}$). As pointed out by CCT, this rate has important implications for the control of the leading term of the bias, which we discuss in detail in Section 4.

Note that both the "pre-asymptotic" conditional variance leading component, $\mathbf{V}_n$, and its limit, $\tilde{\mathbf{V}}$, through Ψ_n and $\tilde{\Psi}$, correctly capture the information about the cluster structure and dependence. In Equation (10), the within-cluster data dependence is captured by $\tilde{\mathbf{V}}$ and would be ignored if conventional bandwidth choice procedures for i.i.d. data were implemented. With these traditional procedures, the researcher minimizes an incomplete MSE, and the resulting bandwidth does not correctly assess the trade-off between bias and variance.

In order to implement the optimal bandwidth, we need to replace $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{B}}$ with feasible estimates. This can be done by replacing $\mathbf{V}_n$ and $\mathbf{B}_n$, using a preliminary bandwidth and obtaining estimates of the unknown parameters $m_+^{(p+1)}$, $m_-^{(p+1)}$ and Σ. CCF point out that using standard errors based on fixed-G calculations, as opposed to first-order asymptotic approximations, can achieve higher-order refinements in terms of coverage error. These authors also discuss the related issue of obtaining bandwidths that provide optimal coverage rates. Finally, refer to Calonico et al. (2016a) for details on the implementation of the feasible optimal bandwidth choice using the "pre-asymptotics" formulas described above.[2]

3.1 Clustering at the Running Variable Level

In this section, we analyze the special case in which the clusters are defined at the running variable level. Note first that the usual asymptotic approximations based on shrinking the bandwidth as the number of clusters increase often fail to adequately capture the dependence structure present in the data. Typically, as the number of clusters grows and bandwidth shrinks to zero the covariance terms in the asymptotic variance vanish, and the clustering issue disappears. This result is similar to the situation described by Bhattacharya (2005) in the context of multi-stage sampling. Intuitively, the proportion of units from a given cluster within the bandwidth goes to zero. However, when the cluster is defined at the running variable level and all units in a cluster share the same value of X_g, the clustering issue remains asymptotically.

This setup encompasses multiple interesting applied cases. In particular, as discussed in the introduction, researchers are often faced with situations where either the running variable is discrete or where the running variable is defined at a higher level of aggregation than the

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

unit of observation. In addition, this setup is important as it provides additional intuition on the theoretical and practical challenges of incorporating data clustering into local identification strategies such as RD designs.

Given this setup, we derive the asymptotic properties of $\hat{\tau}^{(\eta)}$.[3] The results emphasize the similarities with the usual asymptotic results for i.i.d. cases described in Porter (2003), IK, and CCT; making clear the intuition about the pitfalls of ignoring data clustering when performing inference or choosing benchmark bandwidths.[4] The asymptotic variance of the estimator is given by the following:[5]

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

This follows directly from Theorem 3.1 and is analogue to the results for i.i.d data in CCT (Lemma A.1, p. 16). This formula makes clear that ignoring dependence in the data misrepresents the variance of the estimators by ignoring the terms in $\sum_{i=1}^{N_g}\sum_{i\neq s}^{N_g}\sigma_{is}^{+}$ for clusters on both sides of the threshold.

As a secondary point, note that the form of the leading term of the asymptotic bias,

$$E[\hat{\tau}^{(\eta)}|X] = \tau^{(\eta)} + h^{p+1-\eta}C_{1,\eta}\left(m_{+}^{(p+1)} - (-1)^{(p+1+\eta)}m_{-}^{(p+1)}\right) + o_p(h^{p+1-\eta})$$

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 not on the form of the asymptotic variance.

The infeasible MSE-optimal bandwidth choice in this case is then given by the following:

$$h_{opt} = \left[C_{\kappa\eta}\frac{\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{+,is}}{GN_g^2f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{-,is}}{GN_g^2f(0)}}{\left[m_{+}^{(p+1)} - (-1)^{(p+1+\eta)}m_{-}^{(p+1)}\right]^2}\right]^{\frac{1}{2p+3}} \tag{11}$$

These results follow very closely those presented for the local linear estimator by IK and CCT (Lemma 1, p.11) for the i.i.d. data case.[6] If the errors are indeed i.i.d., this bandwidth collapses to the traditional MSE-optimal bandwidth choice.

[3] The asymptotic derivations in this section assume that N_g identical across clusters, the the results can be easily extended to the case where cluster size is allowed to vary.

[4] For simplicity all asymptotic results are presented for the case in which clusters have equal sizes, N_g, general case involves cumbersome notation and formulas and can be similarly derived.

[5] See Corollary A.1 in the Appendix for the exact formulas for the terms $C_{1,\eta}$, $C_{2,\eta}$, and $C_{\kappa,\eta}$, which are constants that depend on the kernel function and polynomial order selected by the researcher.

[6] See Corollary A.1 in the Appendix.

For further insight, consider the case of a linear local estimator ($p = 1$) in the standard RD design ($\eta = 0$) with a constant group-level shock, c_g, and Ω taking 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 (11) can be written as follows:

$$h_{opt} = \left(\frac{C_{2,0}}{4C_{1,0}}\right)^{\frac{1}{5}} \left[\frac{(\sigma_{u,+}^2 + N_g\sigma_{c,+}^2) + (\sigma_{u,-}^2 + N_g\sigma_{c,-}^2)}{f(0)\left[\mu_+^{(2)} - \mu_-^{(2)}\right]^2}\right]^{\frac{1}{5}} N^{-1/5} \tag{12}$$

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 based on 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.

In this case where all units in a cluster share the same value for the running variable, an alternative approach to address within-cluster data dependence is to aggregate the data to the cluster level. This approach, introduced by Moulton (1987) and discussed in Angrist and Pischke (2009), relies on the basic intuition that averaging the data to the cluster level treats each cluster as an independent observation. By collapsing the data one can absorb the within-cluster dependence into each of the group level observations. Hence, one could propose to collapse the data and obtain the benchmark bandwidth using a traditional bandwidth choice procedure based on i.i.d. data. In practice, this approach has some pitfalls. First, it forfeits the use of variation in outcomes (or covariates) at a lower level than the cluster. Also, it requires that the cluster-level observations be weighted by their relative cluster sizes. In addition, this aggregation method would not address any across-cluster heteroskedasticity, which could potentially induced by the averaging of different sized groups. Last, and most importantly, this approach would obviously not be valid if the running variable was allowed to change within clusters. With these caveats, careful implementation of a strategy based on aggregating the data to the cluster level could be a valid alternative. However, given the developments in this paper and the ease of implementation of these procedures, researchers might prefer to use microdata while explicitly accounting for the clustering.

4 Distributional Approximation

In order to perform inference about the parameter of interest, we require an approximation to its distribution. While one could obtain an approximation based on the "usual" studentized statistic $\frac{\hat{\tau}^{(\eta)}-\tau^{(\eta)}}{\sqrt{\frac{1}{Gh^{2\eta+1}}\mathbf{V}_n}}$, this statistic converges to a standard normal distribution only if the bandwidth is allowed to shrink fast enough to assure the leading bias term is negligible ($Gh^{2p+3} \to 0$). However, as seen in Section 3, the MSE-optimal bandwidth is proportional to $G^{\frac{1}{2p+3}}$, which leaves a first-order bias in the approximated distribution, as described by CCT for i.i.d. data.

While a common practice in this situation is to "undersmooth" by choosing an arbitrarily smaller bandwidth than the value suggested by h_{opt}, an alternative proposed by CCT and CCF is to directly remove the leading term of the bias, $h^{p+1-\eta}\mathbf{B}_n$. This bias-corrected studentized statistic behavior is adequately approximated by a standard normal distribution if $Gh^{2p+5} \to 0$, which is compatible with the proposed optimal bandwidth choice discussed above and is described below in Theorem 4.1.

Theorem 4.1. *Under Assumptions 1-6, if $Gh \to \infty$ and $Gh^{2p+5} \to 0$. Then,*

$$\frac{\hat{\tau}^{(\eta)} - \tau^{(\eta)} - h^{p+1-\eta}\mathbf{B}_n}{\sqrt{\frac{1}{Gh^{2\eta+1}}\mathbf{V}_n}} \to_d N(0,1). \tag{13}$$

Notice that Theorem 4.1 is based on the infeasible pre-asymptotic formulas for bias and standard errors, $\mathbf{B}_n$ and $\sqrt{\mathbf{V}_n}$. As pointed out by CCT and CCF for the i.i.d. case, these "fixed-G" standard errors are valid and produce statistics that achieve higher-order refinements relative to approximations based on a first-order asymptotic approximation. We expect that similar refinements would benefit the statistics described above when employing "pre-asymptotic" cluster-robust standard errors. In addition, note that this result can also be cast as the cluster analogue of the point made by Imbens and Lemieux (2008) that the usual parametric heteroskedasticity-robust standard errors may be used in traditional RD designs with i.i.d. data.[7]

Additionally, we propose the use of natural, direct plug-in, feasible estimators for $\mathbf{B}_n$ and $\sqrt{\mathbf{V}_n}$ that replace the unknown quantities $m_+^{(p+1)}$, $m_-^{(p+1)}$ and Σ with consistent estimates. For Σ, it is useful to note that, as it is usual in clustering, we are not interested in estimating each

[7]See Section 3.1 of CCF for a careful treatment of this issue.

σ_{gis} term in the relevant variance-covariance matrix and focus on estimating Ψ_p based on the pre-asymptotic formulas $\hat{\Psi}_{+,p} = R_p' W_{+,p} \hat{\epsilon}\hat{\epsilon}' W_{+,p} R_p$ and $\hat{\Psi}_{-,p} = R_p' W_{-,p} \hat{\epsilon}\hat{\epsilon}' W_{-,p} R_p$.

As described in detail by CCT, consistent estimates of the leading bias term can be obtained by plugging in $\hat{m}_+^{(p+1)}(b)$ and $\hat{m}_-^{(p+1)}(b)$ from a $(p+1)$-order local polynomial around the threshold using a potentially different bandwidth, b.[8] This bias correction strategy introduces non-negligible variability to the bias-corrected statistic and requires that the denominator account for the contribution of both $\hat{\tau}^{(\eta)}$ and the bias estimate.

Recently, CCF argue 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, 2016, 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 meaningfully affected by the cluster dependence, as shown in Theorem 3.1. In addition, the standard error correction proposed in CCT follows similarly with the addition of off-diagonal variance terms that account for cluster dependence. The details of both discussions are beyond the scope of this paper, and the reader is referred to CCT and CCF.

Note that 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. For a detailed discussion on different implementations of both the standard errors, analytical bias correction, and optimal bandwidth choice in RD settings including clustered data, see Calonico et al. (2016a).

5 Extensions

A main conclusion of the analysis above is that cluster dependence should be taken into account not only when performing inference, but also when selecting bandwidths. This insight 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 designs 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

[8] For details on the implementation of the bias estimation and correction, see CCT and Calonico et al. (2016a).

naturally want to allow the outcome and treatment status to be potentially correlated within cluster by allowing $Cov(Y, T|X)$ to have non-zero off diagonal elements within clusters.[9]

Second, instead of relying on optimal-MSE bandwidths, CCF 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, 2016, p. 29), they suggest a rule-of-thumb bandwidth choice based on an adjusted MSE-optimal bandwidth, which provides the correct coverage error rate. A straightforward modification of these procedures to account for cluster dependence can be incorporated for researchers wishing to use these procedures.

6 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). Throughout this section, estimation is performed using a local linear estimator, the preferred method in most applications. All simulation results use the implementation software developed by Calonico et al. (2016a) for STATA.[10] For clarity, the setup follows a random effects structure:

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

The simulations we present are based on a conditional mean function in CCT that modifies the DGP based on Lee (2008)'s data to induce bias near the cutoff. This choice highlights the potential trade-off between bias and variance when examining MSE in a clustering context. This setting demonstrates the properties of our cluster-robust bandwidth selection even in a situation where accounting for bias with small bandwidths is particularly important. For both DGPs presented here, $m(x)$ takes the following form:

$$m_1(x) = \begin{cases} 0.48 + 1.27x - 0.5 * 7.18x^2 + 0.7 * 20.21x^3 + 1.1 * 21.54x^4 + 1.5 * 7.33x^5 \;\; if \;\; x < 0 \\ 0.52 + 0.84x - 0.1 * 3.00x^2 - 0.3 * 7.99x^3 - 0.1 * 9.01x^4 + 3.56x^5 \;\; if \;\; x \geq 0. \end{cases} \tag{14}$$

For the first set of simulations, the running variable is constructed so that x is a linear

[9] Note that the fuzzy RD design implementation is available in STATA through the rdrobust package (Calonico et al., 2016a).

[10] We use the software's version updated in April 2016, available for download at `https://sites.google.com/site/rdpackages/rdrobust`.

combination of x_1 and x_2. x_1 is defined at the group level, while x_2 is defined at the individual level. Hence, while clusters span multiple values of the running variable, there is within-cluster correlation in x. Each x_j is drawn from a half normal distribution, $N_f(0, 4)$ and clusters are constrained so that they do not contain units from both sides of the treatment threshold.[11] 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 $\rho \equiv \frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}$. Simulations are run for various values of within cluster dependence, ρ.

Figure 1 presents MSE in this simulation for three procedures: the cluster-robust and traditional bandwidth selection procedures described in Section 3, a traditional bandwidth selection procedure that does not account for clustering, and a cluster-robust bandwidth choice that optimizes coverage error (CER) as discussed in Section 5. Each panel in Figure 1 plots the empirical MSE of each procedure for different values of ρ, where panels are separated by cluster size and number of clusters. Based on the discussion in Section 3, we expect accounting for clustering to become more important as cluster size (N_g) or within-cluster dependence (ρ) increase. The simulations back this intuition. As N_g or ρ increase, the clustering problem becomes more important, emphasizing the need for approximations and procedures capable of capturing the cluster dependence. In addition, note that the cluster-robust CER bandwidth procedure produces higher MSE than the other procedures, which is to be expected given that it is the only procedure not explicitly designed to minimize MSE. Figure 2 plots coverage for each of the procedures for different levels of ρ. Across all panels, coverage is similar between the three procedures. Note that while each procedure picks a difference bandwidth, all procedures are based on robust bias-corrected confidence intervals with inference that accounts for clustering. Therefore, it is potentially unsurprising that the bandwidth choices lead to similar coverage.

The second simulation showcases the properties of our procedure in a setting where there exists clustering at the level of the running variable. Here, $m(x_g)$ is the mean function shared by all individuals in cluster g, c_g is group-level shock with variance σ_c^2, and u_{ig} is idiosyncratic error term with variance σ_u^2. Note that the functional form of $m(.)$ is the same defined above in Equation (14). Both u and c are normally distributed, and as before the variance of u is set to 0.1295^2 with the variance of c being adjusted to obtain the desired value of ρ.

Additionally, in this set of simulations we report results with data aggregated to the running variable level using the traditional bandwidth choice. This procedure averages all observations

[11] This simulation is based in part on Hagemann (2015), and defines $x = (\sqrt{\rho_x})x_1 + (\sqrt{1-\rho_x})x_2$, with $\rho_x = 0.8$. For clusters below the threshold the signs of x_1 and x_2 are negative, and conversely above the cutoff, guaranteeing there is no crossing within clusters.

in a given cluster and performs estimation with the aggregated data, thereby treating each cluster as a single observation. This 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, as described in Section 3.1.

These simulation results are presented in Figure 3. As expected, higher levels of within-cluster dependence, ρ, lead to situations where the cluster-robust approach dominates procedures using traditional bandwidth selection algorithms in terms of empirical MSE. Moreover, as the size of clusters or rho increase, the cluster-robust procedure far outperforms traditional bandwidth choices using the microdata. In particular, both columns show that the cluster-robust and traditional procedures perform similarly well for small values of ρ, and the former is significantly better as ρ increases, when accounting for clustering becomes more important with larger dependence. As before, the CER bandwidth choice leads to larger empirical MSE across all panels. Last, the *ad hoc* procedure using aggregate data performs quite similar to the cluster-robust procedure in terms of MSE, indicating that this procedure may also be able to account for clustering in RD designs in the special case where clusters are defined at the level of the running variable. Figure 4 presents coverage across these four procedures. While there are some slight differences in coverage between procedures when $G = 250$, the four procedures all produce very similar coverage when $G = 1000$, indicating that accounting for clustering when selecting a bandwidth does little to change coverage.

7 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.[12] As in Baum-Snow and Marion (2009), we restrict to census tracts in metropolitan areas, and exclude Alaska and Hawaii. Table 1 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 2 displays results of four 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 and are weighted to account for differential cluster sizes. 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. The third 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. Last, the fourth column investigates whether including covariates has an effect on the estimates. To do this, we include census region fixed effects as covariates using a cluster-robust procedure based on Calonico et al. (2016b). For each regression, we present coefficient estimates, standard errors in parentheses, robust bias-corrected confidence intervals in brackets,

[12]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`.

effective sample size, and local polynomial bandwidth choice.

The results show that accounting for potential dependence in outcomes within a census tract can substantially change the benchmark optimal-MSE bandwidth. As argued in Sections 3.1 and 6, 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 four 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. Last, note that the census region fixed effects do little to change the estimates, confidence intervals, or bandwidth choice. This result suggests that sorting of census tracts around the cutoff on the basis of census tract is not an issue for the empirical analysis.

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. On the whole, the cluster-robust, aggregated data procedures, and cluster-robust procedure with region fixed effects all provide similar results, and give a different empirical perspective than simply applying the IK bandwidth selection algorithm to the microdata.

8 Conclusion

Even though many researchers utilizing RD designs 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 procedure.

To fill this gap in the literature, we derive the distributional properties of local polynomial estimators in RD designs under a fixed-G approximation and provide accompanying optimal bandwidth selection rules. In addition, we analyze the special case of data clustered at the running variable level, providing further insight into our results. Both sets of results lead to the same implementation of cluster-robust optimal bandwidth selectors, which extend the prior results in IK and CCT to account for clustering. The insights provided by this analysis can also be applied directly to fuzzy and kink RD designs, as well as the robust bias-corrected methods in CCT.

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 fail to capture adequately the bias-variance trade-off. 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.

Angrist, Joshua and Jörn–Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists' Companion.* Princeton, NJ: Princeton Univerisity Press.

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 Max H. Farrell. 2016. "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference." Working Paper, University of Michigan.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocio Titiunik. 2016a. "rdrobust: Software for Regression Discontinuity Designs." Working Paper, University of Michigan.

———. 2016b. "Regression Discontinuity Designs using Covariates." Working Paper, University of Michigan.

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

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.

Hagemann, Andreas. 2015. "Cluster-Robust Bootstrap Inference in Quantile Regression Models." *Journal of the American Statistical Association* 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.

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.

Moulton, Brent R. 1987. "Diagnostics for Group Effects in Regression Analysis." *Journal of Business & Economic Statistics* 5 (2):275–282.

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

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, M.A.: MIT Press, 2nd ed.

Table 1: 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.

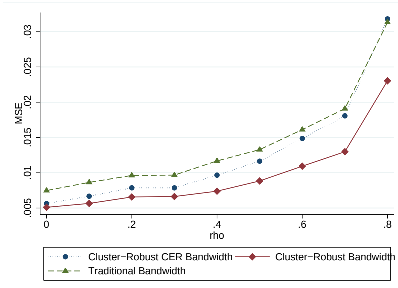
Table 2: Local Linear Estimates of the Effect of QCT Status

Dependent Variable	Cluster-Robust Bandwidth	Traditional Bandwidth	Tract-Level	Cluster-Robust Bandwidth Region FE
Homeownership	-0.0089	-0.0335	-0.0061	-0.0115
	(0.0130)	(0.0217)	(0.0094)	(0.0117)
	[-0.0402, 0.0186]	[-0.0789, 0.0099]	[-0.0281, 0.0162]	[-0.0406, .0121]
	4,537,956	1,179,593	8,761	4,656,057
	w=0.098	w=0.027	w=0.107	w=0.100
Fraction Non-White	0.0024	-0.0048	0.0042	0.0024
	(0.0173)	(0.0386)	(0.0153)	(0.0167)
	[-0.0429, 0.0311]	[-0.0825, 0.0707]	[-0.0361, 0.0320]	[-0.0407, 0.0303]
	5,022,420	945,413	8,600	4,936,642
	w=0.107	w=0.022	w=0.105	w=0.106
High School Diploma or Higher	-0.0073	-0.0006	-0.0105*	-0.0076
	(0.0078)	(0.0123)	(0.0061)	(0.0074)
	[-0.0217, 0.0129]	[-0.0248, 0.0256]	[-0.0243, 0.0043]	[-0.0211, 0.0115]
	3,828,193	1,304,465	10,109	3,941,003
	w=0.129	w=0.052	w=0.121	w=0.131
Bachelors Degree or Higher	0.0039	0.0041	0.0024	0.0039
	(0.0057)	(0.0077)	(0.0064)	(.0058)
	[-0.0098, 0.0177]	[-0.0133, 0.0222]	[-0.0123, 0.0178]	[-0.0096, 0.0181]
	3,285,861	1,747,085	8,336	3,207,979
	w=0.115	w=0.067	w=0.103	w=0.112
Employment Rate	0.0081	0.0218**	-0.0028	0.0051
	(0.0054)	(0.0093)	(0.0045)	(0.0051)
	[-0.0028, 0.0221]	[0.0040, 0.0421]	[-0.014, 0.0075]	[-0.0057, 0.0181]
	2,602,616	811,030	10,104	2,839,725
	w=0.082	w=0.027	w=0.121	w=0.089
Number of LIHTC Units	7.959***	12.729***	4.628**	7.808***
	(2.956)	(5.809)	(2.180)	(2.975)
	[1.543, 14.884]	[1.177, 24.512]	[-0.464, 9.856]	[1.357, 14.770]
	4,618,876	778,252	8,338	4,523,254
	w=0.099	w=0.018	w=0.103	w=0.098
Number of LIHTC Projects	0.0341	-0.0473	0.0662*	0.0351
	(0.0581)	(0.1020)	(0.0258)	(0.0580)
	[-0.0995, 0.1475]	[-0.2484, 0.1579]	[0.0020, 0.1204]	[-0.0985, 0.1482]
	2,961,625	234,210	10,222	2,974,891
	w=0.068	w=0.006	w=0.122	w=0.068
N	30,330,540	30,330,540	45,294	30,330,540
N Clusters	44,716	44,716	45,294	44,716

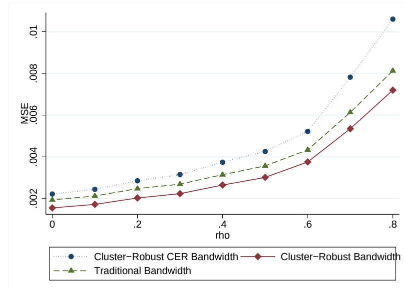
Source: Microdata and tract-level data from the long form of the 2000 decennial census. Standard errors in parentheses are adjusted for clustering at the tract level. Robust bias-corrected confidence intervals following CCT are shown in brackets just above effective sample size used in the local polynomial regression. "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: MSE Performance in Simulated Data – Data Generating Process 1

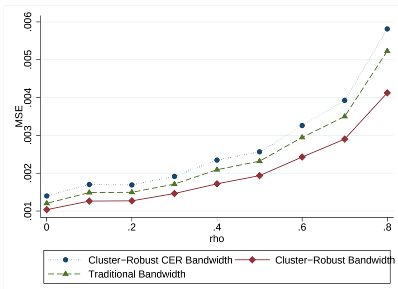
(a) Size = 5, Number of Clusters = 250



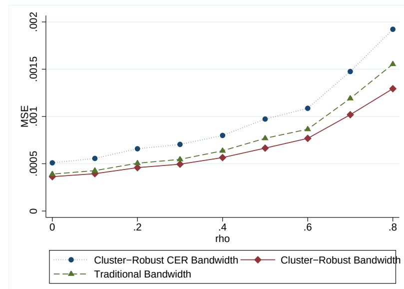
(b) Size = 5, Number of Clusters = 1000



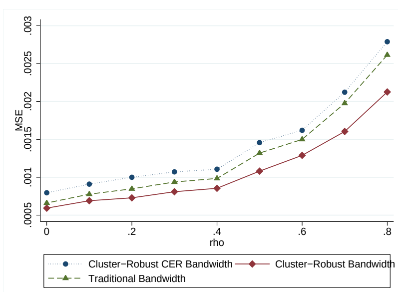
(c) Size = 25, Number of Clusters = 250



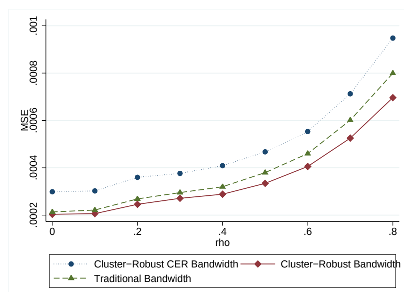
(d) Size = 25, Number of Clusters = 1000



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



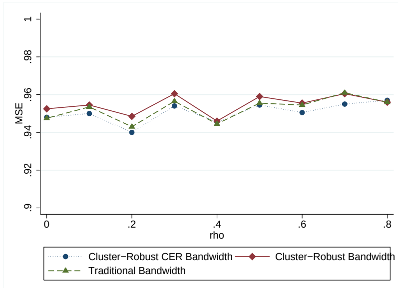
(f) Size = 50, Number of Clusters = 1000



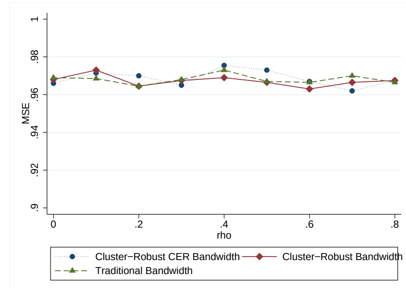
Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.

Figure 2: Coverage Performance in Simulated Data – Data Generating Process 1

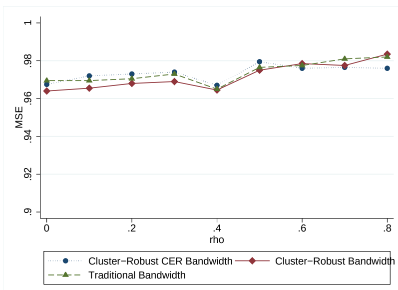
(a) Size = 5, Number of Clusters = 250



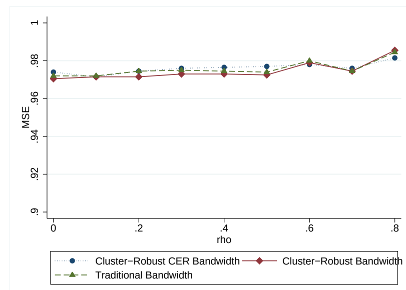
(b) Size = 5, Number of Clusters = 1000



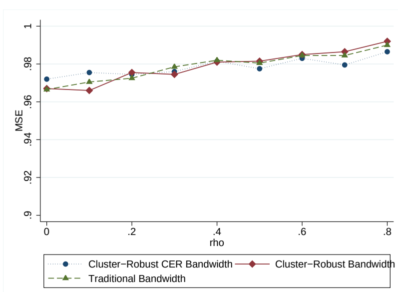
(c) Size = 25, Number of Clusters = 250



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



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



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

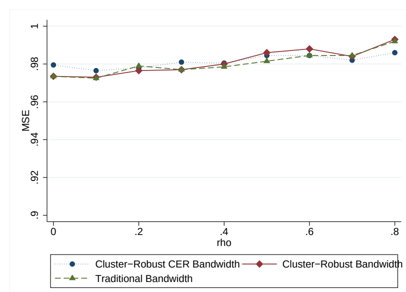
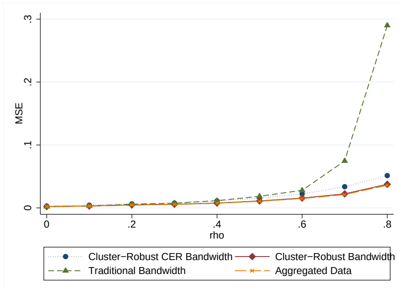
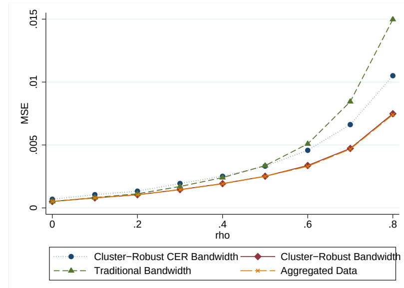


Figure 3: MSE Performance in Simulated Data – Data Generating Process 2

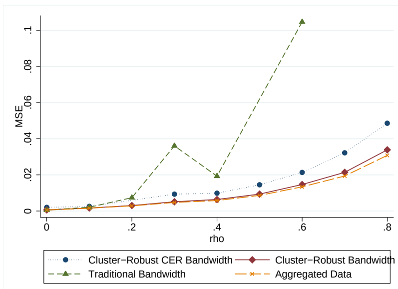
(a) Size = 5, Number of Clusters = 250



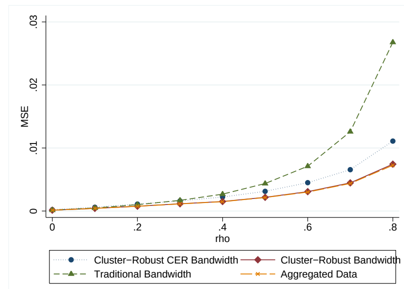
(b) Size = 5, Number of Clusters = 1000



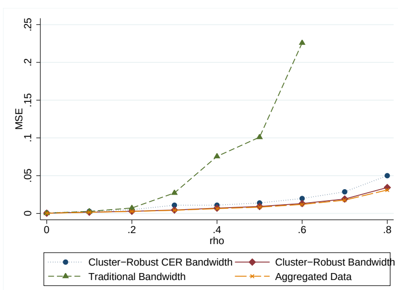
(c) Size = 25, Number of Clusters = 250



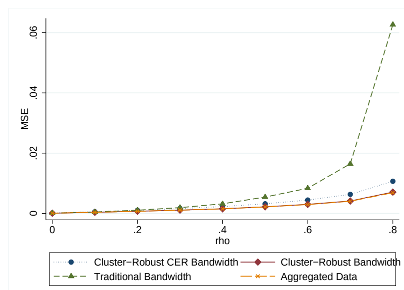
(d) Size = 25, Number of Clusters = 1000



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



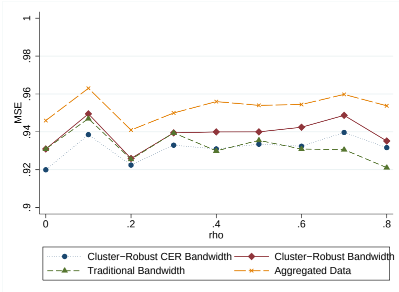
(f) Size = 50, Number of Clusters = 1000



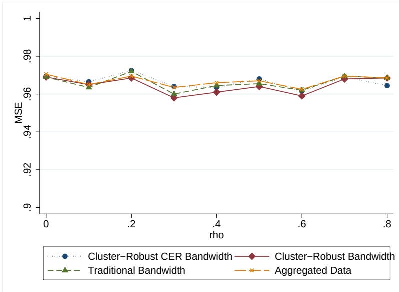
Note: Results are not plotted if the MSE in the traditional bandwidth procedure is more than 25 times the cluster-robust procedure.

Figure 4: Coverage Performance in Simulated Data – Data Generating Process 2

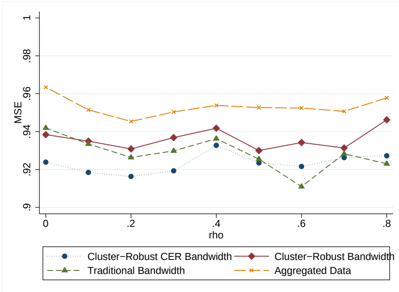
(a) Size = 5, Number of Clusters = 250



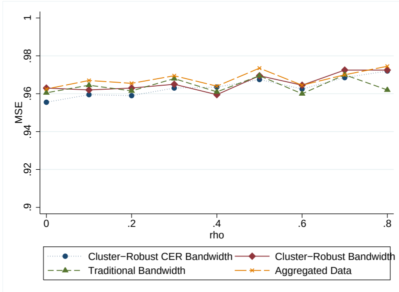
(b) Size = 5, Number of Clusters = 1000



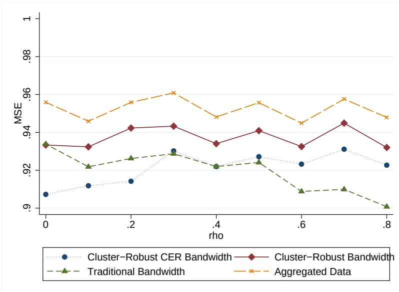
(c) Size = 25, Number of Clusters = 250



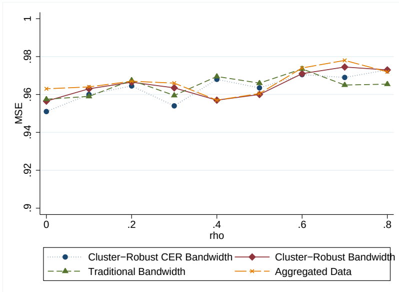
(d) Size = 25, Number of Clusters = 1000



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



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



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 $g = 1, ..., G$.
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-covariance matrix, Σ, is block diagonal with typical block given by $\Omega_g(x) = Var(Y_g|X)$, 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 $\kappa > 0$.
6. Individuals on the same cluster are assumed to be on the same side of the cutoff.

A.2 Results

For analyzing the properties of the estimator $\hat{\tau}^{(\eta)} = \hat{m}_+^{(\eta)} - \hat{m}_-^{(\eta)}$, note that $y_{ig} = m(x_{ig}) + \epsilon_{ig}$. Let $x_{h,ig} = \frac{x_{ig} - \bar{x}}{h}$ where $\bar{x}$ is the treatment cutoff which is set to zero without loss of generality. Also, let $r_p(u) = (1, u, u^2, \ldots, u^p)'$ and $R_p = [r_p(x_{h,11}), r_p(x_{h,21}), \cdots, r_p(x_{h,N_G G})]'$ and e_η' be a vector of zeros except for the $(\eta+1)^{th}$ entry equal to one, e.g., $e_0' = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$. Define the weighting matrix $W_{+,p,g} = diag(h^{-1}\mathbf{1}[x_{1g} > 0]K(x_{h,1g}), h^{-1}\mathbf{1}[x_{2g} > 0]K(x_{h,2g}), \ldots, h^{-1}\mathbf{1}[x_{N_g g} > 0]K(x_{h,N_g g}))$ and $W_{-,p,g} = diag(h^{-1}\mathbf{1}[x_{1g} \leq 0]K(x_{h,1g}), h^{-1}\mathbf{1}[x_{2g} \leq 0]K(x_{h,2g}), \ldots, h^{-1}\mathbf{1}[x_{N_g g} \leq 0]K(x_{h,N_g g}))$. Then, let $W_{+,p}$ and $W_{-,p}$ be the block diagonal matrices with g-th block $W_{+,p,g}$ and $W_{-,p,g}$ for $g = 1, \ldots, G$, respectively. Also, define $H_p = diag(1, h^{-1}, \ldots, h^{-p})$. Finally, let $\Gamma_{+,p} = R_p' W_{+,p} R_p$, $\Gamma_{-,p} = R_p' W_{-,p} R_p$, $\Lambda_{+,p} = R_p' W_{+,p} \left[x_{h,11}^{p+1}, \ldots, x_{h,N_G G}^{p+1}\right]'$ and $\Lambda_{-,p} = R_p' W_{-,p} \left[x_{h,11}^{p+1}, \ldots, x_{h,N_G G}^{p+1}\right]'$. Then,

$$\hat{m}_+^{(\eta)} = \eta! \hat{\beta}_p = \eta! e_\eta' H_p \Gamma_{+,p}^{-1} R_p' W_{+,p} Y \tag{15}$$

And analogously for $\hat{m}_-^{(\eta)}$. Also define $\Psi_{+,p} = R_p' W_{+,p} \Sigma W_{+,p} R_p$ and $\Psi_{-,p} = R_p' W_{-,p} \Sigma W_{-,p} R_p$. It is useful to to establish notation for the (i, s) elements of the variance matrix Ω_g as σ_{gis} and its limits around the threshold $\sigma_{+,is} = lim_{x\to 0^+}\sigma_{gis}$ and $\sigma_{-,is} = lim_{x\to 0^-}\sigma_{gis}$ for describing the cluster at the running variable case.

Moreover, some useful approximations can be obtained by using the fixed-G expectation over the support of X to some of the objects defined above. Following the notation on CCF, let $\tilde{\Lambda}_{+,p} = \mathbb{E}[\Lambda_{+,p}]$, and similarly for $\Lambda_{-,p}$, $\Gamma_{+,p}$, $\Gamma_{-,p}$, $\Psi_{+,p}$, and $\Psi_{-,p}$ (See Lemma A.1 and Lemma A.2).

Theorem 3.1. *Under assumptions 1-6, the conditional Mean Squared Error for $\hat{\tau}^{(\eta)}$ has fixed-G approximation given by,*

$$E[(\hat{\tau}^{(\eta)} - \tau^{(\eta)})^2|x_{11},\dots,x_{N_G G}] = \frac{1}{Gh^{2\eta+1}}\left[\mathbf{V_n}\right] + h^{2(p+1-\eta)}\left[\mathbf{B_n}^2 + o_p(1)\right] \tag{16}$$

$$= \frac{1}{Gh^{2\eta+1}}\left[\tilde{\mathbf{V}} + o_p(1)\right] + h^{2(p+1-\eta)}\left[\tilde{\mathbf{B}}^2 + o_p(1)\right] \tag{17}$$

$$\mathbf{V_n} = \eta!^2 e'_\eta\left[\Gamma^{-1}_{+,p}\Psi_{+,p}\Gamma^{-1}_{+,p} + \Gamma^{-1}_{-,p}\Psi_{-,p}\Gamma^{-1}_{-,p}\right]e_\eta \tag{18}$$

$$\mathbf{B_n} = \frac{\eta!}{(p+1)!}e'_\eta\left[m_+^{(p+1)}\Gamma^{-1}_{+,p}\Lambda_{+,p} - m_-^{(p+1)}\Gamma^{-1}_{-,p}\Lambda_{-,p}\right] \tag{19}$$

$$\textit{and } \tilde{\mathbf{V}} = \eta!^2 e'_\eta\left[\tilde{\Gamma}^{-1}_{+,p}\tilde{\Psi}_{+,p}\tilde{\Gamma}^{-1}_{+,p} + \tilde{\Gamma}^{-1}_{-,p}\tilde{\Psi}_{-,p}\tilde{\Gamma}^{-1}_{-,p}\right]e_\eta \tag{20}$$

$$\tilde{\mathbf{B}} = \frac{\eta!}{(p+1)!}e'_\eta\left[m_+^{(p+1)}\tilde{\Gamma}^{-1}_{+,p}\tilde{\Lambda}_{+,p} - (-1)^{p+1+\eta}m_-^{(p+1)}\tilde{\Gamma}^{-1}_{-,p}\tilde{\Lambda}_{-,p}\right] \tag{21}$$

Additionally, if $\mathbf{B_n} \neq 0$, the approximated MSE-optimal bandwidth is

$$h^{opt} = \left[\frac{2\eta+1}{2(p+1-\eta)}\frac{\tilde{\mathbf{V}}}{\tilde{\mathbf{B}}}\right]^{\frac{1}{2p+3}} G^{-\frac{1}{2p+3}} \tag{22}$$

Proof. Start by rewriting the estimator as,

$$\hat{m}_+^{(\eta)} = \eta!\hat{\beta}_p = \eta! e'_\eta H_p\Gamma_p^{-1}R'_pW_pY = \eta! e'_\eta H_p\Gamma_p^{-1}R'_pW_p\left[m(x)+\epsilon\right] \tag{23}$$

$$= \eta! e'_\eta H_p\Gamma_p^{-1}R'_pW_pm(x) + \eta! e'_\eta H_p\Gamma_p^{-1}R'_pW_p\epsilon \tag{24}$$

For the conditional, fixed-G representation of the bias we proceed as follows. Define $M = (m(x_{11}),\dots,m(x_{N_G G}))'$.

$$\mathbf{E}[\hat{m}_+^{(\eta)}|x_{11},\dots,x_{N_G G}] = \eta! e'_\eta H_p\Gamma_p^{-1}R'_pW_pM \tag{25}$$

Taking a Taylor expansion of $m(\cdot)$ around $\bar{x}=0$, we can rewrite:

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

Where $|T_{ig}| \leq sup_x|m^{(p+2)}(x)x_{h,ig}^{p+2}|$. Then,

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

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

From Lemma A.1 we know that $\eta! e_\eta' H_p \Gamma_p^{-1} R_p' W_p T = o_p(h^{p+1-\eta})$. Then, the conditional bias is given by

$$\mathbb{E}[\hat{m}_+^{(\eta)}|x_{11},\ldots,x_{N_G G}] - m_+^{(\eta)} = h^{p+1-\eta} m_+^{(p+1)} \frac{\eta!}{(p+1)!} e_\eta' \Gamma_{+,p}^{-1}\Lambda_{+,p} + o_p(h^{p+1-\eta})$$

Note that this form of the bias is similar to the one derived by CCT and CCF, and the clustered nature of the data does not affect the basic structure of the bias term as one would expect. Similarly for the conditional variance term, still under a fixed-G approximation:

$$\mathbb{V}[\hat{m}_+^{(\eta)}|x_{11},\ldots,x_{N_G G}] = \frac{1}{G}\eta!^2 e_\eta' H_p \Gamma_{+,p}^{-1} R_p' W_{+,p}\Sigma W_{+,p} R_p \Gamma_{+,p}^{-1} H_p e_\eta \tag{26}$$

$$= \frac{1}{Gh^{2\eta+1}}\eta!^2 e_\eta' \Gamma_{+,p}^{-1} R_p' W_{+,p}\Sigma W_{+,p} R_p \Gamma_{+,p}^{-1} e_\eta \tag{27}$$

And similarly for $\mathbb{E}[\hat{m}_-^{(\eta)}|x_{11},\ldots,x_{N_G G}] - m_-^{(\eta)}$ and $\mathbb{V}[\hat{m}_-^{(\eta)}|x_{11},\ldots,x_{N_G G}]$. Then, the conditional MSE can be described as:

$$\begin{aligned}
E[(\hat{\tau}^{(\eta)} - \tau^{(\eta)})^2|x_{11},\ldots,x_{N_G G}] &= \mathbb{V}[\hat{\tau}^{(\eta)} - \tau^{(\eta)}|x_{11},\ldots,x_{N_G G}] + \{E[\hat{\tau}^{(\eta)} - \tau^{(\eta)}|x_{11},\ldots,x_{N_G G}]\}^2 \\
&= \mathbb{V}[\hat{m}_+^{(\eta)} - m_+^{(\eta)}|x_{11},\ldots,x_{N_G G}] + \mathbb{V}[\hat{m}_-^{(\eta)} - m_-^{(\eta)}|x_{11},\ldots,x_{N_G G}] \\
&+ \{\mathbb{E}[\hat{m}_+^{(\eta)} - m_+^{(\eta)}|x_{11},\ldots,x_{N_G G}] + \mathbb{E}[\hat{m}_-^{(\eta)} - m_-^{(\eta)}|x_{11},\ldots,x_{N_G G}]\}^2 \\
&= \frac{1}{Gh^{2\eta+1}}\left[\mathbf{V_n}\right] + h^{2(p+1-\eta)}\left[\mathbf{B_n}^2 + o_p(1)\right]
\end{aligned}$$

Where,

$$\mathbf{V_n} = \eta!^2 e'_\eta \left[\Gamma^{-1}_{+,p} R'_p W_{+,p} \Sigma W_{+,p} R_p \Gamma^{-1}_{+,p} + \Gamma^{-1}_{-,p} R'_p W_{-,p} \Sigma W_{-,p} R_p \Gamma^{-1}_{-,p}\right] e_\eta \tag{28}$$

$$\mathbf{B_n} = \frac{\eta!}{(p+1)!} e'_\eta \left[m_+^{(p+1)} \Gamma^{-1}_{+,p} \Lambda_{+,p} + m_-^{(p+1)} \Gamma^{-1}_{-,p} \Lambda_{-,p}\right] \tag{29}$$

If we replace the conditional expectations by their nonrandom fixed-G expectations, we obtain:

$$E[(\hat{\tau}^{(\eta)} - \tau^{(\eta)})^2 | x_{11}, \ldots, x_{N_G G}] = \frac{1}{Gh^{2\eta+1}} \left[\tilde{\mathbf{V}} + o_p(1)\right] + h^{2(p+1-\eta)} \left[\tilde{\mathbf{B}}^2 + o_p(1)\right] \tag{30}$$

$$\tilde{\mathbf{V}} = \eta!^2 e'_\eta \left[\tilde{\Gamma}^{-1}_{+,p} \tilde{\Psi}_{+,p} \tilde{\Gamma}^{-1}_{+,p} + \tilde{\Gamma}^{-1}_{-,p} \tilde{\Psi}_{-,p} \tilde{\Gamma}^{-1}_{-,p}\right] e_\eta \tag{31}$$

$$\tilde{\mathbf{B}} = \frac{\eta!}{(p+1)!} e'_\eta \left[m_+^{(p+1)} \tilde{\Gamma}^{-1}_{+,p} \tilde{\Lambda}_{+,p} - (-1)^{p+1+\eta} m_-^{(p+1)} \tilde{\Gamma}^{-1}_{-,p} \tilde{\Lambda}_{-,p}\right] \tag{32}$$

To obtain the formula for the conditional optimal bandwidth, minimize the conditional MSE expression with respect to h conditional on $\mathbf{B}_n$ and $\mathbf{V}_n$. The calculations are exactly the same as presented by CCT Lemma 3.2. □

Theorem 4.1. *Under Assumptions 1-6, if $Gh \to \infty$ and $Gh^{2p+5} \to 0$. Then,*

$$\frac{\hat{\tau}^{(\eta)} - \tau^{(\eta)} - h^{p+1-\eta} \mathbf{B}_n}{\sqrt{\frac{1}{Gh^{2\eta+1}} \mathbf{V}_n}} \to_d N(0,1) \tag{33}$$

Proof.

$$\frac{\hat{m}_+^{(\eta)} - m_+^{(\eta)} - h^{p+1-\eta} B_{+,n}}{\sqrt{\frac{1}{Gh^{2\eta+1}} V_{+,n}}} = \frac{\hat{m}_+^{(\eta)} - E\left[\hat{m}_+^{(\eta)} | X\right] + E\left[\hat{m}_+^{(\eta)} | X\right] - m_+^{(\eta)} - h^{p+1-\eta} B_{+,n}}{\sqrt{\frac{1}{Gh^{2\eta+1}} V_n}}$$

$$= \varepsilon_{1,n} + \varepsilon_{2,n} = \varepsilon_{1,n} + o_p(1)$$

Where $V_{+,n} = \eta!^2 e'_\eta \Gamma^{-1}_{+,p} R'_p W_{+,p} \Sigma W_{+,p} R_p \Gamma^{-1}_{+,p}$ and $B_{+,n} = \frac{\eta!}{(p+1)!} m_+^{(p+1)} e'_\eta \Gamma^{-1}_{+,p} \Lambda_{+,p}$. Since,

$$\varepsilon_{2,n} = \frac{E\left[\hat{m}^{(\eta)} | X\right] - m^{(\eta)} - h^{p+1-\eta} B_{+,n}}{\sqrt{\frac{1}{Gh^{2\eta+1}} V_{+,n}}} = O_p\left(\sqrt{Gh^{5+2p}}\right) = o_p(1).$$

Then,

$$\varepsilon_{1,n} = \left(Var[\hat{m}_+^{(\eta)} - m_+^{(\eta)}|X]\right)^{-\frac{1}{2}} \left(\hat{m}_+^{(\eta)} - E\left[\hat{m}_+^{(\eta)}|X\right]\right) = \left(\frac{1}{Gh^{2\eta+1}}V_{+,n}\right)^{-\frac{1}{2}} \left(\frac{\eta! e_\eta' H_p \Gamma_{+,p}^{-1} R_p' W_{+,p}\epsilon}{G}\right)$$

and, following the same arguments as CCT, $\varepsilon_1 = \tilde{\varepsilon}_{1,n} + o_p(1)$, where $\tilde{\varepsilon}_{1,n} = \sum_{g=1}^{G} \omega_g' \epsilon_g$ where $\epsilon_g = [\epsilon_{g1}, \ldots, \epsilon_{gN_g}]$ for all $g = 1, \ldots, G$. Let $R_{p,g} = \left[r_p(x_{h,1g}), r_p(x_{h,2g}), \cdots, r_p(x_{h,N_g g})\right]'$ and

$$\omega_g = \left(\frac{1}{Gh^{2\eta+1}}V_n\right)^{-\frac{1}{2}} \frac{\eta! e_\eta' H_p \Gamma_{+,p}^{-1} R_{p,g}' W_{+,p,g}}{G}$$

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

The following Corollary presents the asymptotic results for the special case in which the clusters are defined at the running variable level.

Corollary A.1. *(Clustering at the Running Variable) Suppose assumptions 1-6 hold and the clusters are defined at the running variable level, formally,* $X_{ig} = X_g \forall i = 1, \ldots, G$. *Also, let* $Gh \to \infty$.

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

$$E[\hat{\tau}^{(\eta)}|X] = \tau^{(\eta)} + h^{p+1-\eta} C_{1,\eta} \left(m_+^{(p+1)} - (-1)^{(p+1+\eta)} m_-^{(p+1)}\right) + o_p(h^{p+1-\eta})$$

   with $C_{1,\eta} \frac{\eta!}{(p+1)!} e_\eta' \Gamma_p^{-1} \Lambda_p$.

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

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

   with $C_{2,\eta} = \eta!^2 e_\eta' \Gamma_p^{-1} \Psi_p \Gamma_p^{-1} e_\eta$.

3. ***(MSE)***

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

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

$$h_{opt} = \left[C_{\kappa\eta} \frac{\frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{+,is}}{GN_g^2 f(0)} + \frac{\sum_{i=1}^{N_g}\sum_{s=1}^{N_g}\sigma_{-,is}}{GN_g^2 f(0)}}{\left[m_+^{(p+1)} - (-1)^{(p+1+\eta)} m_-^{(p+1)} \right]^2} \right]^{\frac{1}{2p+3}} \tag{34}$$

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

Proof. The proof follows directly from Theorem 3.1 by replacing $\Gamma_{+,p}$, $\Gamma_{-,p}$, $\Lambda_{+,p}$, $\Lambda_{-,p}$, $\Psi_{+,p}$ and $\Psi_{-,p}$ by their asymptotic approximations obtained in Lemma A.1 and Lemma A.2. □

A.3 Auxiliary Lemmas

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 is equivalent to Lemma SA1 on CCT's Technical Supplement.

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_p and Ψ_p as $(p+1)\times(p+1)$ matrices with element (i,j) given by ν_{i+j-2} and π_{i+j-2}, respectively. Finally, Λ_p is the column vector with typical element $(j,1)$ given by ν_{j+p+1}.

Lemma A.1. *Under assumptions 1-6 above and $Gh \to \infty$.*

$$G^{-1}\Lambda_{+,p} = \tilde{\Lambda}_{+,p} + o_p(1) \text{ where, } \tilde{\Lambda}_{+,p} = N_g \int_0^\infty K(u) r_p(u) u^{p+1} f(uh) du$$
$$G^{-1}\Lambda_{-,p} = (-1)^{p+1} H_p(-1)\tilde{\Lambda}_{-,p} + o_p(1) \text{ where, } \tilde{\Lambda}_{-,p} = N_g \int_0^\infty K(u) r_p(u) u^{p+1} f(-uh) du$$
$$G^{-1}\Gamma_{+,p} = \tilde{\Gamma}_{+,p} + o_p(1) \text{ where, } \tilde{\Gamma}_{+,p} = N_g \int_0^\infty K(u) r_p(u) r_p(u)' f(uh) du$$
$$G^{-1}\Gamma_{-,p} = H_p(-1)\tilde{\Gamma}_{-,p} H_p(-1) + o_p(1) \text{ where, } \tilde{\Gamma}_{-,p} = N_g \int_0^\infty K(u) r_p(u) r_p(u)' f(-uh) du$$

Where $H_p(-1) = diag(1, (-1)^{-1}, \ldots, (-1)^{-p})$. *Also, if* $h \to 0$.

$$\tilde{\Lambda}_{+,p} = N_g f(0)\Lambda_p + o(1)$$
$$\tilde{\Lambda}_{-,p} = N_g f(0)\Lambda_p + o(1)$$
$$\tilde{\Gamma}_{+,p} = N_g f(0)\Gamma_p + o(1)$$
$$\tilde{\Gamma}_{-,p} = N_g f(0)\Gamma_p + o(1)$$

Proof. To obtain the results, note that $\Lambda_{+,p}$, $\Lambda_{-,p}$, $\Gamma_{+,p}$ and $\Gamma_{-,p}$ can be written as functions of $F_{+,j}$ and $F_{-,j}$ with $F_{+,j} = \frac{1}{Gh}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{1}[X_{ig} > 0]K(X_{h,ig})X_{h,ig}^j = \frac{1}{G}\sum_{g=1}^{G} N_g \frac{1}{N_g h}\sum_{i=1}^{N_g}\mathbf{1}[X_{ig} > 0]K(X_{h,ig})X_{h,ig}^j = \frac{1}{G}\sum_{g=1}^{G} N_g A_{+,jg}$, where $A_{+,jg} = \frac{1}{N_g h}\sum_{i=1}^{N_g}\mathbf{1}[X_{ig} > 0]K(X_{h,ig})X_{h,ig}^j$. And similarly for $F_{-,j}$ with $\mathbf{1}[X_{ig} \leq 0]$ replacing $\mathbf{1}[X_{ig} > 0]$. If N_g is equal for all G clusters, then $F_{+,j} = \frac{1}{G}\sum_{g=1}^{G} A_{jg}$. Under Assumptions 1-6, (i) for non-negative integer j

$$\begin{aligned} E\left[A_{+,jg}\right] &= E\left[\frac{1}{N_g h}\sum_{i=1}^{N_g}\mathbf{1}[X_{ig} > 0]K(X_{h,ig})X_{h,ig}^j\right] = h^{-1}\int_0^{\infty} K\left(\frac{x}{h}\right)\left(\frac{x}{h}\right)^j f(x)dx \\ &= \int_0^{\infty} K\left(u\right) u^j f(uh)du = \tilde{A}_{jg} \end{aligned}$$

and $\tilde{F}_{+,j} = N_g \int_0^{\infty} K\left(u\right) u^j f(uh)du$. When $h \to 0$,

$$\tilde{A}_{+,jg} = f(0)v_j + O(1)$$

Then,

$$\tilde{F}_{+,j} = E[F_{+,j}] = \frac{1}{G}\sum_{g=1}^{G} N_g \tilde{A}_{+,jg} = N_g f(0) v_j + O(1)$$

For the variance,

$$\begin{aligned} Var\left[A_{+,jg}\right] &= E\left[A_{+,jg}^2\right] - E\left[A_{+,jg}\right]^2 \\ &\leq \frac{1}{N_g^2 h^2} E\left[\sum_{i=1}^{N_g}\mathbf{1}[X_{ig} > 0]K^2(X_{h,ig})X_{h,ig}^{2j}\right] = \frac{1}{N_g h}\int_0^{\infty} K^2(u)u^{2j} f(uh)du = O\left(h^{-1}\right) \end{aligned}$$

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

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

Then,

$$F_{+,j} = \tilde{F}_{+,j} + o_p(1)$$

Noting that $G^{-1}\Gamma_{+,p}$ and $G^{-1}\Gamma_{-,p}$ have typical element (i,j) given by $F_{+,i+j-2}$ and $F_{-,i+j-2}$, respectively gives the result. Similarly, $G^{-1}\Lambda_{+,p}$ and $G^{-1}\Lambda_{-,p}$ are column vectors with typical element $(j,1)$ given by $F_{+,p+1+j}$ and $F_{-,p+1+j}$, respectively. □

The following discussion concerns the variance terms which are altered to allow for clustering.

Lemma A.2. *Under assumptions 1-6 above and $Gh \to \infty$.*

$$G^{-1}\Psi_{+,p} = \tilde{\Psi}_{+,p} + o_p(1) \text{ where, } \tilde{\Psi}_{+,p} = \int_0^\infty \int_0^\infty K(u)K(w)r_p(u)\Sigma r_p(w)' f(uh, wh) du dw$$
$$G^{-1}\Psi_{-,p} = H_p(-1)\tilde{\Psi}_{-,p}H_p(-1) + o_p(1) \text{ where, } \tilde{\Psi}_{-,p} = \int_0^\infty \int_0^\infty K(u)K(w)r_p(u)\Sigma r_p(w)' f(-uh, -wh) du dw$$

If the cluster is defined at the tunning variable level, i.e., $X_{ig} = X_g$ for all $i = 1, \ldots, N_g$, and $h \to 0$. Then,

$$\tilde{\Psi}_{+,p} = f(0)\Psi_p \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{+,is} + o(1)$$
$$\tilde{\Psi}_{-,p} = f(0)\Psi_p \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \sigma_{-,is} + o(1)$$

Proof. Start by noticing that the typical element (t,j) in matrix $G^{-1}\Psi_{+,p}$ is given by $Q_{+,tj} = \frac{1}{Gh^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \mathbf{1}[x_{ig} > 0]\mathbf{1}[x_{sg} > 0]K(x_{h,ig})K(x_{h,sg})x_{h,ig}^t x_{h,sg}^j \sigma_{gis}$, where σ_{gtj} is the term in the $t-th$ line and $j-th$ column in Ω_g. Then,

$$\begin{aligned}
E\left[Q_{+,tj}\right] &= E\left[\frac{1}{Gh^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{s=1}^{N_g} \mathbf{1}[x_{ig} > 0]\mathbf{1}[x_{sg} > 0]K(x_{h,ig})K(x_{h,sg})x_{h,ig}^t x_{h,sg}^j \sigma_{gis}\right] \\
&= \frac{1}{h^2}\int_0^\infty \int_0^\infty \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(x_{h,ig})K(x_{h,sg})x_{h,ig}^t x_{h,sg}^j \sigma_{gis} f(x_i, x_s) dx_i dx_s \\
&= \int_0^\infty \int_0^\infty \sum_{i=1}^{N_g}\sum_{s=1}^{N_g} K(u)K(w)u^t w^j \sigma_{gis} f(uh, wh) du dw = \tilde{Q}_{+,tj}
\end{aligned}$$

The $Var(Q_{+,tj})$ can be bounded by arguments similar to the ones used in lemma A.1. Then,

$$Q_{+,tj} = \tilde{Q}_{+,tj} + o_p(1)$$

Furthermore, if the cluster is defined at the tunning variable level, i.e., $X_{ig} = X_g$ for all $i = 1, \ldots, N_g$, and $h \to 0$. Then,

$$\tilde{Q}_{+,tj} = E\left[Q_{tj}\right] = f(0)\pi_{t+j} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{+,is}(0) + O(1), \text{ and } \quad = f(0)\pi_{t+j} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \sigma_{+,is} + o_p(1)$$

Noting that $G^{-1}\Psi_{+,p}$ and $G^{-1}\Psi_{-,p}$ have typical element (i, j) given by $Q_{+,tj}$ and $Q_{-,tj} = \frac{1}{Gh^2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{s=1}^{N_g} \mathbf{1}[x_{ig} \leq 0]\mathbf{1}[x_{sg} \leq 0]K(x_{h,ig})K(x_{h,sg})x_{h,ig}^{t} x_{h,sg}^{j} \sigma_{gis}$, respectively gives the result. $\square$