Heterogeneous Groups and Covariate Adjustment in Regression Discontinuity Designs

Preliminary Version — Please do not circulate

Tobias Großbölting*

March 26, 2026

Abstract

I study covariate adjustment in regression discontinuity designs estimated with data from heterogeneous groups. I demonstrate that, when groups are pooled, such heterogeneity induces bias unless the group composition is constant across the running variable. Standard estimators pool groups together and therefore generally suffer from pooling bias. To address this issue, I propose a stratified RD estimator, which first estimates the group-specific treatment effects and group shares at the treatment cut-off and then forms a weighted average. This fully nonparametric covariate adjustment reduces variance and worst-case bias, while converging at the minimax nonparametric rate. Consequently, the stratified RD estimator achieves lower asymptotic worst-case MSE than the standard local polynomial RD estimator over a common parameter space. An application to the Head Start program shows significant reductions in MSE. Resulting honest confidence intervals are 7% to 11% shorter than those of the standard RD estimator in a range of scenarios.

Keywords: RD Designs, Heterogeneity, Grouped Data, Discrete Covariates, Covariate Adjustment, Stratification, Honest Inference

JEL Codes: C14

*Universität Mannheim.

The author thanks Christoph Rothe for invaluable supervision. Further, the author thanks Paul Goldsmith-Pinkham, Max Cytrynbaum, Soonwoo Kwon, P Aronow, Harrison Zhou, Jeffrey Wooldridge, Claudia Noack, Tomasz Olma, Mengshan Xu, Pablo C. Zarate and Bjarne Horst for useful feedback and comments.

1 Introduction

The Regression Discontinuity Design (RDD) estimates the causal effect of policies assigned by a discontinuous rule: units with a running variable above a cutoff receive treatment, those below do not. Comparing units just above and below the cutoff identifies the average treatment effect at the cutoff under continuity assumptions. This local comparison is intuitively understood as approximating a randomized experiment, in the sense that covariates are balanced between treatment and control units at the cutoff. In practice, estimation requires extrapolating away from the cutoff, which introduces smoothing bias: units further away may differ systematically in observed or unobserved covariates, even when the continuity assumptions hold.

To partly adjust for these systematic differences, researchers intuitively consider including predetermined covariates to improve comparability near the cutoff. This intuition draws on standard causal reasoning for observational studies, where covariate imbalance between treated and control units induces bias (Imbens and Rubin 2015). Yet the existing literature on covariate adjustment in RDDs is limited to variance reduction. No theoretical results establish that covariate adjustment systematically reduces bias. Current theory recommends unadjusted estimators, such as the canonical local linear RD estimator, as the primary approach, treating covariate adjustment purely as a tool for efficiency gains, with point estimates expected to remain unaffected (Cattaneo et al. 2023).

In this paper I show that covariate adjustment can systematically reduce bias in empirical RD studies with heterogeneous groups. Heterogeneous groups are defined by discrete covariates, such as education type, sex, or region, that partition the sample into mutually exclusive subpopulations with heterogeneous regression functions. In this setting, the base RD estimator implicitly pools observations across heterogeneous groups. A decomposition of its bias reveals two components: the usual smoothing bias within each group, and an additive pooling bias term that arises whenever the composition of groups varies along the running variable. Under a uniform asymptotic framework, I show that this pooling bias is part of the leading term of the worst-case bias of the base RD estimators.

Building on this insight, I propose the Stratified RD estimator, which estimates group-specific treatment effects using standard local polynomial methods within each group and then averages them using appropriate weights to obtain an overall estimate of the treatment effect at the cutoff. I show that stratification eliminates pooling bias and leads to a variance

reduction under appropriate bandwidth choices.[1] The Stratified RD estimator thus achieves lower worst-case asymptotic MSE than the base RD estimator across a common parameter space. As long as the researcher is willing to put an explicit bound on treatment effect heterogeneity between groups and smoothness of the group composition along the running variable, the Stratified RD estimator provides more efficient inference on the same target parameter.

To illustrate the practical relevance of the proposed method, I apply the Stratified RD estimator to data from the Head Start program (Ludwig and Miller 2007). The estimator yields a substantial MSE reduction relative to the base RD estimator and leads to meaningfully shorter honest confidence intervals. The R package *RDStratified* **(Work in Progress)** provides an implementation of the suggested estimator.

Related Literature. The seminal work by Hahn et al. (2001) established identification of the RD parameter under continuity assumptions on the potential outcomes. They proposed estimating it by approximating the regression functions on each side of the cutoff using local polynomial estimators (Fan and Gijbels 1996). Because this estimation relies on extrapolation using observations away from the cutoff, these estimators exhibit smoothing bias. Under MSE-optimal bandwidths, this bias is of the same asymptotic order as the standard deviation, which distorts conventional inference.

To address this issue, two main approaches have emerged in the literature. Calonico et al. (2014b) use pointwise asymptotic theory to characterize the leading bias of the base RD estimator and use a bias correction to construct robust confidence intervals. In contrast, a recent strand of literature has shifted to a uniform framework which offers precise control of the finite-sample bias by providing a worst-case upper bound valid for any function within an assumed smoothness class (Armstrong and Kolesár 2018; Kolesár and Rothe 2018; Armstrong and Kolesár 2020). This allows the construction of honest confidence intervals that account for smoothing bias and guarantee nominal coverage (Li 1989). While these approaches focus on controlling bias to ensure valid inference, this paper shows that covariate adjustment reduces worst-case smoothing bias by removing the component induced by pooling heterogeneous groups.

The existing literature on covariate adjustment in RDDs primarily establishes results on variance reduction within the pointwise asymptotic framework (Calonico et al. 2019; Kramer and Kreiß 2023) with extensions to high-dimensional settings in Kreiss and Rothe (2023),

[1] Due to the second-step averaging over a discrete covariate space, the estimator does not suffer from the curse of dimensionality that arises in multivariate nonparametric estimation. The asymptotic analysis assumes a fixed number of groups, providing a good approximation in empirical settings where the number of groups is small relative to the sample size.

Noack et al. (2023), and Arai et al. (2025). Frölich and Huber (2019) propose a fully nonparametric multivariate approach that can also reduce variance. Although they note that bias could in principle be removed to restore identification when covariate distributions are discontinuous at the cutoff, they do not present results under continuous covariate distributions. In contrast, this paper establishes bias reduction by adjusting for predetermined covariates with continuous distributions, where all identifying continuity assumptions hold. Ma and Yu (2024) directly address a potential covariate imbalance between treatment and control groups using a balancing condition, but their results likewise center on variance reduction. Overall, the existing literature has not established systematic bias reduction through covariate adjustment, a gap attributable to the reliance on pointwise asymptotic theory.

In particular, the pointwise asymptotic framework has two major weaknesses in this context. First and most importantly, it does not precisely control the smoothing bias of the underlying local polynomial estimators. As a result, the pointwise asymptotic theory suggests that odd-order base RD estimators, such as the canonical local-linear choice, have asymptotically negligible bias due to covariates if their distribution is sufficiently smooth at the cutoff.[2] Relying on this pointwise asymptotic analysis leads to inaccurate bias approximations in finite samples, as documented in the simulation exercise in appendix J. The simulation reports a bias term for the local linear base RD estimator due to covariates that converges proportional to the bandwidth at empirically relevant sample sizes, even though the pointwise theory predicts a negligible bias. This highlights the practical relevance of the uniform asymptotic results developed in this paper.

Secondly, even if the bias due to covariates does not vanish in the asymptotic analysis, removing it has an ambiguous effect on overall bias. This ambiguity arises because the total smoothing bias can be decomposed into bias due to smoothing the long regression functions and the additional pooling bias arising from heterogeneous groups. The net impact of covariate adjustment therefore depends on the relative signs and magnitudes of these two components. These problems add to a growing list of drawbacks of the pointwise asymptotic framework for nonparametric regression using local polynomial estimators (Tsybakov 2009; Armstrong and Kolesár 2018; Armstrong and Kolesár 2020), motivating the uniform framework adopted in this paper.

The Stratified RD estimator is inspired by the literature on regression adjustment and post-stratification in randomized control trials (Lin 2013; Miratrix et al. 2013; Negi and Wooldridge 2021; Chang et al. 2024). Specifically, Negi and Wooldridge (2021) compare the asymptotic variance of the simple difference in means, pooled regression adjustment and full regression adjustment estimators. They find that the asymptotic variance of the

[2]That is at least twice continuously differentiable.

full regression adjustment estimator, which estimates separate regression functions for the control and treatment groups, weakly dominates those of the other two estimators. Because the Stratified RD estimator has the same structure as the full regression adjustment estimator in the case of a discrete covariate, the results on variance reduction can be understood as a generalization of Negi and Wooldridge (2021) from constant estimation in RCTs to local polynomial estimation in RDDs.

2 Setup and Motivation

This section introduces the setup, notation, and estimators used throughout the paper. Let Y_i denote the outcome variable, X_i the running variable, and $T_i := \mathbb{1}\{X_i \geq c\}$ the discontinuous treatment rule at the cutoff c, corresponding to a sharp RDD. Observed outcomes relate to potential outcomes via the treatment rule $Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]T_i$, where $Y_i(0)$ is the baseline outcome absent of treatment and $Y_i(1) - Y_i(0)$ the treatment effect. To simplify notation, I present the case of a single binary covariate Z_i that partitions the sample into two mutually exclusive groups. A generalization to K groups is given in appendix K. If the covariate is predetermined and hence unaffected by treatment (Assumption 1), one would expect the regression function of the group indicator given the running variable $g(x) := E[Z_i|X_i = x]$ to vary smoothly[3] around the cutoff.

To formalize how heterogeneous groups affect estimation in RDDs, consider the long regression function $E[Y_i \mid X_i = x, Z_i = z] = m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq c\}}$, where $m(x,z) := E[Y_i(0) \mid X_i = x, Z_i = z]$ and $\tau(x,z) := E[Y_i(1) - Y_i(0) \mid X_i = x, Z_i = z]$. The pooled regression function, which averages over group membership, is then $E[Y_i \mid X_i = x] = \overline{m}(x) + \overline{\tau}(x)\mathbb{1}_{\{x\geq c\}}$, with $\overline{m}(x) := E[m(x, Z_i) \mid X_i = x]$ and $\overline{\tau}(x) := E[\tau(x, Z_i) \mid X_i = x]$. Groups are heterogeneous if their regression functions differ, i.e. $\Delta m(x) := m(x,1) - m(x,0) \neq 0$ or $\Delta\tau(x) := \tau(x,1) - \tau(x,0) \neq 0$ for some x.

The parameter of interest is the conditional average treatment effect at the cutoff, denoted by τ. In the presence of two groups, it can be decomposed into a weighted average of the group-specific treatment effects, $\tau(c,z)$, using the law of iterated expectations. The appropriate weights for the canonical RD parameter are the conditional densities of each group at the cutoff $f_{Z|X}(z|c)$.

$$\tau := \overline{\tau}(c) = \sum_{z\in\{0,1\}} \tau(c,z) f_{Z|X}(z|c) \tag{1}$$

[3]Formally, smoothness will be characterized by Lipschitz conditions on the regression functions or their derivatives.

This parameter is typically estimated using the base RD estimator. The estimator fits local polynomial regressions on the treatment and control sides, implicitly pooling observations across groups. It then reports the difference in the estimated values of the regression function on either side of the cutoff as an estimate of $\tau(c)$.

Base RD Estimator. Formally the base RD estimator is defined to be the solution to the following weighted least squares problem.

$$\hat{\tau}(h) = e_1^\top \arg\min_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K_h(X_i)(Y_i - S_q(X_i)^\top \beta)^2$$

$K_h(u)$ denotes a chosen kernel function, h the bandwidth, $S_q(X_i) = (T_i, 1, X_i, T_i X_i, ..., X_i^q, X_i^q T_i)^\top$ collects the necessary terms of the q-th order polynomial of the running variable X_i interacted with the treatment indicator T_i and $e_1^\top := (1, 0, ...)$ selects the coefficient on the treatment variable.

The local polynomial regressions underlying the base RD estimator approximate the jump at the cutoff by extrapolating the pooled regression function. This extrapolation leads to smoothing bias, whose magnitude depends crucially on the derivatives of the pooled regression function. In the presence of heterogeneous groups, these derivatives include interactions between group heterogeneity and changes in group composition across x. Thus, extrapolation over heterogeneous groups induces pooling bias whenever the group composition varies with the running variable.

To illustrate this mechanism, I focus on baseline outcomes and relate the pooled regression function $m(x)$ to the group-specific regression functions $m(x, z)$, making the role of group heterogeneity and composition explicit. The share of group 1 at point x is given by $g(x) = f_{Z|X}(1|x)$ so that the pooled regression function can be written as

$$m(x) = m(x, 0) + \Delta m(x) g(x), \tag{2}$$

For the stylized example considered below, the leading bias of the base RD estimator depends on the second derivative of $m(x)$,

$$m''(x) = \underbrace{m''(x, 1)g(x) + m''(x, 0)(1 - g(x))}_{A} + \underbrace{2\Delta m'(x)g'(x) + \Delta m(x)g''(x)}_{B} \tag{3}$$

Term (A) reflects smoothness of the group-specific regression functions, while term (B) arises from the interaction of group heterogeneity and compositional changes. This decomposition shows that the bias of the base RD estimator will contain two components: smoothing bias

of the long regression functions, and bias from pooling heterogeneous groups.

To avoid this pooling bias I propose the Stratified RD estimator, which stratifies the sample by group, estimates the group-specific treatment effects $\tau(c, z)$ within each subsample, and averages them using an estimate of the weights in equation (1). Unlike the base RD estimator, this procedure directly targets the long regression functions and thereby prevents pooling bias.

Stratified RD Estimator. Let $\hat{\tau}_j(h)$ denote the within-group RD estimator computed for group j:

$$\hat{\tau}_j(h) = e_1^\top \arg\min_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K_h(X_i)(Y_i - S_q(X_i)^\top \beta)^2 \mathbb{1}_{\{Z_i = j\}}. \tag{4}$$

I estimate the group weights $g(x) = f_{Z|X}(1, x)$ and $(1 - g(x)) = f_{Z|X}(0, x)$ using a local polynomial regression for an interior point with bandwidth h_z:

$$\hat{g}(h_z) = e_1^\top \arg\min_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K_h(X_i)(\mathbb{1}_{\{Z_i = 1\}} - \tilde{S}_q(X_i)^\top \beta)^2, \tag{5}$$

where $\tilde{S}_q(X_i) = (1, X_i, ...X_i^q)^\top$. The Stratified RD estimator in the two group case is then given by

$$\hat{\tau}_S(h, h_z) = \hat{\tau}_1(h)\hat{g}(h_z) + \hat{\tau}_0(h)(1 - \hat{g}(h_z)). \tag{6}$$

2.1 Pooling Bias: A Stylized Example

To illustrate how group heterogeneity and compositional changes generate pooling bias, consider a stylized example in which the group-specific regression functions are constant. Applied researchers often expect regression functions to become smoother when important covariates are controlled for. Assuming constant group-specific regression functions is the limiting case of this intuition: all variation in the pooled regression function is driven by changes in group composition rather than within-group curvature. Under this assumption the long regression functions are $m(X_i, Z_i) = \alpha + \beta Z_i$ and $\tau(X_i, Z_i) = \gamma + \delta Z_i$, where β and δ capture group heterogeneity in baseline outcomes and treatment effects respectively. The corresponding pooled regression functions are $m(x) = \alpha + \beta g(x)$ and $\tau(x) = \gamma + \delta g(x)$.

To estimate the pooled regression functions nonparametrically, one has to control the smoothing bias by placing constraints on the derivatives of $g(x)$. A standard way to enforce smoothness is to assume that $g(x)$ is p-times continuously differentiable and that the absolute value of its p-th derivative is globally bounded by a known constant L, picked by the researcher. For technical reasons, it will be important to consider the closure (in sup norm)

of this class of functions. This closure includes all limiting functions that can be arbitrarily well approximated by such smooth functions and may be discontinuous on a set of measure zero. This leads to the Hölder class used in Armstrong and Kolesár (2020).

Definition 1 (Hölder Class of Order p)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ and $p \geq 1$ be an integer, then the (integer) Hölder class of order p is defined as:*

$$\mathcal{F}_p(L) := \left\{ f : |f^{(p-1)}(x) - f^{(p-1)}(y)| \leq L|x - y| \quad \forall x, y \in \mathbb{R} \right\}$$

I assume $g(x) \in \mathcal{F}_2(L)$, so that L bounds the second derivative of $g(x)$. To fully control pooling bias in the stylized example, I additionally assume an upper bound on the heterogeneity between groups in baseline outcomes $|\beta| \leq C_{\Delta m}$ and treatment effects $|\delta| \leq C_{\Delta\tau}$.

Bias of the Base RD Estimator. Under the maintained smoothness assumptions, the optimal polynomial order for estimation is $q = p - 1$, which yields the canonical local linear estimator in this example (Armstrong and Kolesár 2020). I assume the use of the uniform kernel, which assigns equal weight to all units that fall within the bandwidth. For analytical convenience, the base RD estimator can be written as a weighted sum of outcomes:

$$\hat{\tau}(h) = \sum_{X_i \geq 0} w_+(X_i; h) Y_i - \sum_{X_i < 0} w_-(X_i; h) Y_i$$

Since the weights depend on the data only through the running variable X_i, taking expectations conditional on the sample realizations $\mathbf{X} := \{X_i\}_{i=1}^n$ treats the weights as fixed. Standard calculations then yield the exact finite-sample bias, which depends on the second derivatives of the pooled regression functions:

$$E[\hat{\tau}(h)|\mathbf{X}] - \tau = \sum_{X_i \geq 0} w_+(X_i; h) \int_0^{X_i} [m''(s) + \tau''(s)] ds - \sum_{X_i < 0} w_-(X_i; h) \int_{X_i}^0 m''(s) ds \quad (7)$$

After plugging in the derivatives $m''(x) = \beta g''(x)$ and $\tau''(x) = \delta g''(x)$, equation (7) shows that the bias is entirely driven by group heterogeneity β, δ and compositional changes along the running variable $g''(x)$. To upper bound this bias, I take the supremum over the assumed smoothness and heterogeneity bounds. Applying the results in Armstrong and Kolesár (2020) then gives the leading-order expression for the worst-case finite sample bias, with the $o(h^2)$ term capturing randomness in the realizations of the running variables:

$$\overline{Bias}(\hat{\tau}(h)) := \sup_{\{g(x), \beta, \tau_\beta\}} E[\hat{\tau}(h)|\mathbf{X}] - \tau = h^2 L [2C_{\Delta m} + C_{\Delta\tau}] \frac{1}{12} + o(h^2) \quad (8)$$

Equation (8) shows that the leading bias term converges at rate h^2 and is scaled by the product of the bounds on compositional changes and group heterogeneity in baseline outcomes and treatment effects. The term $L[2C_{\Delta m} + C_{\Delta\tau}]$ upper bounds this bias due to pooling heterogeneous groups. This pooling bias is always present from a worst-case perspective unless the researcher can safely assume that groups are homogeneous, $C_{\Delta m} = C_{\Delta\tau} = 0$, or that the composition of groups does not change across the running variable, $L = 0$.[4]

Bias of the Stratified RD Estimator. The general worst-case bias for the Stratified RD estimator is established in proposition 3. Because the regression functions are assumed constant within groups, the group-specific treatment effect estimates $\hat{\tau}_j(h)$ are unbiased. Therefore, the upper bound on the bias equals the worst-case bias of the weight estimate $\hat{g}$ multiplied with treatment effect heterogeneity. Using the worst case bound on treatment effect heterogeneity $|\delta| \leq C_{\Delta\tau}$ and the smoothness assumptions on $g(x)$ gives:

$$\overline{\text{Bias}}\left(\hat{\tau}_S(h, h_z)\right) = h_z^2 L\, C_{\Delta\tau} \frac{1}{6} + o(h_z^2) \tag{9}$$

The worst-case bias converges at rate h_z^2, the bandwidth used for estimating the weights, and the product $L\, C_{\Delta\tau}$ appears as in equation (8) for the base RD estimator. However, it is important to note that the bias of the Stratified RD estimator does not suffer from bias due to heterogeneity in baseline outcomes $C_{\Delta m}$.

Comparison of Bias Expressions. Lastly, to compare the bias expressions in equation (8) and equation (9), it is useful to relate the bandwidth choice for the weight estimation h_z to the bandwidth used for estimating treatment effects h. Since both estimators converge at the same rate in this setting, the theoretical analysis considers bandwidth sequences that shrink proportionally, parameterized as $h_z = \kappa h$ for some constant $\kappa > 0$ (Assumption 3). Substituting $h_z = \kappa h$ yields:

$$\overline{\text{Bias}}(\hat{\tau}(h)) - \overline{\text{Bias}}(\hat{\tau}_S(h, \kappa h)) = h^2 L\, C_{\Delta m} \frac{1}{6} + h^2 L\, C_{\Delta\tau} \frac{1}{6} \left(\frac{1}{2} - \kappa^2\right) + o(h^2) \tag{10}$$

As long as $\kappa \leq \sqrt{\frac{1}{2}}$, the bias of the Stratified RD estimator is weakly smaller than that of the base RD estimator, regardless of the assumed constants $C_{\Delta m}$ and $C_{\Delta\tau}$.

[4] If the researcher places a smoothness assumption directly on the pooled regression function the resulting bias expression does not contain an explicit term for pooling bias. However, as shown in equation (3) the pooled regression function includes terms that generate pooling bias. Thus, any assumptions on the pooled regression function implicitly make an assumption on the size of pooling bias.

In section 3, the general analysis establishes an interval of bandwidth ratios that simultaneously guarantee both lower worst-case bias and lower variance for the Stratified RD estimator, with $\kappa = \sqrt{\frac{1}{2}}$ serving as its upper bound. Stratification under these bandwidth choices thus reduces worst-case MSE overall by eliminating bias arising from pooling heterogeneous groups while simultaneously improving estimation precision.

3 General Results

This section presents the asymptotic distribution and MSE of the base RD and Stratified RD estimators in the two-group setting. Results generalized to several groups can be found in Appendix K. The analysis closely follows Armstrong and Kolesár (2020) and the smoothness assumptions used in the following theorems are standard in nonparametric estimation. A notable exception is the common parameter space $\mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$, which restricts the pooled and stratified estimators to the same underlying smoothness class and thereby enables a meaningful MSE comparison in Theorem 6. Its construction will be discussed in Section 3.4.

The pooling approach and the base RD estimator have been studied extensively in the literature. For reasons outlined in the introduction, I rely on the existing work using the uniform asymptotic framework. To fix ideas and establish notation, I first restate a probabilistic version of the results of Armstrong and Kolesár (2020) in Theorem 1. The main results of this paper are then presented in Theorem 2, Proposition 3 and Theorem 6.

3.1 Pooled Estimation - Base RD

The standard approach to estimate $\tau(c)$ in RDDs is to use the base RD estimator that pools all units together, effectively ignoring the group structure. The choice of polynomial order should depend on the smoothness assumptions made by the researcher. A common choice is the local linear estimator which sets $q = 1$. This is justified and optimal when the researcher assumes the regression function to lie in a second order Hölder smoothness class ($p = 2$).

Theorem 1 restates a probabilistic version of the results in Armstrong and Kolesár (2020), that characterize the asymptotic bias and variance of the base-RD estimator. Their honest inference approach assumes that the targeted regression function lies in a p-th order Hölder function class $\mathcal{F}_p(M_P)$. This implies that the p-th derivative of the regression function is bounded (a.e.) by a constant M_P chosen by the researcher. The subscript indicates that the bound is chosen for the pooled regression function. The upper bound on the p-th derivative is used to bound the bias of the nonparametric estimator by the worst-case bias over the function class, determined by the constant M_P. I denote this worst case bias as

$\overline{\text{Bias}}(\hat{\tau}(h)) := \sup_{\mathcal{F}} E[\hat{\tau}(h)|\mathbf{X}] - \tau$.

Theorem 1 (Asymptotic Worst-Case MSE - Base-RD (Armstrong and Kolesár 2020))**.** *Assume assumption 2 to assumption 5. Additionally require the following smoothness assumption on the regression function,* $m(x), m(x) + \tau(x) \in \mathcal{F}_p(M_P)$.

Then

$$V[\hat{\tau}(h)|\mathbf{X}] = \frac{\mathcal{V}_B}{nh}(1 + o(1)) \tag{11}$$

$$\overline{\text{Bias}}(\hat{\tau}(h)) = h^p \frac{2M_P}{p!} B_{p,q}(k)(1 + o(1)) \tag{12}$$

with $\mathcal{V}_B = \frac{\overline{k_q^2}\left(\sigma^2_{P,+}(c) + \sigma^2_{P,-}(c)\right)}{f_X(c)}$, $\overline{k_q^2} = \int_0^\infty k_q^*(u)^2 du$ *and* $B_{p,q}(k)$ *kernel constants tabulated in table S.1 of the online appendix in Armstrong and Kolesár (2020).*

Theorem 1 shows that the asymptotic worst-case bias of the base RD estimator shrinks with the bandwidth at the optimal nonparametric rate h^p. The size of the bias crucially depends on the bound M_P on the p-th derivative that the researcher assumed for the pooled regression function. Note that this upper bound implicitly also bounds any pooling bias. If the data stem from heterogeneous groups the smoothness bound has to be inflated accordingly. The asymptotic variance, besides a kernel constant and the density of the running variable at the cut-off, depends on the variance of the error terms in the pooled regression model $\sigma^2_P(x) = V(u_i|X_i = x)$. $\sigma^2_{P,+}(c)$ and $\sigma^2_{P,-}(c)$ denote the respective right- and left-sided limits at the cut-off.

3.2 Pooling Heterogeneous Groups

Pooling heterogeneous groups leads to pooling bias and additional variance.

Pooling Bias. As demonstrated in section 2, pooling bias occurs because group heterogeneity enters the pooled regression through the LIE as demonstrated in **??**. Taking the first derivative of the pooled regression function reveals the terms responsible for pooling bias.

$$m'(x) = \underbrace{m'(x,1)g(x) + m'(x,0)(1 - g(x))}_{I)} + \underbrace{\Delta m(x)\, g'(x)}_{II)} \tag{13}$$

For instance, in the case of the base RD estimator with local constant estimation, the bias depends on the first derivative of the regression function. In this setting, the two

multiplicative terms under II) correspond to pooling bias. Here, $\Delta m(x) = [m(x,1) - m(x,0)]$ captures level differences between the two groups. The derivative of the conditional density function $g'(x)$ describes how the group composition changes along the running variable. The product of these two terms thus represents the contribution of group heterogeneity and compositional changes to the slope of the pooled regression function. The terms under I) correspond to standard smoothing bias due to the slopes in the long regression functions.

The base RD framework assumes a single bound on the first derivative of the pooled regression function ($|m'(x)| \leq M_P$), which implicitly bounds all terms together. This makes it impossible to directly quantify the extent of pooling bias. However, for the bound to be valid and the resulting confidence intervals to achieve nominal coverage, the upper bound has to account for the group heterogeneity and potential compositional changes. Only when either of the two terms, or both, are exactly zero pooling bias does not occur. This would be the case if the groups were exactly homogeneous ($\Delta m(x) = 0$) or if the group distribution were as-if randomly assigned ($g'(x) = 0$). To theoretically quantify the extent of pooling bias, I explicitly bound the between group heterogeneity and changes in the conditional density of groups in the common parameter space constructed in Section 3.4 .

To generally study RD estimators based on q-th order local polynomial estimators, one has to look at higher order derivatives. In general, the j-th derivative of $m(x)$ contains multiple cross-terms that describe heterogeneity in the slopes, curvature, and higher-order variation between groups.

$$m^{(j)}(x) = m^{(j)}(x,1)g(x) + m^{(j)}(x,0)(1 - g(x)) + R_j(x), \tag{14}$$

with $R_j(x) := \sum_{k=0}^{j-1} \binom{j}{k} \Delta m^{(k)}(x) g^{(j-k)}(x)$ and $\Delta m^{(k)}(x) := [m^{(k)}(x,1) - m^{(k)}(x,0)]$, and $m^{(l)}(x)$ denoting the l-th derivative of a function. The term $R_j(x)$, is the part of the derivative that is due heterogeneity between groups and compositional changes that induces pooling bias for the base RD estimator.

In principle, the analysis for q-th order local polynomial estimators follows immediately from the arguments above. However, due to the increasing number of terms the presentation becomes cluttered and is therefore omitted. Formal generalizations can be found in Appendix K.

Variance due to Pooling. Well-known arguments in the literature on covarite adjustments for RCTs (Negi and Wooldridge 2021) or RDDs (Calonico et al. 2019) demonstrate that omitted variables will be absorbed into the error term. This leads to additional variance in the error term relative to the error term of a regression function that includes the covariates. This can be seen by relating the pooled and long regression functions.

$$
\begin{aligned}
Y_i &= [m(X_i,1) + \tau(X_i,1)T_i]Z_i + [m(X_i,0) + \tau(X_i,0)T_i](1 - Z_i) + e_i \\
&= m(X_i) + \tau(X_i)T_i + \underbrace{([m(X_i,1) - m(X_i,0)] + [\tau(X_i,1) - \tau(X_i,0)]T_i)\,(Z_i - f_{Z|X}(1|x)) + e_i}_{u_i} \\
&= m(X_i) + \tau(X_i)T_i + u_i
\end{aligned}
$$

where $u_i = ([m(X_i,1) - m(X_i,0)] + [\tau(X_i,1) - \tau(X_i,0)]T_i)\,(Z_i - f_{Z|X}(1|x)) + e_i$ is the error for the pooled regression function. This error term now contains additional terms that depend on the heterogeneity between groups. Let $\sigma_P^2(X_i) = V(u_i|X_i)$ the conditional variance of the error terms u_i in the pooled regression model. Further, denote $\sigma_L^2(X_i, Z_i) = V(e_i|X_i, Z_i)$, $\sigma_L^2(X_i) = V(e_i|X_i) \overset{\text{LIE}}{=} E[E[e_i^2|X_i, Z_i]|X_i = x] = \sum_{z\in(0,1)} \sigma_L^2(x,z) f_{Z|X}(z|x)$ the conditional variances of the error term e_i in the long regression model.

Lemma 3.1 (Pooled and Long Variance Comparison - Law of Total Variance)**.** *Define* $\Delta m(X_i) := [m(X_i,1) - m(X_i,0)]$ *and* $\Delta\tau(X_i) := [\tau(X_i,1) - \tau(X_i,0)]$*, then*

$$\sigma_P^2(X_i) - \sigma_L^2(X_i) \geq [\Delta m(X_i) + \Delta\tau(X_i)T_i]^2 \sigma_Z^2(X_i)$$

Lemma A.1 proves that the variance of the error term in the long regression model is whenever group heterogeneity is present. This provides the foundation for the variance-reduction properties of covariate-adjusted estimators documented in the literature.

3.3 Stratified Estimation - Stratified RD

The stratified RD estimator solves the problem of pooling heterogeneous groups by directly targeting the long regression functions. It requires smoothness assumptions on the long regression functions, the regression function of the covariate and a bound on treatment effect heterogeneity. It does not require any additional bounds on baseline heterogeneity or other cross-terms in equation (14). Optimal bandwidth choices for h and h_z, under smoothness assumptions on $m(x,z), \tau(x,z)$ and $g(x)$ target the $(p-1)$-th derivative, both scale proportional to $n^{-\frac{1}{2(p-1)+3}}$. To model this I assume the supplementary bandwidth h_z to be proportional to h up to a constant $h_z = \kappa h$ (assumption 3). Theorem 2 then gives a characterization of the asymptotic distribution of the stratified RD estimator.

Theorem 2 (Asymptotic Normality of the Stratified RD Estimator)**.** *Assume assumption 2, assumption 3, and assumption 6 to assumption 8. Further assume* $m(x,z), m(x,z)+\tau(x,z) \in \mathcal{F}_p(M_L)$*,* $g_1(x) \in \mathcal{F}_p(L)$ *and* $|\tau(x,1)-\tau(x,0)| \leq C_{\Delta\tau} \quad \forall x$*, then*

$$\sqrt{nh}\left(\hat{\tau}_S(h) - \tau - LB\right) \xrightarrow{d} N(0, \mathcal{V}_S)$$

where $LB := g_1 E[\hat{\tau}_1(h) - \tau_1 | \mathbf{X}, \mathbf{Z}] + (1-g_1) E[\hat{\tau}_0(h) - \tau_0 | \mathbf{X}, \mathbf{Z}] + [\tau_1 - \tau_0] E[\hat{g}_1(h_z) - g_1 | \mathbf{X}]$ *denotes the leading-order bias terms and the asymptotic variance is given by* $\mathcal{V}_S = \mathcal{V}_O + [\tau_1 - \tau_0]^2 \frac{1}{\kappa} \mathcal{V}_g$*. With* $\mathcal{V}_O$ *given in Lemma D.4 and* $\mathcal{V}_g$ *being the asymptotic variance of the weight estimation using a local linear estimator at an interior point using* Z_i *as outcome (see section 3.2 in Armstrong and Kolesár (2020)).*

Proof. See appendix F. □

The group-wise estimation and second stage averaging step avoids any pooling bias from cross-terms in the derivative of the regression function in equation (14). The only heterogeneity that contributes to its leading bias is treatment effect heterogeneity. Consequently, we get the following asymptotic worst-case bias.

Proposition 3 (Worst-Case Bias of the Stratified RD estimator)**.** *Assume assumption 2, assumption 3, and assumption 6 to assumption 8. Further assume* $m(x,z), m(x,z)+\tau(x,z) \in \mathcal{F}_p(M_L)$*,* $g_1(x) \in \mathcal{F}_p(L)$ *and* $|\tau(x,1)-\tau(x,0)| \leq C_{\Delta\tau} \quad \forall x$*, then*

$$\overline{\text{Bias}}(\hat{\tau}_S(h, h_z)) := \sup_{\{m(x,z), m(x,z)+\tau(x,z), g_1(x), \Delta\tau\}} |LB| \tag{15}$$

$$= h^p \frac{2M_L}{p!} B_{p,q}(k) + h_z^p C_{\Delta\tau} \frac{L}{p!} B_{p,q}^{int}(k) + o(h^p) \tag{16}$$

Proof. See appendix D.V. □

The worst-case bias of the strat-RD estimator primarily depends on the constant M_L and converges at the optimal rate h^p. Because the target weight $g(c)$ is estimated the expression contains an additional bias term, which depends on the product of the bound on treatment effect heterogeneity and second derivative of the conditional density function $C_{\Delta\tau} L$.

3.4 Common Parameter Space

Given the results on the asymptotic MSE of the base RD and stratified RD estimator, how do the two estimation approaches compare to each other in empirically relevant settings? In general they target different regression function objects. Theorem 1 and Theorem 2 invoke different smoothness assumptions, none of them implying the other. Most importantly, the base RD estimator requires a bound on the p-th derivative of the short regression function, while the Stratified RD estimator requires one on the long regression function. To demonstrate the relation between the sets of assumptions I collect the regression object in broader function classes.

Definition 2 (Base RD Smoothness Class)**.**

$$\mathcal{F}_p^{Base}(M_P) := \left\{ \begin{array}{l} m(x) + \tau(x)\mathbb{1}_{\{x\geq 0\}}, \\ m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq 0\}}, \\ g_1(x) \end{array} \middle| \; m(x), m(x)+\tau(x) \in \mathcal{F}_p(M_P) \right\}$$

Definition 3 (Stratified RD Smoothness Class)**.**

$$\mathcal{F}_p^{Strat}(\mathbf{M},\mathbf{C},\mathbf{L}) := \left\{ \begin{array}{l} m(x) + \tau(x)\mathbb{1}_{\{x\geq 0\}}, \\ m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq 0\}}, \\ g_1(x) \end{array} \middle| \; \begin{array}{l} m(x,z), m(x,z)+\tau(x,z) \in \mathcal{F}_p(M_L), \\ g_z(x) \in \mathcal{F}_p(L), \\ |\tau(x,1) - \tau(x,0)| \leq C_{\Delta\tau}, \forall x \end{array} \right\}$$

The function classes $\mathcal{F}^{\text{Base}}$ and $\mathcal{F}^{\text{Strat}}$ formally collect the required smoothness assumptions for Theorem 1 and Theorem 2. The two function classes are overlapping but not nested. For example, $\mathcal{F}^{\text{Base}}$ does not impose a bound on $m'(x,z)$ and therefore the worst-case bias of $\hat{\tau}_S$ is infinite over $\mathcal{F}^{\text{Base}}$. Likewise, $\mathcal{F}^{\text{Strat}}$ has no bound on $m'(x)$ and the worst-case bias of $\hat{\tau}$ is infinite over $\mathcal{F}^{\text{Strat}}$. Consequently, it is impossible to compare the worst-case MSE of the estimators over either of the two function classes.

To make the estimators comparable, a common parameter space, that links the bound M_P on the pooled and M_L on the long regression function, is necessary. For the construction of this common parameter space I apply the triangle inequality to equation (13). The application of the triangle inequality is sensible here, since the analysis follows a worst-case perspective as in Armstrong and Kolesár (2020). The triangle inequalitys only rules out edge cases in which different bias terms cancel each other. Then I place constraints on the terms in the derivatives of the pooled regression functions $m(x)$ and $\tau(x)$. These additional assumption taken together with the assumptions in theorem 2 then give a bound on the pooled regression function.

$$|m'(x)| = |\sum_{z\in\{0,1\}} m'(x,z)g_z(x) + \Delta m(x) g_1'(x)| \tag{17}$$

$$\leq \sum_{z\in\{0,1\}} |m'(x,z)|g_z(x) + |\Delta m(x)||g_1'(x)| \tag{18}$$

$$\leq M + C_{\Delta m} L := M_P \tag{19}$$

I bound all the terms individually by assuming the regression objects to satisfy the smoothness assumptions with constants $M, L, C_{\Delta m}$ to construct $\mathcal{F}^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$. By placing bounds on all the individual objects we can directly control how each of the different components contributes to the derivative. Note that since Z_i is a binary variable, $g_z(x)$ is naturally bounded between 0 and 1. Further $g_1(x) + g_0(x) = 1$. Likewise I derive the smoothness bound for $m(x) + \tau(x)$.

Definition 4 (Restricted Smoothness Class)**.**

$$\mathcal{F}_1^R(\mathbf{M}, \mathbf{C}, \mathbf{L}) = \left\{ \begin{array}{l} m(x) + \tau(x)\mathbb{1}_{\{x\geq 0\}} \\ m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq 0\}} \\ g_1(x) \end{array} \middle| \begin{array}{l} m(x,z), m(x,z) + \tau(x,z) \in \mathcal{F}_1(M), \\ g_z(x) \in \mathcal{F}_1(L), \\ |m(x,1) - m(x,0)| \leq C_{\Delta m}, \forall x \\ |\tau(x,1) - \tau(x,0)| \leq C_{\Delta \tau}, \forall x \end{array} \right\}$$

The construction of $\mathcal{F}_1^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$ ensures that $\mathcal{F}_1^{\text{Base}}(M_P) \subseteq \mathcal{F}_1^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$ and $\mathcal{F}_1^{\text{Strat}}(\mathbf{M}, \mathbf{C}, \mathbf{L}) \subseteq \mathcal{F}_1^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$. In general, one can construct such a common parameter space also for higher order local polynomial estimators. Such a general parameter space $\mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$ will include additional cross-terms of the higher order derivatives. In total a number of $2 + 2p$ restrictions and bounds will be necessary. To avoid notational clutter I omit an explicit derivation for $\mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$.

3.5 Estimator Comparison

Given the common parameter space $\mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$ theorem 6 proofs that the feasible stratified RD estimator attains lower worst-case asymptotic MSE than the base RD estimator. The base RD estimator will have additional variance if groups are not homogeneous ($(C_{\Delta m}, C_{\Delta \tau}) \neq (0,0)$) and additional worst-case bias if groups are heterogeneous and the conditional density of the group indicator is non-constant across the support of the running variable $L \neq 0$).

Proposition 4 (Leading Term Worst-Case Bias Comparison). *Assume assumption 2 to assumption 8. Additionally assume the smoothness constraints* $m(x), m(x) + \tau(x), m(x,z), m(x,z) + \tau(x,z), g_1(x), [\tau(x,1) - \tau(x,0)] \in \mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$ *Following assumption 3 one chooses* $h_z = \kappa h$.

Denote the leading bias terms of the asymptotic expansion as, $\overline{\text{Bias}}_{\text{lead}}(\hat{\tau}) := \lim_{n\to\infty} \frac{\overline{\text{Bias}}(\hat{\tau})}{h^p}$. *As long as* $\kappa > 0$, *is chosen smaller than a kernel specific constant* $\kappa \leq U_{p,q}(k)$, *that depends on p and polynomial order q, then the leading terms of the bias of the estimators can be ranked*

$$\overline{\text{Bias}}_{\text{lead}}(\hat{\tau}_S) \leq \overline{\text{Bias}}_{\text{lead}}(\hat{\tau})$$

or equivalently,

$$\overline{\text{Bias}}(\hat{\tau}_S) \leq \overline{\text{Bias}}(\hat{\tau}) + o(h^p)$$

Proposition 5 (Asymptotic Variance Comparison). *Let* $\mathcal{V}_B$ *denote the asymptotic variance of the base RD estimator and* $\mathcal{V}_S$ *denote the asymptotic variance of the Stratified RD estimator respectively. For an appropriate scaling of the secondary bandwidth* $h_z = \kappa h$ *(assumption 3) with* $\kappa \geq L_{p,q}(k)$ *the following holds:*

$$\mathcal{V}_S \leq \mathcal{V}_B \tag{20}$$

with strict inequality if there is group heterogeneity.

Proposition 4 and proposition 5 establish that the stratified estimator improves both bias and variance. Taken together the results imply Theorem 6, where the respective bounds in the previous results form the interval $\kappa \in I_{p,q}(k) := [L_{p,q}(k), U_{p,q}(k)]$. Theorem 6 presents the main claim of the paper: stratification reduces the overall asymptotic worst-case MSE.

Theorem 6 (Asymptotic Worst-Case MSE Comparison). *Assume assumption 2 to assumption 8. Additionally assume the smoothness constraints* $m(x), m(x)+\tau(x), m(x,z), m(x,z)+\tau(x,z), g_1(x), [\tau(x,1)-\tau(x,0)] \in \mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$. *The bandwidths are chosen according to assumption 3,* $h_z = \kappa h$. *Define the asymptotic worst-case MSE of the two estimators as*

$$\overline{AMSE}(\hat{\tau}(h)) := \left(h^p \overline{\text{Bias}}_{\text{lead}}(\hat{\tau}(h))\right)^2 + \frac{V_B}{nh},$$
$$\overline{AMSE}(\hat{\tau}_S(h, h_z)) := \left(h^p \overline{\text{Bias}}_{\text{lead}}(\hat{\tau}_S(h, h_z))\right)^2 + \frac{V_S}{nh}.$$

Then if κ is chosen from a kernel specific interval $\kappa \in I_{p,q}(k)$, that depends on the used kernel, the order of smoothness p and polynomial order q, the asymptotic worst case MSE of the Stratified RD estimator is uniformly smaller than that of the base RD estimator over $\mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$.

$$\overline{AMSE}(\hat{\tau}_S(h, h_z)) \leq \overline{AMSE}(\hat{\tau}(h))$$

With strict inequality whenever groups are heterogeneous.

Remark 1 (Kernel and Polynomial Order Specific Intervals)**.** *The kernel specific intervals for the local constant and local linear estimator using a uniform kernel are:*

- $I_{1,0}(uniform) = \{1\}$
- $I_{2,1}(uniform) = [\frac{1}{4}, \sqrt{\frac{1}{2}}] \approx [0.25, 0.707]$

Note that the interval for possible choices of κ is quite conservative due the inequality approach in the variance comparison. The MSE improvement occurs due to variance reduction by fully nonparametric covariate adjustment, as well as a reduction of the worst-case bias compared to the base RD estimator.

Due to the nature of a worst-case analysis, one of the benefits of the stratified estimator, namely its ability to avoid potential pooling bias, is guaranteed. Regardless of whether pooling bias is actually present in a given study, as long as the researcher cannot a priori rule out pooling bias we have to account for its possibility in the bias aware confidence interval. As long as the researcher is not willing to assume full group-homogeneity or perfectly balanced covariates across the running variable, she has to inflate the confidence intervals of the Base-RD estimator accordingly (Armstrong and Kolesár 2020). The stratified RD estimator will prevent pooling bias in any case. Consequently the worst-case bias is smaller and we inflate the confidence interval less.

Lastly, corollary 7 shows that both estimators are rate-optimal over the restricted smoothness class. MSE improvements are achieved by reducing worst-case bias and variance constants.

Corollary 7 (Rate Optimality)**.** *As long as $m(x, z) \in \mathcal{F}_p$ and $g(x) \in \mathcal{F}_p$ lie in a Hölder class of same order p and given bandwidth choices consistent with assumption 3, the estimators $\hat{\tau}(h)$ and $\hat{\tau}_s(h, h_z)$ achieve the mini-max optimal rate $n^{-\frac{2p}{2p+1}}$ over the smoothness class $\mathcal{F}^R(\vec{M})$.*

3.6 Differing Order of Smoothness Between Outcome and Covariate Regression

[Work in Progress]

In settings most relevant for empirical work, both estimators are rate optimal under the given smoothness assumptions and MSE improvements are achieved by reducing worst-case bias and variance constants. Special cases in which the Stratified RD estimator achieves the mini-max rate, while the base RD estimator, and in fact any other pooling estimator, fails to do so, are discussed in section 3.6.

4 Implementation

Given the result in theorem 6, we know that under appropriate bandwidth choices the stratified estimator has lower asymptotic worst-case MSE than the base-RD estimator. For the construction of theorem 6 we have assumed the stratified RD estimator to use the same bandwidth h, as the base RD estimator, to estimate the group-specific treatment effects $\hat{\tau}_z(h)$. However, additional MSE improvements can be made by optimizing the bandwidths h_1, h_0, h_z specifically for $\hat{\tau}_S(h_1, h_0, h_z)$ using a similar approach to Armstrong and Kolesár (2020).

4.1 MSE Optimal Bandwidths

As long as the numerical optimization correctly determines the optimal bandwidths h_1^*, h_0^*, h_z^* we know that, $\overline{AMSE}(\hat{\tau}_S(h_1^*, h_0^*, h_z^*)) \leq \overline{AMSE}(\hat{\tau}_S(h_{Base}^*, \kappa h_{Base}^*))$. This holds, because h_{Base}^* is also included in the set we are optimizing over. It follows that $\overline{AMSE}(\hat{\tau}_S(h_1^*, h_0^*, h_z^*)) \leq \overline{AMSE}(\hat{\tau}(h_{Base}^*))$, which leads to shorter bias aware confidence intervals in the empirical applications in section 5.

Additional reductions in asymptotic MSE are achieved, because the separate bandwidth choices h_1, h_0, h_z optimize the bias-variance trade-off with respect to group-specific heteroskedasticity and smoothness assumptions. I follow the approach in the literature, that given bounds on the second derivatives, uses an algorithm to pick the MSE optimal bandwidths. Specifically, I adopt the procedure outlined in section 3.3.2 in Armstrong and Kolesár (2020) to choose the MSE optimal bandwidth.

$$\min_{h_1,h_0,h_z} \overline{AMSE}(h_1,h_2,h_z) = \min_{h_1,h_0,h_z} \overline{\text{Bias}}(\hat{\tau}_s;h_1,h_0,h_z)^2 + V(\hat{\tau}_s;h_1,h_0,h_z) \tag{21}$$

$$= \min_{h_1,h_0,h_z} \left(g_1(c)\overline{\text{Bias}}(\hat{\tau}_1,h_1) + g_0(c)\overline{\text{Bias}}(\hat{\tau}_0,h_0) + K_{\Delta\tau}\overline{\text{Bias}}(\hat{g}_1,h_z)\right)^2 \tag{22}$$

$$+ g_1^2(c)V(\hat{\tau}_1;h_1) + g_0^2(c)V(\hat{\tau}_0;h_0) + \Delta\tau^2 V(\hat{g}_1;h_z) \tag{23}$$

The general strategy is to plug in the theoretical worst-case bias for $\overline{\text{Bias}}(\hat{\tau}_s;h_1,h_0,h_z)^2$ and use a consistent pilot variance estimator $\hat{V}(\hat{\tau}_s;h_1,h_0,h_z)$ to proxy $V(\hat{\tau}_s;h_1,h_0,h_z)$ and then numerically optimize the bandwidths. Since in this problem $g_1(c)$ is unknown, I propose an iterative procedure.

If we approximate, $\overline{\text{Bias}}(\hat{\tau}_s;h_1,h_0,h_z)^2 \approx \left(g_1(c)\overline{\text{Bias}}(\hat{\tau}_1,h_1) + g_0(c)\overline{\text{Bias}}(\hat{\tau}_0,h_0)\right)^2 + \left(K_{\Delta\tau}\overline{\text{Bias}}(\hat{g}_1,h_z)\right)^2$, which drops the cross term $2\left(g_1(c)\overline{\text{Bias}}(\hat{\tau}_1,h_1) + g_0(c)\overline{\text{Bias}}(\hat{\tau}_0,h_0)\right)\left(K_{\Delta\tau}\overline{\text{Bias}}(\hat{g}_1,h_z)\right)$, we get the following minimization

$$\min_{h_1,h_0,h_z} \left(g_1(c)\overline{\text{Bias}}(\hat{\tau}_1,h_1) + g_0(c)\overline{\text{Bias}}(\hat{\tau}_0,h_0)\right)^2 + \left(K_{\Delta\tau}\overline{\text{Bias}}(\hat{g}_1,h_z)\right)^2 \tag{24}$$

$$+ g_1^2(c)V(\hat{\tau}_1;h_1) + g_0^2(c)V(\hat{\tau}_0;h_0) + \Delta\tau^2 V(\hat{g}_1;h_z) \tag{25}$$

$$= \min_{h_1,h_0} \left(g_1(c)\overline{\text{Bias}}(\hat{\tau}_1,h_1) + g_0(c)\overline{\text{Bias}}(\hat{\tau}_0,h_0)\right)^2 + g_1^2(c)V(\hat{\tau}_1;h_1) + g_0^2(c)V(\hat{\tau}_0;h_0) \tag{26}$$

$$+ \min_{h_z} \left(K_{\Delta\tau}\overline{\text{Bias}}(\hat{g}_1,h_z)\right)^2 + \Delta\tau^2 V(\hat{g}_1;h_z) \tag{27}$$

Now we can first pick h_z to estimate $\hat{g}_1(c)$ and then plug $\hat{g}_1(c)$ into the second minimization. This gives slightly larger bandwidths, as the approximate solution drops a cross-term of the biases in the square.[5] Now one can simply plug-in the expressions for the bias and variance of the within-group estimators into the expression in Armstrong and Kolesár (2020). For simplicity we set $K''_{m,1} = K''_{m,0}$. But if the researcher has strong priors on different curvatures in both groups one can specify different smoothness bounds and choose the bandwidth in each group adaptively.

Algorithm

1. First compute pilot estimates for the variance for the outcome in each group, as well

[5] In the empirical application this approximation works well as the bias-standard error ratio is close to the theoretical optimum of 0.5 for local linear regression (see Armstrong and Kolesár (2020)).

as the group dummy

2. Given the bound on the second derivative K_z'' use the procedure in Armstrong and Kolesár (2020) to pick h_z and estimate the weight $\hat{g}_1(c)$

3. Plug $\hat{g}_1(c)$ into equation (24) and then solve numerically for h_1 and h_0.

5 Empirical Applications

I replicate the analysis by (Barcellos and Jacobson 2015) of the effect of Medicare on financial stress of individuals and the analysis by (Ludwig and Miller 2007) of the Head Start program on child mortality. I pick the smoothness constants to match the bandwidth choices made in the paper. Further, I do not use the survey weights used in the papers and rely on the raw data provided in the replication dataset. I consider four different scenarios under which pooling bias is responsible for a share s of the upper bound on the second derivative of the regression function. As s increases the stratified estimator will have less worst-case bias for fixed h. Hence for larger s the reduction in CI length using the stratified estimator is larger. The first column in the tables with $s = 0$ assumes group homogeneity and reductions in CI length are purely due to variance reduction.

The tables presenting estimates and CIs for the base-RD and stratified-RD estimators and the reduction in CI length can be found in section 7. In the replicated studies CIs based on the stratified RD estimator are between 3% to 11% shorter than CIs for the base RD estimator, depending on how much of overall bias is attributed to pooling bias. This is expressed through the ratio s.

6 Conclusion

This paper shows that the standard estimation approach for RD designs suffers from pooling bias and increased variance if groups are heterogeneous. To prevent additional bias due to pooling of heterogeneous groups, this paper suggests the stratified RD estimator. Since the stratified RD estimator does not suffer from pooling bias, the bias-aware confidence intervals have to be inflated less to guarantee uniform coverage. Additionally, the fully nonparametric covariate adjustment results in a lower asymptotic variance. Lastly, separate bandwidth choices for each group allow the estimator to adapt to heteroskedastic variances between the groups further reducing MSE. Taken together, these effects lead to a meaningful reduction in CI length making statistical inference more precise, whenever groups are heterogeneous.

7 Tables

7.1 Barcellos and Jacobson (2015)

Table 1: Honest CIs with shared BW h

Method	s	Estimate (SE)	Max Bias	t	Honest CI	Length	Rel. Length
Panel A: Group Homogeneity							
Base		-202.06 (55.13)	25.83	0.47	[-83.15, -320.98]	237.83	1.00
Stratified	0	-214.99 (54.60)	25.83	0.47	[-97.03, -332.94]	235.91	0.99
Panel B: Group Heterogeneity							
Stratified	0.1		23.49	0.43	[-98.79, -331.19]	232.40	0.98
Stratified	0.2		21.14	0.39	[-100.43, -329.55]	229.12	0.96
Stratified	0.3		18.79	0.34	[-101.93, -328.04]	226.11	0.95

MSE optimal BW (Base RD) h=10 with effective sample size: 22006. Pooled derivative bound K_PR =1.55

Table 2: Honest CIs with optimized bandwidths h1, h0, hz

Method	s	Estimate (SE)	Max Bias	t	Honest CI	Length	Rel. L	h1	h0	hz
Panel A: Group Homogeneity										
Base		-202.06 (55.13)	25.83	0.47	[-83.15, -320.98]	237.83	1.00	10.00	10.00	
Stratified	0	-216.88 (52.65)	26.67	0.51	[-101.76, -332.00]	230.25	0.97	7.75	10.50	3.5
Panel B: Group Heterogeneity										
Stratified	0.1	-211.47 (51.60)	25.59	0.50	[-99.07, -323.86]	224.79	0.95	8.50	10.75	3.5
Stratified	0.2	-232.35 (51.12)	24.96	0.49	[-121.31, -343.39]	222.09	0.93	8.75	11.25	3.5
Stratified	0.3	-234.37 (48.85)	24.83	0.51	[-127.49, -341.26]	213.77	0.90	9.00	12.00	3.5

Effective sample size: 22006. Pooled derivative bound K_PR =1.55

7.2 Ludwig and Miller (2007)

Smoothness bound $K_P^R = 0.05$

Table 3: Honest CIs with shared BW h

Method	Estimate (SE)	Max Bias	t	Honest CI	Length	Rel. Length
Panel A: Group Homogeneity						
Base	-2.00 (1.04)	0.49	0.47	[0.24, -4.24]	4.48	1.00
Stratified (s = 0)	-1.94 (0.99)	0.49	0.50	[0.21, -4.09]	4.29	0.96
Panel B: Group Heterogeneity						
Stratified (s = 0.1)		0.44	0.45	[0.17, -4.05]	4.22	0.94
Stratified (s = 0.2)		0.40	0.40	[0.14, -4.02]	4.16	0.93
Stratified (s = 0.3)		0.35	0.36	[0.11, -3.99]	4.10	0.91

MSE optimal BW (Base RD): 7.661 with effective sample size: 467

Table 4: Honest CIs with optimized bandwidths h1, h0, hz

Method	Estimate (SE)	Max Bias	t	Honest CI	Length	Rel. L	h1	h0	hz
Panel A: Group Homogeneity									
Base	-2.00 (1.04)	0.49	0.47	[0.24, -4.24]	4.48	1.00	7.66	7.66	NA
Stratified (s = 0)	-1.79 (0.94)	0.50	0.54	[0.29, -3.86]	4.15	0.93	6.96	8.58	4.63
Panel B: Group Heterogeneity									
Stratified (s = 0.1)	-1.84 (0.91)	0.50	0.55	[0.18, -3.87]	4.05	0.90	7.18	9.05	4.63
Stratified (s = 0.2)	-1.61 (0.89)	0.51	0.58	[0.39, -3.61]	4.00	0.89	7.57	9.76	4.63
Stratified (s = 0.3)	-1.49 (0.90)	0.49	0.54	[0.50, -3.47]	3.97	0.89	7.66	10.17	4.63

References

Arai, Yoichi, Taisuke Otsu, and Myung Hwan Seo (Feb. 2025). "Regression Discontinuity Design with Potentially Many Covariates". en. In: *Econometric Theory*, pp. 1–36.

Armstrong, Timothy B. and Michal Kolesár (2018). "Optimal Inference in a Class of Regression Models". en. In: *Econometrica* 86.2, pp. 655–683.

— (2020). "Simple and honest confidence intervals in nonparametric regression". en. In: *Quantitative Economics* 11.1, pp. 1–39.

Barcellos, Silvia Helena and Mireille Jacobson (Nov. 2015). "The Effects of Medicare on Medical Expenditure Risk and Financial Strain". en. In: *American Economic Journal: Economic Policy* 7.4, pp. 41–70.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik (Dec. 2014a). "Robust Data-Driven Inference in the Regression-Discontinuity Design". en. In: *The Stata Journal* 14.4, pp. 909–946.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik (2014b). "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs". en. In: *Econometrica* 82.6, pp. 2295–2326.

Calonico, Sebastian et al. (July 2019). "Regression Discontinuity Designs Using Covariates". In: *The Review of Economics and Statistics* 101.3, pp. 442–451.

Cattaneo, Matias D., Luke Keele, and Rocío Titiunik (2023). "Covariate Adjustment in Regression Discontinuity Designs". In: *Handbook of Matching and Weighting Adjustments for Causal Inference*. Chapman and Hall/CRC.

Chang, Haoge, Joel A. Middleton, and P. M. Aronow (2024). "Exact Bias Correction for Linear Adjustment of Randomized Controlled Trials". en. In: *Econometrica* 92.5, pp. 1503–1519.

Fan, Jianqing and Irène Gijbels (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. New York: Routledge.

Frölich, Markus and Martin Huber (Oct. 2019). "Including Covariates in the Regression Discontinuity Design". In: *Journal of Business & Economic Statistics* 37.4, pp. 736–748.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design". In: *Econometrica* 69.1, pp. 201–209.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Kolesár, Michal and Christoph Rothe (Aug. 2018). "Inference in Regression Discontinuity Designs with a Discrete Running Variable". en. In: *American Economic Review* 108.8, pp. 2277–2304.

Kramer, Patrick and Alexander Kreiß (Oct. 2023). *Estimating Average Treatment Effects in Regression Discontinuity Designs with Covariates under Minimal Assumptions*.

Kreiss, Alexander and Christoph Rothe (May 2023). "Inference in regression discontinuity designs with high-dimensional covariates". en. In: *The Econometrics Journal* 26.2, pp. 105–123.

Li, Ker-Chau (Sept. 1989). "Honest Confidence Regions for Nonparametric Regression". In: *The Annals of Statistics* 17.3, pp. 1001–1008.

Lin, Winston (Mar. 2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique". In: *The Annals of Applied Statistics* 7.1, pp. 295–318.

Ludwig, Jens and Douglas L. Miller (Feb. 2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design*". In: *The Quarterly Journal of Economics* 122.1, pp. 159–208.

Ma, Jun and Zhengfei Yu (May 2024). *Empirical Likelihood Covariate Adjustment for Regression Discontinuity Designs*.

Miratrix, Luke W., Jasjeet S. Sekhon, and Bin Yu (Mar. 2013). "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75.2, pp. 369–396.

Negi, Akanksha and Jeffrey M. Wooldridge (Apr. 2021). "Revisiting regression adjustment in experiments with heterogeneous treatment effects". In: *Econometric Reviews* 40.5, pp. 504–534.

Noack, Claudia, Tomasz Olma, and Christoph Rothe (May 2023). *Flexible Covariate Adjustments in Regression Discontinuity Designs*.

Tsybakov, Alexandre B. (2009). *Introduction to Nonparametric Estimation*. en. Springer Series in Statistics. New York, NY: Springer.

Appendix

This appendix includes assumptions, definitions and proofs for the results presented in the paper. Because the paper is still work in progress some proofs are only available upon request. The theorems are proven for the case of local-linear estimation. But the results generalize to local constant and higher order local polynomial estimation. Proofs are available upon request.

Without loss of generality I set the cut-off $c = 0$.

Notation

- Following the notation in Armstrong and Kolesár (2020) I define the domain of the regression function to be $\mathcal{X} := \mathbb{R}$ for the case of interior point/RD estimation and $\mathcal{X} := [0, \infty)$ for the case of a (left) boundary point.

Bounding Constants

- Constants using M denote bounds on the p-th derivatives of a regression function of the otucome variable Y_i.
- M_P is a bound on the p-th derivative of the pooled regression function
- M_L is a bound on the p-th derivative of the long regression function. When necessary $M_{L,j}$ denotes the bound on the p-th derivative of the long regression function for group j. If not stated otherwise it is assumed that $M_{L,j} = M_L, \quad \forall j$
- Constants using C denote bounds on group heterogeneity
- This heterogeneity is always a difference in the baseline regression function $\Delta m(x) := m(x,1) - m(x,0)$ and the treatment effect regression function $\Delta\tau(x) := \tau(x,1) - \tau(x,0)$
- $C_{\Delta m}$ is a bound on the difference in the baseline regression function between groups, i.e., $|\Delta m(x)| \leq C_{\Delta m}, \quad \forall x \in \mathcal{X}$
- $C_{\Delta\tau}$ is a bound on the difference in the treatment effect regression function between groups, i.e., $|\Delta\tau(x)| \leq C_{\Delta\tau}, \quad \forall x \in \mathcal{X}$
- In some parts of the paper it will also be necessary to apply smoothness assumption on the p-th derivative of the difference in regression functions $\Delta m(x), \Delta\tau(x)$. For

example $\Delta m(x) \in \mathcal{F}_p(C_{\Delta m,p})$ assumes $\Delta m(x)$ to lie in a p-th order smoothness class with constant $C_{\Delta m,p}$. This constant gives a uniform upper bound on the p-th derivative of the difference in regression functions, i.e. $|\Delta m^{(p)}(x)| \leq C_{\Delta m,p} \quad \forall \in \mathcal{X}$.

- L_p denotes the bound on the p-th derivative of the group regression function $g(x)$. Since, for higher smoothness order (e.g. $p > 1$) bounds on all the p derivatives will be considered I explicitly denote the order as a subscript of the bound whenever necessary.

Functionals and Worst Case Bias

I summarize the main notation used in the paper for functionals and their supremum over the Hölder classes.

Exact Finite Sample Bias

- Formally, the expectation of the base RD estimator $\mathbb{E}_{m,m+\tau}[\hat{\tau}(h) - \tau | X]$ is taken under the DGP determined by the regression functions $m(x)$ and $\tau(x)$ for the pooled model. The expectation of the within-group RD estimators

$$\mathbb{E}_{m,m+\tau}[\hat{\tau}_j(h) - \tau \mid X]$$

  is taken under the DGP given the long regression functions $m(x,z), \tau(x,z)$. For notational simplicity, I suppress the explicit subscripts $m, m+\tau$ and write $\mathbb{E}[\cdot]$ throughout.

- For a given estimator $\hat{\tau}(h)$, the bias, conditional on the realizations of the running variable, is a linear functional of the regression functions, $Bias_X : \mathcal{F} \to \mathbb{R}$ and is defined as,

$$Bias_X(\hat{\tau}(h); m, m+\tau) := \mathbb{E}[\hat{\tau}(h) - \tau \mid X],$$

  where $m(x)$ and $\tau(x)$ are the pooled baseline and treatment effect regression functions respectively. Similarly I denote the bias functional for the long regression functions $m(x,z), m(x,z) + \tau(x,z)$ and the within-group estimator for group j as

$$Bias_{X,Z}(\hat{\tau}_j(h); m, m+\tau) := \mathbb{E}[\hat{\tau}_j(h) - \tau \mid X, Z].$$

- Asymptotic Worst-Case Bias

$$\overline{Bias}(\hat{\tau}(h)) := \sup_{m,\, m+\tau \in \mathcal{F}} \Big| B_X(\hat{\tau}(h); m, m+\tau) \Big|,$$

$$\overline{Bias}(\hat{\tau}_j(h)) := \sup_{m,\, m+\tau \in \mathcal{F}} \Big| B_{X,Z}(\hat{\tau}_j(h); m, m+\tau) \Big|.$$

Worst-Case Bias of Leading Terms in Linearization

$$\overline{Bias}(\hat{\tau}_S(h)) := \sup_{\mathcal{F}_{Strat}} \Big| LB \Big|$$

Remark 2 (Asymptotic leading bias)**.** *As $n \to \infty$ with $h \to 0$, and under a smooth running variable distribution $f_X(x)$, the finite-sample conditional bias is well-approximated by the asymptotic expansion given in Armstrong and Kolesár (2020) and does not depend on the realized X_i. Consequently, in the asymptotic worst-case bias, the subscript X can be omitted.*

A Pooled and Long Regression Models

A.I Pooled Regression Function

$$Y_i = m(X_i) + \tau(X_i)T_i + u_i$$

with $m(X_i) := E[Y_i(0)|X_i = x]$, $\tau(X_i) := E[Y_i(1) - Y_i(0)|X_i]$ and $E[u_i|X_i] = 0$ by CEF property, $\sigma_P^2(X_i) = V(u_i|X_i)$

A.II Long Regression Function

$$\begin{aligned} Y_i &= m(X_i, Z_i) + \tau(X_i, Z_i)T_i + e_i \\ Y_i &= [m(X_i, 1) + \tau(X_i, 1)T_i]Z_i + [m(X_i, 0) + \tau(X_i, 0)T_i](1 - Z_i) + e_i \end{aligned} \tag{28}$$

with $m(X_i, z) := E[Y_i|X_i, Z_i = z]$, $\tau(X_i, z) := E[Y_i(1) - Y_i(0)|X_i, Z_i = z]$ and $E[e_i|X_i, Z_i] = 0$ by CEF property, $\sigma_L^2(X_i, Z_i) = V(e_i|X_i, Z_i)$, $\sigma_L^2(X_i) = V(e_i|X_i) = E[e_i^2|X_i] = E[E[e_i^2|X_i, Z_i]|X_i = x] = \sum_{z\in(0,1)} \sigma_L^2(x, z) f_{Z|X}(z|x)$

A.III Relating Error Terms of the Pooled and Long Regression Function

by rearranging equation 28 we get the following relation between the pooled and long regression model.

$$\begin{aligned} Y_i &= [m(X_i, 1) + \tau(X_i, 1)T_i]Z_i + [m(X_i, 0) + \tau(X_i, 0)T_i](1 - Z_i) + e_i \\ &= m(X_i) + \tau(X_i)T_i + ([m(X_i, 1) + \tau(X_i, 1)T_i]Z_i + [m(X_i, 0) + \tau(X_i, 0)T_i](1 - Z_i) - [m(X_i) + \tau(X_i)T_i]) \\ &= m(X_i) + \tau(X_i)T_i + \underbrace{([m(X_i, 1) - m(X_i, 0)] + [\tau(X_i, 1) - \tau(X_i, 0)]T_i)\,(Z_i - f_{Z|X}(1|x)) + e_i}_{u_i} \\ &= m(X_i) + \tau(X_i)T_i + \underbrace{(\Delta m(X_i) + \Delta\tau(X_i)T_i)\,r_i + e_i}_{u_i} \\ &= m(X_i) + \tau(X_i)T_i + u_i \end{aligned}$$

where $u_i = ([m(X_i, 1) - m(X_i, 0)] + [\tau(X_i, 1) - \tau(X_i, 0)T_i])\,(Z_i - f_{Z|X}(1|x)) + e_i$ is the error in the pooled regression model. Note that the law of total variance implies $V[u_i|X_i] \geq$

$E[V[e_i|X_i, Z_i]|X_i]$, with strict inequality if $[m(X_i, 1) - m(X_i, 0)] + [\tau(X_i, 1) - \tau(X_i, 0)]T_i > 0$ and $V(Z_i|X_i) > 0$.

A.IV Variance Comparison

Lemma A.1 (Pooled and Long Variance Comparison)**.** *Define* $\Delta m(X_i) := [m(X_i, 1) - m(X_i, 0)]$ *and* $\Delta\tau(X_i) := [\tau(X_i, 1) - \tau(X_i, 0)]$*, then*

$$\sigma_P^2(X_i) = [\Delta m(X_i) + \Delta\tau(X_i)T_i]^2\sigma_Z^2(X_i) + \sigma_L^2(X_i)$$

Proof. Note that $T_i := \mathbb{1}\{X_i \geq c\}$ is just an indicator function of X_i.

$$\begin{aligned}
\sigma_P^2(X_i) :=& Var(u_i|X_i) \\
=& E[u_i^2|X_i] \\
=& [\Delta m(X_i) + \Delta\tau(X_i)T_i]^2 E[r_i^2|X_i] + E[e_i^2|X_i] \\
=& [\Delta m(X_i) + \Delta\tau(X_i)T_i]^2 E[r_i^2|X_i] + E[E[e_i^2|X_i, Z_i]|X_i] \\
=& [\Delta m(X_i) + \Delta\tau(X_i)T_i]^2 E[r_i^2|X_i] + E[\sigma^2(X_i, Z_i)|X_i] \\
=& [\Delta m(X_i) + \Delta\tau(X_i)T_i]^2\sigma_Z^2(X_i) + \sigma_L^2(X_i)
\end{aligned}$$

Line 3 follows because after squaring $u_i^2 = ((\Delta m(X_i) + \Delta\tau(X_i)T_i)r_i + e_i)^2$ the cross-terms cancel out. That is because of the CEF property $E[e_i r_i|X_i] = E[r_i \underbrace{E[e_i|X_i, Z_i]}_{=0}|X_i] = 0$. □

A.V Group Variable Regression Function

I assume, as usual in the covariate adjustment RD literature, that the covariates are pre-determined in the sense that the treatment has no effect on the covariate distribution. This gives:

$$Z_i = g(X_i) + r_i$$

with $E[r_i|X_i] = 0$ by the CEF property.

B General Assumptions and Definitions

Assumption 1 (Predetermined covariates)**.** *The covariates are pre-determined, i.e., realized before treatment such that their values are invariant to treatment assignment*

$$Z_i(1) = Z_i(0) \quad \forall i.$$

Remark 3. *The unit-level formulation of Assumption 1 implies that the distribution of covariates is invariant to treatment, $Z_i(0) \overset{d}{=} Z_i(1)$, as assumed in Calonico et al. (2019).*

Assumption 2 (Kernel Function)**.** *Assume that the kernel function k is bounded with bounded support.*

Assumption 3 (Bandwidths)**.** *Assume the bandwidth sequence to satisfy $0 < \liminf_n h_n(nM^2)^{1/(2p+1)} < \limsup_n h_n(nM^2)^{1/(2p+1)} < \infty$ as in assumption 3.1 in Armstrong and Kolesár (2020). Further, assume the bandwidth choices h_z and h to be proportional, $h_n \asymp h_n^Z$. Specifically, assume the bandwidths to satisfy $h_z = \kappa h$, for an appropriate $\kappa > 0$.[6] The notation is as in Calonico et al. (2014a).*

B.I Assumptions for the Base RD Estimator

Assumption 4 (Marginal Density)**.** *Assume the marginal (pooled) density $f_X(x)$ is continuous and bounded away from 0 and infinity*

Assumption 5 (Regularity conditions on the pooled error terms)**.** *The random variables $\{u_i\}_{i=1}^n$ are independent with $E[|u_i|^{2+\eta}|X_i] \leq \frac{1}{\eta}$, for some $\eta > 0$. Denote the variance by $Var(u_i|X_i) := \sigma_P^2(X_i)$ and assume that the one-sided limits at the cutoff exist and are strictly positive*

$$\sigma_+^2(0) = \lim_{x \downarrow 0} \sigma_P^2(x) > 0, \quad \sigma_-^2(0) = \lim_{x \uparrow 0} \sigma_P^2(x) > 0$$

B.II Assumptions for the Within-group and Stratified RD Estimators

Assumption 6 (Joint Density)**.** *Assume the joint density $f_{X,Z}(x, z)$ is continuous in x and bounded away from 0 and infinity*

Assumption 7 (Regularity conditions on the long error terms)**.** *The random variables $\{e_i\}_{i=1}^n$ are independent with $E[|e_i|^{2+\eta}|X_i, Z_i] \leq \frac{1}{\eta}$, for some $\eta > 0$. Denote the variance*

[6] This is necessary to get valid comparisons between the two estimators

by $Var(e_i|X_i, Z_i) := \sigma_L^2(X_i, Z_i)$ *and assume that the one-sided limits at the cutoff exist and are strictly positive*

$$\sigma_+^2(0, z) \;=\; \lim_{x \downarrow 0} \sigma_L^2(x, z) \;>\; 0, \quad \sigma_-^2(0, z) \;=\; \lim_{x \uparrow 0} \sigma_L^2(x, z) \;>\; 0 \quad \textit{for } z \in \{0, 1\}$$

Assumption 8 (Regularity conditions on the group-indicator error terms)**.** *The random variables* $\{r_i\}_{i=1}^n$ *are independent. Note that* $E[|r_i|^{2+\eta}|X_i] \leq \frac{1}{\eta}$*, for some* $\eta > 0$*, holds by the fact that* Z_i *is binary, which implies* $|r_i| \leq 1$*. Denote the variance by* $Var(r_i|X_i) := \sigma_Z^2(X_i)$ *and assume that the one-sided limits at the cutoff exist and are strictly positive*

$$\sigma_{Z,+}^2(0) \;=\; \lim_{x \downarrow 0} \sigma_Z^2(x) \;>\; 0, \quad \sigma_{Z,-}^2(0) \;=\; \lim_{x \uparrow 0} \sigma_Z^2(x) \;>\; 0$$

B.III Definitions

I define the estimators for the right-hand side estimation using the kernel $k_+(v) := k(v)\mathbb{1}_{\{v\geq 0\}}$. The respective expressions for the left-hand side estimation are defined quivalently using $k_-(v) := k(v)\mathbb{1}_{\{v\leq 0\}}$. Let $m_q(u) := (1, u, ..., u^q)$, denote a polynomial of u. To simplify notation I supress the chosen polynomial order q for the local polynomial estimator. It is implicitly defined in the weights..

The definition for left boundary points follows analogously

Definition 5 (General Equivalent Kernel)**.** *For estimation of an interior point the domain of integration is $\mathcal{X} := \mathbb{R}$. For estimation of a right boundary point the domain of integration is $\mathcal{X} := [0, \infty)$.*

$$k_q^*(u) = e_1' \left(\int_{\mathcal{X}} m_q(t) m_q(t)' k(t)\, dt \right)^{-1} m_q(u) k(u), \tag{29}$$

Definition 6 (Base RD Local Polynomial Estimator)**.** *The local polynomial estimator takes the form $\hat{\tau}(h) = \frac{1}{nh}\sum_{i=1}^n \tilde{k}_n(X_i/h)$ given in Armstrong and Kolesár (2020) with $\tilde{k}_n(X_i/h) = \tilde{k}_{n,+}(X_i/h) - \tilde{k}_{n,-}(X_i/h)$*

$$\begin{aligned}
\tilde{k}_{n,+}(u) =& e_1^\top \left(\frac{1}{nh} \sum_{i=1}^n k_+(x_i/h) m_q(x_i/h) m_q(x_i/h)' \right)^{-1} m_q(u) k_+(u) && (30)\\
:=& e_1^\top (\tilde{Q}_{n,+})^{-1} m_q(u) k_+(u) && (31)
\end{aligned}$$

$\tilde{k}_{n,-}$ is defined similarly. If k is bounded with bounded support, then under Assumption 3.1 this sequence satisfies Assumption B.4 with

$$\begin{aligned}
\tilde{k}_+(u) =& e_1^\top \left(f(0) \int_{\mathcal{X}} k_+(t) m_q(t) m_q(t)'\, dt \right)^{-1} m_q(u) k_+(u) && (32)\\
:=& e^\top (\tilde{Q}_+)^{-1} m_q(u) k_+(u) && (33)\\
=& \frac{1}{f(0)} k_q^*(u) \mathbb{1}_{\{u\geq 0\}} && (34)
\end{aligned}$$

where $k_+^*(u)$ is the equivalent kernel as defined above. Theorem 3.1 and equation (15) in Armstrong and Kolesár (2020) then follow immediately by applying Theorem B.4 with this choice of $\tilde{k}_n$ and $\tilde{k}$.

For simplicity, I restrict the analysis to local polynomial estimators, with equal bandwidth choices on both sides of the cut-off, and a kernel that is bounded and has bounded

support (as done in Armstrong and Kolesár (2020)).

The local polynomial estimator with a bounded kernel implies that assumption B.4 hold (Armstrong and Kolesár 2020). Assuming lemma D.1 (or lemma D.2) allows the application of the results in theorem B.4.

Definition 7 (Within-Group Local Polynomial Estimator)**.** *For the groupwise case we have an additional sample restriction yielding:*

$$\begin{aligned}\tilde{k}_{j,n,+}(u,z) =& e_1' \left(\underbrace{\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i=j\}} k_+(x_i/h) m_q(x_i/h) m_q(x_i/h)'}_{:=\tilde{Q}_{n,j,+}} \right)^{-1} m_q(u) k_+(u) \mathbb{1}_{\{z=j\}} \\ =& e_1' (\tilde{Q}_{n,j,+})^{-1} m_q(u) k_+(u) \mathbb{1}_{\{z=j\}}\end{aligned}$$

with the repsective equivalent kernel:

$$\begin{aligned}\tilde{k}_{j,+}(u,z) =& e_1' \left(f_{X,Z}(0,z) \int_{\mathcal{X}} k_+(t) m_q(t) m_q(t)' dt \right)^{-1} m_q(u) k_+(u) \mathbb{1}_{\{z=j\}} \\ :=& e_1' \left(\tilde{Q}_{j,+} \right)^{-1} m_q(u) k_+(u) \mathbb{1}_{\{z=j\}} \\ =& \frac{1}{f_{X,Z}(0,j)} k_q^*(u) \mathbb{1}_{\{u\geq 0\}} \mathbb{1}_{\{z=j\}}\end{aligned}$$

C Smoothness Classes

Following Armstrong and Kolesár (2020), let $\mathcal{F}_p(M)$ denote the closure of the family of p-times differentiable functions with the p-th derivative bounded by M uniformly over the domain of the function.

Definition 8 (Hölder Class of Order p)**.**

$$\mathcal{F}_p(M) := \{f(x) \quad : \quad |f^{(p-1)}(x) - f^{(p-1)}(y)| \leq M|x-y| \quad \forall x, y\}$$

Note for $p = 1$ this becomes a lipschitz class.

C.I Smoothness Classes for Base RD and Stratified RD

Definition 9 (Sharp RD Smoothness Class)**.**

$$\mathcal{F}_p^{Base}(M_P) := \left\{ \begin{array}{l} m(x) + \tau(x)\mathbb{1}_{\{x\geq 0\}}, \\ m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq 0\}}, \\ g(x) \end{array} \middle| \, m(x), m(x) + \tau(x) \in \mathcal{F}_p(M_P) \right\}$$

Definition 10 (Stratified RD Smoothness Class)**.**

$$\mathcal{F}_p^{Strat}(\mathbf{M}) := \left\{ \begin{array}{l} m(x) + \tau(x)\mathbb{1}_{\{x\geq 0\}}, \\ m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq 0\}}, \\ g(x) \end{array} \middle| \begin{array}{l} m(x,z), m(x,z) + \tau(x,z) \in \mathcal{F}_p(M_L), \\ g_z(x) \in \mathcal{F}_p(L), \\ |\tau(x,1) - \tau(x,0)| \leq C_{\Delta\tau}, \forall x \end{array} \right\}$$

with $\mathbf{M} := (M_L, M_z'', M_{\Delta\tau})$ *capturing all relevant smoothness constants.*

C.II Restricted Hoelder Smoothness Class

The estimators analyzed in this project all require different smoothness classes, which are not naturally comparable.

By noting that $m(x), \tau(x)$ are pooled conditional expectation functions, one can apply the law of iterated expectations. This gives several additive terms that all contribute to the roughness of $m(x)$. In the spirit of a worst case bias analysis I bound these terms using the triangle inequality to ensure that they do not cancel out.

Definition 11 (Restricted Hölder Class)**.** *Let* $m(x) = \sum_{z\in\{0,1\}} m(x,z)g_z(x)$ *be a pooled CEF. Further, let* $\vec{M} = \{M''_m, M''_\tau, M_{\Delta m}, M'_{\Delta m}, M_{\Delta\tau}, M'_{\Delta\tau}, M'_Z, M''_Z\}$

$$\mathcal{F}^R(\vec{M}) = \left\{ m(x) + \tau(x)\mathbb{1}_{\{x\geq 0\}}, m(x,z) + \tau(x,z)\mathbb{1}_{\{x\geq 0\}}, g_1(x) \;\middle|\; \begin{array}{l} m(x,z) \in \mathcal{F}_2(M''_m), \\ \tau(x,z) \in \mathcal{F}_2(M''_\tau), \\ g_z(x) \in \mathcal{F}_1(M'_Z), \\ g_z(x) \in \mathcal{F}_2(M''_Z), \\ m(x,1) - m(x,0) \leq M_{\Delta m}, \forall x \\ \tau(x,1) - \tau(x,0) \leq M_{\Delta\tau}, \forall x \\ m(x,1) - m(x,0) \in \mathcal{F}_1(M'_{\Delta m}) \\ \tau(x,1) - \tau(x,0) \in \mathcal{F}_1(M'_{\Delta\tau}) \end{array} \right\}$$

D Estimators and Asymptotic Properties

This section reviews and presents the asymptotic properties of the Base-RD estimator, the group-wise RD estimator, and two variants of the stratified-RD estimator, one assuming the group-weights are known (oracle) and a feasible version that estimates the group weights.

Throughout the paper we assume assumption 2 and assumption 3. Further, for the Base-RD estimator we additionally assume assumption 4 and assumption 5. Denote the worst case bias over the function class $\mathcal{F}$, conditional on the realizations of the running variable $\mathbf{X}$, as $\bar{B}_X(\hat{\tau}(h)) := \sup_{\mathcal{F}} E[\hat{\tau}(h)|\mathbf{X}] - \tau$:

D.I Base-RD

Denote $u_i = \frac{X_i}{h}$, further following Armstrong and Kolesár (2020) and $k_+(v) = k(v)\mathbb{I}\{v \geq 0\}$, then we get the following definition for the gram matrix:

$$\tilde{Q}_{n,+} := \frac{1}{nh}\sum_{i=1}^{n} k_+(u_i) m_q(u_i) m_q(u_i)^\top \tag{35}$$

which equals exactly the expression $\Gamma_{+,1}(h)$ in Calonico et al. (2014a). Then the local polynomial weights are given by $\tilde{k}_{n,+}(v) := e_1^\top (\tilde{Q}_{n,+})^{-1} m_q(v) k_+(v)$. The kernel weight, gram matrix and weights are defined similarly for the control side. The Base-RD estimator then takes the form:

$$\hat{\tau}(h) = \frac{1}{nh}\sum_{i=1}^{n} [\tilde{k}_{n,+}(X_i/h) - \tilde{k}_{n,-}(X_i/h)] Y_i \tag{36}$$

$$:= \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_n(X_i/h) Y_i \tag{37}$$

with $\tilde{k}_n(v) := [\tilde{k}_{n,+}(v) - \tilde{k}_{n,-}(v)]$

Lemma D.1. *Assume assumption 4 and that the function $g(\cdot)$ is bounded and has bounded support $\mathcal{S}$. A probabilistic version of assumption 3.1 in Armstrong and Kolesár (2020) holds, as long as $n \to \infty$, $h \to 0$ and $nh \to \infty$*

$$\operatorname*{plim}_{h\to 0}\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right) = f(0)\int_{\mathcal{S}} g(u)du = d\int_{\mathcal{S}} g(u)du$$

Proof. First rewrite

$$\begin{aligned}
\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right) &= E\left[\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right] + \frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right) - E\left[\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right] \\
&= E\left[\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right] + O_p\left(\sqrt{V\left(\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right)}\right) \\
&= E\left[\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right] + o_P(1)
\end{aligned}$$

Since

$$V\left(\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right) = \frac{1}{nh^2}V\left(g\left(\frac{X_i}{h}\right)\right) = \frac{1}{nh}\frac{1}{h}V\left(g\left(\frac{X_i}{h}\right)\right) = \frac{1}{nh}O(1) = o(1)$$

(since nh diverges to infinity)

because $V\left(g\left(\frac{X_i}{h}\right)\right) = E\left[g\left(\frac{X_i}{h}\right)^2\right] - E\left[g\left(\frac{X_i}{h}\right)\right]^2 = hO(1) + h^2O(1) = hO(1)$, and:

$$\begin{aligned}
E\left[g\left(\frac{X_i}{h}\right)^2\right] &= \int_{\mathbb{R}} g\left(\frac{x}{h}\right)^2 f_X(x)dx \\
&= \int_{\mathcal{S}} g\left(\frac{x}{h}\right)^2 f_X(x)dx && \text{(Compact Support)} \\
&= h\int_{\mathcal{S}} g\left(u\right)^2 f_X(uh)du && \text{(Change of Variable)} \\
&= h\int_{\mathcal{S}} g\left(u\right)^2 f(0)du + h\int_{\mathcal{S}} g\left(u\right)^2 \left[f_X(uh) - f(0)\right]du \\
&= h\int_{\mathcal{S}} g\left(u\right)^2 f(0)du + ho(1) \\
&= hO(1)
\end{aligned}$$

where the fourth to fifth line follows by using the fact that boundedness of $g(\cdot), f_X(\cdot)$

and the compactness of the area of integration allows the application of the dominated convergence theorem to interchange limits and integration. Finally the continuity of $f_X(\cdot)$ implies that $\lim_{h\to 0}[f_X(uh) - f_X(0)] = 0$. The convergence of $E\left[g\left(\frac{X_i}{h}\right)\right]^2 = h^2 O(1)$ follows a similar argument.

Further,

$$\begin{aligned} E\left[\frac{1}{nh}\sum_{i=1}^{n} g\left(\frac{X_i}{h}\right)\right] &= \frac{1}{h}\int_{\mathbb{R}} g\left(\frac{X_i}{h}\right) f(x)dx \\ &= \int_{\mathcal{S}} g(u)f(uh)du \qquad \text{(Change of Variable, Compact Support)} \\ &= f(0)\int_{\mathcal{S}} g(u)du + \int_{\mathcal{S}} g(u)[f(uh) - f(0)]du \\ &= f(0)\int_{\mathcal{S}} g(u)du + o(1) \end{aligned}$$

where the last line follows by assuming continuity and boundedness of $f(x)$ and applying the dominated convergence theorem to change order of integration and limit.

□

Theorem 1 (Asymptotic Worst-Case MSE - Base-RD (Armstrong and Kolesár 2020))**.** *Assume assumption 2 to assumption 5. Additionally require the following smoothness assumption on the regression function, $m(x), m(x) + \tau(x) \in \mathcal{F}_p(M_P)$.*

Then

$$V[\hat{\tau}(h)|\mathbf{X}] = \frac{\mathcal{V}_B}{nh}(1 + o(1)) \tag{11}$$

$$\overline{\text{Bias}}(\hat{\tau}(h)) = h^p \frac{2M_P}{p!} B_{p,q}(k)(1 + o(1)) \tag{12}$$

with $\mathcal{V}_B = \frac{\overline{k_q^2}\left(\sigma^2_{P,+}(c) + \sigma^2_{P,-}(c)\right)}{f_X(c)}$, $\overline{k_q^2} = \int_0^\infty k_q^*(u)^2 du$ and $B_{p,q}(k)$ kernel constants tabulated in table S.1 of the online appendix in Armstrong and Kolesár (2020).

Proof. Since by lemma D.1 assumptions 3.1 and 3.2 hold by hypothesis the result follows by the arguments in Armstrong and Kolesár (2020).

□

D.II Within-group RD estimator

The within-group RD estimator is the base RD estimator applied to the subsample restricted to a given group and is defined as follows. For the within-group estimation we get the following gram matrices, that include an indicator to restrict the sum to within-group units.

$$\tilde{Q}_{n,j,+} := \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i=j\}} k_+(u_i) m_q(u_i) m_q(u_i)^\top \tag{38}$$

Then the local polynomial weights are given by $\tilde{k}_{n,j,+}(v,z) := e_1^\top(\tilde{Q}_{n,j,+})^{-1} m_1(v) k_+(v) \mathbb{1}_{\{z=j\}}$. The kernel weight, gram matrix and weights are defined similarly for the control side. The within-group RD estimator then takes the form:

$$\hat{\tau}_j(h) = \frac{1}{nh}\sum_{i=1}^{n} [\tilde{k}_{n,j,+}(X_i/h, Z_i) - \tilde{k}_{n,j,-}(X_i/h, Z_i)] Y_i \tag{39}$$

$$= \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_{n,j}(X_i/h, Z_i) Y_i \tag{40}$$

with $\tilde{k}_{n,j}(v) := [\tilde{k}_{n,j,+}(v) - \tilde{k}_{n,j,-}(v)]$ as before.

To make the within-group and stratified estimation procedures well defined, assume assumption 6 and assumption 7. Further, denote the worst case bias $\bar{B}_{X,Z}(\hat{\tau}_z(h)) := \sup_{\mathcal{F}} E[\hat{\tau}_z(h)|\mathbf{X},\mathbf{Z}] - \tau$, which now additionally conditions on the sample realization of Z to make the conditional expectation well-defined.

Lemma D.2 (Groupwise Assumption 3.1)**.** *Assume assumption 6 and that the function $g(\cdot)$ is bounded and has bounded support $\mathcal{S}$, a probabilistic version of assumption 3.1 in Armstrong and Kolesár (2020) holds, as long as $n \to \infty$, $h \to 0$ and $nh \to \infty$*

$$\frac{1}{nh}\sum_{i=1}^{n} g(X_i/h)\mathbb{1}_{(Z_i=z)} \xrightarrow{P} d\int_{\mathcal{X}} g(u)du$$

with $d = f_{X,Z}(c,z)$

Proof. Analogous to the proof of lemma D.1. □

Lemma D.3 (Groupwise-RD Asymptotic Properties)**.** *Assume the regression function* $E[Y_i|X_i = x, Z_i = z] \in \mathcal{F}_{Strat}(M_L), \quad \forall z \in \{0,1\}$. *Then*

$$V[\hat{\tau}_z(h)|\mathbf{X},\mathbf{Z}] = \frac{\mathcal{V}_z}{nh}(1+o(1)), \tag{41}$$

$$\mathcal{V}_z = \frac{\overline{k_q^2}\left(\sigma^2_{L,+}(c,z) + \sigma^2_{L,-}(c,z)\right)}{f_{X,Z}(c,z)}, \tag{42}$$

$$\overline{\text{Bias}}(\hat{\tau}_z(h)) = h^p M_L \frac{2}{p!} B_{p,q}(k)(1+o(1)), \tag{43}$$

Proof. Lemma D.2 implies assumption 3.1 under assumed conditions. The result then follows as argued in Armstrong and Kolesár (2020). □

D.III Stratified RD (Oracle)

The asymptotic properties of the oracle version of the stratified estimator, $\hat{\tau}_O(h) = \sum_z \hat{\tau}_z(h) g_z(c)$, follow immediately after characterizing the properties of $\hat{\tau}_z(h)$ and then taking the corresponding weighted average. Note that $\hat{\tau}_z(h)$ are estimated on disjoint subsamples and are independent conditional on the sample realization of regressors $(\mathbf{X},\mathbf{Z})$.

Lemma D.4 (Stratified RD (Oracle) - Asymptotic Properties)**.** *Assume the regression function*
$E[Y_i|X_i = x, Z_i = z] \in \mathcal{F}_{Strat}(M_L) \quad \forall z \in \{0,1\}$, *then*

$$V[\hat{\tau}_O(h)|\mathbf{X},\mathbf{Z}] = \frac{\mathcal{V}_O}{nh}(1+o(1)), \tag{44}$$

$$\begin{aligned}\mathcal{V}_O =& \frac{\overline{k_q^2}(k)\left(E[\sigma^2_{L,+}(0,Z_i) + \sigma^2_{L,-}(0,Z_i)|X_i = c]\right)}{f_X(c)} &(45)\\ :=& \frac{\overline{k_q^2}(k)\left(\sigma^2_{L,+}(0) + \sigma^2_{L,-}(0)\right)}{f_X(c)} &(46)\end{aligned}$$

$$\begin{aligned}\overline{\text{Bias}}(\hat{\tau}_O(h)) &= g_1(c)\overline{\text{Bias}}(\hat{\tau}_1) + g_0(c)\overline{\text{Bias}}(\hat{\tau}_0) &(47)\\ &= h^p M_L \frac{2}{p!} B_{p,q}(1+o(1)) &(48)\end{aligned}$$

with kernel constants as before.

Proof. Equation 47 follows by the linearity of the expectation operator, the results for the groupwise-RD estimator, the symmetry of the function class, and by the fact that $g_1(c) + g_0(c) = 1$.

Equation 44 follows by the fact that $f_{X,Z}(x,z) = f_{Z|X}(x|z)f_X x$, hence $\frac{g_1(c)^2}{f_{X,Z}(x,z)} = \frac{f_{Z|X}(z|x)^2}{f_{Z|X}(z|x)f_X(x)} = \frac{f_{Z|X}(z|x)}{f_X(x)}$. Plugging everything in yields,

$$
\begin{aligned}
V[\hat{\tau}_O|\mathbf{X},\mathbf{Z}] &= V[g_1(c)\hat{\tau}_1 + g_0(c)\hat{\tau}_0|\mathbf{X},\mathbf{Z}] \\
&= g_1(c)^2 V[\hat{\tau}_1|\mathbf{X},\mathbf{Z}] + g_0(c)^2 V[\hat{\tau}_0|\mathbf{X},\mathbf{Z}] && (\ \hat{\tau}_1 \perp\!\!\!\perp \hat{\tau}_0|\mathbf{X},\mathbf{Z}) \\
&= \frac{1}{nh}\left(g_1(c)^2 \frac{\overline{k_q^2}\left(\sigma_+^2(c,1) + \sigma_-^2(c,1)\right)}{f_{X,Z}(c,1)} + g_0(c)^2 \frac{\overline{k_q^2}\left(\sigma_+^2(c,0) + \sigma_-^2(c,0)\right)}{f_{X,Z}(c,0)} + o(1) \right) \\
& && \text{(by lemma D.3)} \\
&= \frac{1}{nh}\left(f_{Z|X}(1|c)^2 \frac{\overline{k_q^2}\left(\sigma_+^2(c,1) + \sigma_-^2(c,1)\right)}{f_{Z|c}(1|x)f_X(c)} + f_{Z|X}(0|c)^2 \frac{\overline{k_q^2}\left(\sigma_+^2(c,0) + \sigma_-^2(c,0)\right)}{f_{Z|X}(0|c)f_X(c)} + o(1) \right) \\
&= \frac{1}{nh}\left(f_{Z|X}(1|x) \frac{\overline{k_q^2}\left(\sigma_+^2(c,1) + \sigma_-^2(c,1)\right)}{f_X(c)} + f_{Z|X}(0|x) \frac{\overline{k_q^2}\left(\sigma_+^2(c,0) + \sigma_-^2(c,0)\right)}{f_X(c)} + o(1) \right) \\
&= \frac{1}{nh}\left(\frac{\overline{k_q^2} E\left[\left(\sigma_+^2(c,Z_i) + \sigma_-^2(c,Z_i)\right)|X_i = c\right]}{f_X(c)} + o(1) \right)
\end{aligned}
$$

where the last line follows by the law of total expectation □

D.IV Weight Estimation

The weight estimator $\hat{g}(h_z)$ is a local polynomial estimator of order q at the interior point c using Z_i as outcome and bandwidth h_z. The definition and results directly follows from theorem 3.1 in Armstrong and Kolesár (2020).

Denote the estimator as,

$$\hat{g}(h_z) := \frac{1}{nh_z}\sum_{i=1}^{n} \tilde{k}_n^z(X_i/h_z)Z_i$$

Lemma D.5 (Asymptotic Properties of the Weight Estimator)**.** *Assume assumption 2,*

assumption 3, assumption 8, and assumption 6. Then

$$V[\hat{g}(h_z)|\mathbf{X}] = \frac{\mathcal{V}_g}{nh_z}(1+o(1)), \tag{49}$$

$$\overline{\text{Bias}}(\hat{g}(h_z)) = h_z^p \frac{L}{p!} B_{p,q}^{int}(k)(1+o(1)), \tag{50}$$

where

$$\mathcal{V}_g = \frac{\widetilde{k}_q^2(k)\, g(c)(1-g(c))}{f_X(c)}, \qquad \widetilde{k}_q^2(k) = \int_{-\infty}^{\infty} k_q^*(u)^2 \,\mathrm{d}u, \tag{51}$$

and $B_{p,q}^{int}(k)$ is the interior point analogue of the bias kernel constant, with L bounding the p-th derivative of $g(x)$ and $\widetilde{k}_q^2(k)$ the interior point analogue of the variance kernel constant.

D.V Stratified-RD (feasible)

Let $\hat{\tau}_S(h, h_z) = \hat{\tau}_1(h)\hat{g}_1(h_z) + \hat{\tau}_0(h)(1-\hat{g}_1(h_z))$ be the stratified RD estimator, a weighted average of estimators defined above.

Theorem 2 (Asymptotic Normality of the Stratified RD Estimator)**.** *Assume assumption 2, assumption 3, and assumption 6 to assumption 8. Further assume $m(x,z), m(x,z)+\tau(x,z) \in \mathcal{F}_p(M_L)$, $g_1(x) \in \mathcal{F}_p(L)$ and $|\tau(x,1) - \tau(x,0)| \leq C_{\Delta\tau} \quad \forall x$, then*

$$\sqrt{nh}\,(\hat{\tau}_S(h) - \tau - LB) \xrightarrow{d} N(0, \mathcal{V}_S)$$

where $LB := g_1 E[\hat{\tau}_1(h) - \tau_1|\mathbf{X},\mathbf{Z}] + (1-g_1)E[\hat{\tau}_0(h) - \tau_0|\mathbf{X},\mathbf{Z}] + [\tau_1 - \tau_0]E[\hat{g}_1(h_z) - g_1|\mathbf{X}]$ denotes the leading-order bias terms and the asymptotic variance is given by $\mathcal{V}_S = \mathcal{V}_O + [\tau_1 - \tau_0]^2 \frac{1}{\kappa}\mathcal{V}_g$. With $\mathcal{V}_O$ given in Lemma D.4 and $\mathcal{V}_g$ being the asymptotic variance of the weight estimation using a local linear estimator at an interior point using Z_i as outcome (see section 3.2 in Armstrong and Kolesár (2020)).

Proof. See appendix F.

□

Since the leading bias terms LB are generally unknown we want to upper bound them and derive the worst-case bias, given the assumptions, to construct honest confidence intervals.

Proposition 3 (Worst-Case Bias of the Stratified RD estimator)**.** *Assume assumption 2, assumption 3, and assumption 6 to assumption 8. Further assume* $m(x,z), m(x,z)+\tau(x,z) \in \mathcal{F}_p(M_L)$*,* $g_1(x) \in \mathcal{F}_p(L)$ *and* $|\tau(x,1)-\tau(x,0)| \leq C_{\Delta\tau} \quad \forall x$*, then*

$$\overline{\text{Bias}}(\hat{\tau}_S(h,h_z)) := \sup_{\{m(x,z),m(x,z)+\tau(x,z),g_1(x),\Delta\tau\}} |LB| \tag{15}$$

$$= h^p \frac{2M_L}{p!} B_{p,q}(k) + h_z^p C_{\Delta\tau} \frac{L}{p!} B_{p,q}^{int}(k) + o(h^p) \tag{16}$$

Denote the asymptotic worst case bias $\bar{B}_a(\hat{\tau}_S(h,h_z))$, which puts a worst-case upper bound on the leading terms of the asymptotic bias in theorem 2.

Proof. I use the shorthand notation $\tau_1 - \tau_0 := \tau(c,1) - \tau(c,0) =: \Delta\tau(c)$ and $g_1 := g_1(c)$, that is the evaluation of functionals at point c. First putting everything into explicit functional notation and note that the supremum is using the long regression function objects,

$$
\begin{aligned}
&\sup_{\{m(x,z),m(x,z)+\tau(x,z),g_1(x),\Delta\tau(x))\in\mathcal{F}_{Strat}\}} |LB| \\
=&\sup_{\{m(x,z),m(x,z)+\tau(x,z),g_1(x),\Delta\tau(x))\in\mathcal{F}_{Strat}\}} \\
&\left|g_1 E[\hat{\tau}_1(h)-\tau_1|\mathbf{X},\mathbf{Z}]+(1-g_1)E[\hat{\tau}_0(h)-\tau_0|\mathbf{X},\mathbf{Z}]+[\tau_1-\tau_0]E[\hat{g}_1(h_z)-g_1|\mathbf{X}]\right| \\
:=&\sup_{\{m(x,z),m(x,z)+\tau(x,z),g_1(x),\Delta\tau(x)\ \in\mathcal{F}_{Strat}\}} \\
&\left|g_1 B_{X,Z}(\hat{\tau}_1(h);m,m+\tau)+(1-g_1)B_{X,Z}(\hat{\tau}_0(h);m,m+\tau)+[\tau_1-\tau_0]B_X(\hat{g}(h_z);g)\right| \\
=&\sup_{\{m(x,z),m(x,z)+\tau(x,z)\}}\ \sup_{\{g_1(x)\}} \\
&\quad\sup_{\{|\Delta\tau(x)|\le C_{\Delta\tau}\}}\left|g_1 B_{X,Z}(\hat{\tau}_1(h);m,m+\tau)+(1-g_1)B_{X,Z}(\hat{\tau}_0(h);m,m+\tau)+[\tau_1-\tau_0]B_X(\hat{g}(h_z);g)\right| \\
=&\sup_{\{m(x,z),m(x,z)+\tau(x,z)\}}\ \sup_{\{g_1(x)\}} \\
&\quad\sup_{\{|\tau_1-\tau_0|\le C_{\Delta\tau}\}}\left|g_1 B_{X,Z}(\hat{\tau}_1(h);m,m+\tau)+(1-g_1)B_{X,Z}(\hat{\tau}_0(h);m,m+\tau)+[\tau_1-\tau_0]B_X(\hat{g}(h_z);g)\right| \\
=&\sup_{\{m(x,z),m(x,z)+\tau(x,z)\}} \\
&\sup_{\{g_1(x)\}}\left|g_1 B_{X,Z}(\hat{\tau}_1(h);m,m+\tau)+(1-g_1)B_{X,Z}(\hat{\tau}_0(h);m,m+\tau)|+C_{\Delta\tau}|B_X(\hat{g}(h_z);g)\right| \\
=&\sup_{\{m(x,z),m(x,z)+\tau(x,z)\}}\left|g_1 B_{X,Z}(\hat{\tau}_1(h);m,m+\tau)+(1-g_1)B_{X,Z}(\hat{\tau}_0(h);m,m+\tau)|\right. \\
&+C_{\Delta\tau}\frac{L}{p!}h_z^p B_{p,q}^{int}(k)+o(h^p) \\
=&h^p 2M_L\frac{1}{p!}B_{p,q}(k)+h_z^p C_{\Delta\tau}L\frac{1}{p!}B_{p,q}^{int}(k)+o(h^p)
\end{aligned}
$$

The third line applies Lemma E.1 the principle of iterated suprema , which is valid because the objective is bounded over $\mathcal{F}_{Strat}$. The fourth line replaces $\sup_{\{|\Delta\tau(x)|\le C_{\Delta\tau}\}}$ with $\Delta\tau(c) := \tau_1 - \tau_0$. Since the objective depends on $\Delta\tau(x)$ only through the real-valued evaluation $\Delta\tau(c)$ at the point c, and every value in $[-C_{\Delta\tau}, C_{\Delta\tau}]$ is attainable by some admissible function in $\mathcal{F}_{Strat}$, the inner supremum reduces to the supremum over a real variable over a symmetric and compact set. Likewise the two other suprema consider real variable over a symmetric and compact set. The fifth line then follows after an application of Lemma E.3 and Lemma E.4. The sixth and seventh line follow after plugging in the results from Armstrong and Kolesár (2020). □

E Supremum/Maximum of Real Numbers

Lemma E.1 (Principle of the Iterated Suprema)**.** *Let X and Y be nonempty sets and let $h : X \times Y \to \mathbb{R}$ have bounded range in $\mathbb{R}$. Let $F : X \to \mathbb{R}$ and $G : Y \to \mathbb{R}$ be defined by*

$$F(x) := \sup\{h(x, y) : y \in Y\}, \quad G(y) := \sup\{h(x, y) : x \in X\}$$

*Then the **Principle of the Iterated Suprema** states:*

$$\sup\{h(x, y) : x \in X, y \in Y\} = \sup\{F(x) : x \in X\} = \sup\{G(y) : y \in Y\}$$

Or equivalently

$$\sup_{x,y} h(x, y) = \sup_{x} \sup_{y} h(x, y) = \sup_{y} \sup_{x} h(x, y)$$

Lemma E.2 (Maximum of Univariate Convex Functions on Compact Intervals)**.** *Let $T = [t_1, t_2]$ be a compact interval on the real line, where $t_1 < t_2$. Any convex function $f : T \to \mathbb{R}$ attains its maximum at t_1, t_2, or both.*

Proof. Since f is a convex function on the compact interval T, it is bounded. We aim to show that for any $p \in T$, $f(p) \leq \max(f(t_1), f(t_2))$.

Note that any point $p \in T$ can be expressed as a convex combination of the endpoints t_1 and t_2. That is, $p = \lambda t_1 + (1 - \lambda)t_2$ for some $\lambda \in [0, 1]$.

Hence, for any point p, we have:

$$\begin{aligned} f(p) &= f(\lambda t_1 + (1 - \lambda)t_2) \\ &\leq \lambda f(t_1) + (1 - \lambda) f(t_2) && \text{by convexity} \\ &\leq \max(f(t_1), f(t_2)) && \text{by } \lambda \in [0, 1] \end{aligned}$$

Since the function value at any interior point p is less than or equal to the maximum of the values at the endpoints, the global maximum of f on T is attained at either t_1, t_2, or both. □

Lemma E.3. *Let $a, b \in \mathbb{R}$ and $t \geq 0$. Then*

$$\sup_{c \in [-t,t]} |a + cb| = |a| + \sup_{c \in [-t,t]} |c||b| = |a| + t|b|$$

Proof. Note that $f(c) = |a + cb|$ is a convex function since for any $\lambda \in [0, 1]$,

$$
\begin{aligned}
f(\lambda x + (1-\lambda)y) &= |a + (\lambda x + (1-\lambda)y)b| \\
&= |\lambda(a+xb) + (1-\lambda)(a+yb)| \\
&\leq \lambda|a+xb| + (1-\lambda)|a+yb| \quad \text{by triangle inequality} \\
&= \lambda f(x) + (1-\lambda)f(y),
\end{aligned}
$$

The function $f(c)$ is defined and evaluated over the compact and symmetric interval $[-t, t]$. Over a compact set the supremum equals its maximum. Further, by lemma E.2 any univariate convex function attains its maximum at one of the end points, which for the symmetric sets considered here are just the negative of each other. The endpoint that maximizes the function value depends on the signs of a and b. Specifically, it is easy to verify that $c^* := \arg\max_{c\in[-t,t]} |a + cb|$ with

$$
c^* \begin{cases} t & \text{if} \quad ab \geq 0 \\ -t & \text{if} \quad ab < 0 \end{cases} \tag{52}
$$

Now with this construction of c^* one gets that $|a + c^*b| = |a| + |c^*b| = |a| + |c^*||b|$, since a and c^*b are by construction of c^* of the same sign. Further, by construction of c^* it is straighforward to see that $|c^*| = t$

Hence,

$$
\sup_{c\in[-t,t]} |a + cb| = |a| + |c^*||b| = |a| + t|b| = |a| + \sup_{c\in[-t,t]} |c||b| \tag{53}
$$

□

Lemma E.4. *Let $a, b, c \in \mathbb{R}$ and $r, s, t \geq 0$. Define the cartesian product of intervals $S := [-r, r] \times [-s, s] \times [-t, t]$. Then*

$$
\sup_{(a,b,c)\in S} |a + cb| = \sup_{a\in[-r,r]} |a| + \sup_{c\in[-t,t]} |c| \sup_{b\in[-s,s]} |b| \tag{54}
$$

Proof. Since clearly $|a + cb|$ is bounded on S applying lemma E.1 one can rewrite rewrite

$$
\sup_{(a,b,c)\in S} |a + cb| = \sup_{a\in[-r,r]} \left(\sup_{b\in[-s,s]} \left(\sup_{c\in[-t,t]} |a + cb| \right) \right) \tag{55}
$$

Now applying lemma E.3 to the inner-most sup over c one gets,

$$\sup_{(a,b,c)\in S} |a+cb| = \sup_{a\in[-r,r]} \left(\sup_{b\in[-s,s]} |a| + t|b| \right) \tag{56}$$

Clearly $\sup_{b\in[-s,s]} |a| + t|b| = |a| + ts$ and finally $\sup_{a\in[-r,r]} |a| + ts = r + ts$. It follows that

$$\sup_{(a,b,c)\in S} |a+cb| = r + ts = \sup_{a\in[-r,r]} |a| + \sup_{c\in[-t,t]} |c| \sup_{b\in[-s,s]} |b| \tag{57}$$

□

F Asymptotic Normality of the Stratified RD Estimator

Theorem 2 (Asymptotic Normality of the Stratified RD Estimator)**.** *Assume assumption 2, assumption 3, and assumption 6 to assumption 8. Further assume* $m(x,z), m(x,z)+\tau(x,z) \in \mathcal{F}_p(M_L)$, $g_1(x) \in \mathcal{F}_p(L)$ *and* $|\tau(x,1)-\tau(x,0)| \leq C_{\Delta\tau} \quad \forall x$, *then*

$$\sqrt{nh}\left(\hat{\tau}_S(h) - \tau - LB\right) \xrightarrow{d} N(0, \mathcal{V}_S)$$

where $LB := g_1 E[\hat{\tau}_1(h) - \tau_1 | \mathbf{X}, \mathbf{Z}] + (1-g_1) E[\hat{\tau}_0(h) - \tau_0 | \mathbf{X}, \mathbf{Z}] + [\tau_1 - \tau_0] E[\hat{g}_1(h_z) - g_1 | \mathbf{X}]$ *denotes the leading-order bias terms and the asymptotic variance is given by* $\mathcal{V}_S = \mathcal{V}_O + [\tau_1 - \tau_0]^2 \frac{1}{\kappa} \mathcal{V}_g$. *With* $\mathcal{V}_O$ *given in Lemma D.4 and* $\mathcal{V}_g$ *being the asymptotic variance of the weight estimation using a local linear estimator at an interior point using* Z_i *as outcome (see section 3.2 in Armstrong and Kolesár (2020)).*

Proof. See appendix F.

□

Proof Sketch: To apply the Lindeberg CLT I need to verify **Lindeberg CLT Conditions:**

i $E[Z_{n,i}] = 0$

ii $\sum_{i=1}^n Var(Z_{i,n}) = \sum_{i=1}^n E[Z_{i,n}^2] \xrightarrow[n\to\infty]{h\to 0} \sigma^2$ [7]

iii $L_n(\epsilon) := \sum_{i=1}^n E[Z_{n,i}^2 \mathbb{1}_{\{|Z_{n,i}|>\epsilon\}}] \to 0$

Steps:

1. Linearize the estimator to focus on leading terms
2. Decompose the estimator into bias and variance using the long regression model
3. Replace empirical gram matrix with its limiting matrix, in regression weights to control weight of individual error terms
    - This operation vanishes asymptotically, following arguments in Lemma.S.A.1 in Calonico et al. (2014a)

[7] Often assumed that $\sum_{i=1}^n Var(Z_{i,n}) = 1$ WLOG (see Billingsley (1995))

4. It then suffices to show the CLT result for the expressions using the deterministic local polynomial weights based on the limiting gram matrix

5. Verify Lindeberg Conditions

   - The relevant regularity assumption on the distribution of X_i is given by lemma D.2

F.I Stratified RD Linearization

Lemma F.1. *Let $\hat{\tau}_S$ be the statified RD estimator. Then one can write:*

$$\hat{\tau}_S = \tau + g[\hat{\tau}_1 - \tau_1] + (1-g)[\hat{\tau}_0 - \tau_0] + [\tau_1 - \tau_0][\hat{g} - g] + o_p((nh)^{-\frac{1}{2}}) \tag{58}$$

$$= \hat{\tau}_O + [\tau_1 - \tau_0][\hat{g} - g] + o_p((nh)^{-\frac{1}{2}}) \tag{59}$$

where $\hat{\tau}_O$ is the Stratified RD Oracle estimator.

Proof. Linearizing the stratified estimator gives:

$$\begin{aligned}\hat{\tau}_S &= \sum_{z\in\{0,1\}} [\tau_z + (\hat{\tau}_z - \tau_z)][g_z + (\hat{g}_z - g_z)] \\ &= \tau + \sum_{z\in\{0,1\}} g_z[\hat{\tau}_z - \tau_z] + [\hat{g}_z - g_z]\tau_z + [\hat{g}_z - g_z][\hat{\tau}_z - \tau_z]\end{aligned}$$

Now from standard results on local polynomial regressions (Armstrong and Kolesár 2020), and given that bandwidths are assumed to be proportional $h_n \asymp h_n^z$, we know that $[\hat{g}_z - g_z] = O(h^p) + O_p(\frac{1}{\sqrt{nh}})$ and $[\hat{\tau}_z - \tau_z] = O(h^p) + O_p(\frac{1}{\sqrt{nh}})$. Plugging this in and expanding the expression $[\hat{g}_z - g_z][\hat{\tau}_z - \tau_z] = [O(h^p) + O_p(\frac{1}{\sqrt{nh}})][O(h^p) + O_p(\frac{1}{\sqrt{nh}})] = O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh})$. Note since the uniform framework in Armstrong and Kolesár (2020) controls the bias uniformly over the function class, the linearization argument also holds uniformly over the assumed function classes.

$$\hat{\tau}_S = \tau + \sum_{z\in\{0,1\}} g_z[\hat{\tau}_z - \tau_z] + [\hat{g}_z - g_z]\tau_z + o(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh})$$

Further, noting that $\hat{g}_0 = (1 - \hat{g}_1)$ and $g_0 = (1 - g_1)$. Plugging this in we get:

$$\sum_{z\in(0,1)} [\hat{g}_z - g_z]\tau_z = [\hat{g}_1 - g_1][\tau_1 - \tau_0]$$

So putting everything together and writing out the sum over both groups gives:

$$\hat{\tau}_S = \tau + g_1[\hat{\tau}_1 - \tau_1] + g_0[\hat{\tau}_0 - \tau_0] + [\hat{g}_1 - g_1][\tau_1 - \tau_0] + O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh}) \quad (60)$$

Plugging in the definition of the CATE $\tau = g_1\tau_1 + g_0\tau_0$ and of the Oracle estimator $\hat{\tau}_O = g_1\hat{\tau}_1 + g_0\hat{\tau}_0$ gives the final display.

$$\hat{\tau}_S = \hat{\tau}_O + [\hat{g}_1 - g_1][\tau_1 - \tau_0] + O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh}) \quad (61)$$

Under the assumptions on bandwidths choices in assumption 3 $O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh}) = o_p\left(\frac{1}{\sqrt{nh}}\right)$, which proves the claim. For notational simplicity I define $g_1 := g$ and $g_0 := (1-g)$ in the main text for the two group case, which gives the final display.

$$\begin{aligned}\hat{\tau}_S &= \tau + g[\hat{\tau}_1 - \tau_1] + (1-g)[\hat{\tau}_0 - \tau_0] + [\tau_1 - \tau_0][\hat{g}_1 - g_1] + o_p((nh)^{-\frac{1}{2}}) && (62)\\ &= \hat{\tau}_O + [\tau_1 - \tau_0][\hat{g} - g] + o_p((nh)^{-\frac{1}{2}}) && (63)\end{aligned}$$

□

F.II Bias and Variance Decomposition

Starting from equation (60) one can derive a standard bias-variance decomposition. Plugging in definitions (c.f. appendix D) of groupwise RD and the local-linear weight estimator into equation (60) and inserting the respective regression models gives the bias-variance decomposition.

$$\begin{aligned}\hat{\tau}_j(h) &= \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_{n,j}(X_i/h, Z_i)Y_i \\ &= \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_{n,j}(X_i/h, Z_i)m(X_i, Z_i) + \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_{n,j}(X_i/h, Z_i)e_i \quad \text{(Long Regression Model)}\end{aligned}$$

Note that bias $E[\hat{\tau}_j(h)|\mathbf{X},\mathbf{Z}] - \tau_j = \frac{1}{nh}\sum_{i=1}^n \tilde{k}_{n,j}(X_i/h, Z_i)m(X_i, Z_i) - \tau_j$

$$[\hat{\tau}_j(h) - \tau_j] - E[\hat{\tau}_j - \tau_j|\mathbf{X},\mathbf{Z}] = \frac{1}{nh}\sum_{i=1}^n \tilde{k}_{n,j}(X_i/h, Z_i)e_i$$

Likewise

$$\begin{aligned}\hat{g}_1(h_z) &= \frac{1}{nh_z}\sum_{i=1}^n \tilde{k}_n^z(X_i/h_z)Z_i \\ &= \frac{1}{nh_z}\sum_{i=1}^n \tilde{k}_n^z(X_i/h_z)g_1(X_i) + \frac{1}{nh_z}\sum_{i=1}^n \tilde{k}_n^z(X_i/h_z)r_i\end{aligned}$$

$$[\hat{g}_1(h) - g_1] - E[\hat{g}_1 - g_1|\mathbf{X}] = \frac{1}{nh_z}\sum_{i=1}^n \tilde{k}_n^z(X_i/h_z)r_i$$

Now subtracting the leading bias terms from equation (60):

$$\begin{aligned}&\hat{\tau}_S - \tau - g_1 E[\hat{\tau}_1 - \tau_1|\mathbf{X},\mathbf{Z}] - (1-g_1)E[\hat{\tau}_0 - \tau_0|\mathbf{X},\mathbf{Z}] - [\tau_1 - \tau_0]E[\hat{g}_1 - g_1|\mathbf{X}] \\ =&g_1\frac{1}{nh}\sum_{i=1}^n \tilde{k}_{n,1}(X_i/h, Z_i)e_i + (1-g_1)\frac{1}{nh}\sum_{i=1}^n \tilde{k}_{n,0}(X_i/h, Z_i)e_i + [\tau_1 - \tau_0]\frac{1}{nh_z}\sum_{i=1}^n \tilde{k}_n^z(X_i/h_z)r_i \\ &O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh})\end{aligned}$$

F.III Replace Stochastic Local Polynomial Weights with their Deterministic Limit

Now we replace the the empirical gram matrices in the regression weights, that depend on the sample realization, with their limits. This helps to control the error terms to show asymptotic normality. In the next sections, we will show that this replacement is asymptotically negligible in the sense that $\sqrt{nh}[\tilde{k}_n - \tilde{k}] = o_P(1)$, where $\tilde{k}_n$ are the regression weights based on the empirical gram matrices, depending on the sample realization, and $\tilde{k}$ are the regression weights based on the limit of the gram matrices.

$$
\begin{aligned}
&\hat{\tau}_S - \tau - g_1 E[\hat{\tau}_1 - \tau_1|\mathbf{X},\mathbf{Z}] - (1-g_1)E[\hat{\tau}_0 - \tau_0|\mathbf{X},\mathbf{Z}] - [\tau_1 - \tau_0]E[\hat{g}_1 - g_1|\mathbf{X}] \\
=&g_1 \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_1(X_i/h, Z_i)e_i + (1-g_1)\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_0(X_i/h,Z_i)e_i + [\tau_1-\tau_0]\frac{1}{nh_z}\sum_{i=1}^{n}\tilde{k}^z(X_i/h_z)r_i \\
&+g_1\frac{1}{nh}\sum_{i=1}^{n}[\tilde{k}_{n,1}(X_i/h,Z_i) - \tilde{k}_1(X_i/h,Z_i)]e_i \\
&+(1-g_1)\frac{1}{nh}\sum_{i=1}^{n}[\tilde{k}_{n,0}(X_i/h,Z_i) - \tilde{k}_0(X_i/h,Z_i)]e_i \\
&+[\tau_1-\tau_0]\frac{1}{nh_z}\sum_{i=1}^{n}[\tilde{k}_n^z(X_i/h_z) - \tilde{k}^z(X_i/h_z)]r_i \\
&+O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh})
\end{aligned}
$$

Recall that $\tilde{k}_{n,j}(v.z) = \tilde{k}_{n,j,+}(v,z) - \tilde{k}_{n,j,-}(v,z)$, where $\tilde{k}_{n,j,+}(v,z) := e_1^\top(\tilde{Q}_{n,j,+})^{-1}m_1(v)k_+(v)\mathbb{1}_{\{z=j\}}$. The deterministic limiting local polynomial weights is given by, $\tilde{k}_j(v,z) = \tilde{k}_{j,+}(v,z) - \tilde{k}_{j,-}(v,z)$,
where $\tilde{k}_{j,+}(v,z) := e_1^\top(\tilde{Q}_{j,+})^{-1}m_1(v)k_+(v)\mathbb{1}_{\{z=j\}}$. The expressions for the regression on the control side are analogue. $\tilde{Q}_{n,j,+}$ and $\tilde{Q}_{j,+}$ are as defined in appendix B.

Now group the terms in line 2,3 and 4 together and collect them under $R := g_1\frac{1}{nh}\sum_{i=1}^{n}[\tilde{k}_{n,1}(X_i/h,Z_i) - \tilde{k}_1(X_i/h,Z_i)]e_i + (1-g_1)\frac{1}{nh}\sum_{i=1}^{n}[\tilde{k}_{n,0}(X_i/h,Z_i) - \tilde{k}_0(X_i/h,Z_i)]e_i + [\tau_1-\tau_0]\frac{1}{nh_z}\sum_{i=1}^{n}[\tilde{k}_n^z(X_i/h_z) - \tilde{k}^z(X_i/h_z)]r_i$. This gives the following expression:

$$
\begin{aligned}
&\hat{\tau}_S - \tau - g_1 E[\hat{\tau}_1 - \tau_1|\mathbf{X},\mathbf{Z}] - (1-g_1)E[\hat{\tau}_0 - \tau_0|\mathbf{X},\mathbf{Z}] - [\tau_1 - \tau_0]E[\hat{g}_1 - g_1|\mathbf{X}] \\
=&g_1 \frac{1}{nh}\sum_{i=1}^{n} \tilde{k}_1(X_i/h, Z_i)e_i + (1-g_1)\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_0(X_i/h,Z_i)e_i + [\tau_1-\tau_0]\frac{1}{nh_z}\sum_{i=1}^{n}\tilde{k}^z(X_i/h_z)r_i \\
&+R + O(h^{2p}) + O_p(\frac{h^p}{\sqrt{nh}}) + O_p(\frac{1}{nh})
\end{aligned}
$$

The goal is to show that even after pre-multiplying $\sqrt{nh}$, the replacement of the empirical gram matrices is asymptotically negligible, in the sense that $\sqrt{nh}R = o_P(1)$. The terms remaining from the linearization are also asymptotically negligible under the

assumptions on bandwidth choice as shown in lemma F.1. Note that under assumption 3, $\sqrt{nh}\frac{1}{nh_z} = \frac{1}{\kappa}\frac{1}{\sqrt{nh}}$. We want to show:

$$\begin{aligned}&\sqrt{nh}(\hat{\tau}_S - \tau - g_1 E[\hat{\tau}_1 - \tau_1|\mathbf{X},\mathbf{Z}] - (1-g_1)E[\hat{\tau}_0 - \tau_0|\mathbf{X},\mathbf{Z}] - [\tau_1 - \tau_0]E[\hat{g}_1 - g_1|\mathbf{X}])\\ =&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left[g_1\tilde{k}_1(X_i/h, Z_i)e_i + (1-g_1)\tilde{k}_0(X_i/h, Z_i)e_i + [\tau_1 - \tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)r_i\right] + o_P(1)\end{aligned}$$

Under assumption 3 it is straightforward to verify $\sqrt{nh}\left(O(h^4) + O_p(\frac{h^2}{\sqrt{nh}}) + O_P(\frac{1}{nh})\right) = o_P(1)$.

To verify that $\sqrt{nh}R = o_P(1)$, show that each element in the sum R is $o_P(1)$. We will show the result for one term $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}[\tilde{k}_{n,j}(X_i/h, Z_i) - \tilde{k}_j(X_i/h, Z_i)]e_i$. The convergence of the additional terms can be proofed in the same manner. Recall that the estimator includes a regression on the treatment and control side respectively,

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left[\underbrace{\left\{\tilde{k}_{n,j,+}(X_i/h, Z_i) - \tilde{k}_{j,+}(X_i/h, Z_i)\right\}}_{T} - \underbrace{\left\{\tilde{k}_{n,j,-}(X_i/h, Z_i) - \tilde{k}_{j,-}(X_i/h, Z_i)\right\}}_{C}\right]e_i.$$

We will proof the result for the term T and the result for C can be proven analogously.

F.IV Random Denominator Problem - Invertibility of Gram Matrix

Lemma F.2 (Limiting Gram Matrix Replacement)**.**

$$\begin{aligned}\tilde{Q}_{n,+}^{-1} - \tilde{Q}_{+}^{-1} &= o_p(1)\\ \tilde{Q}_{n,j,+}^{-1} - \tilde{Q}_{j,+}^{-1} &= o_p(1)\end{aligned}$$

Proof. The invertibility of the gram matrix for the Base-RD estimator is covered by Lemma S.A.1 in Calonico et al. (2014a). Since the groupwise RD estimator is simply the Base-RD estimator in a subsample the results follow immediately.

Basic matrix algebra shows that the expression in Calonico et al. (2014a), $\Gamma_+(h) := \tilde{Q}_{n,+} = \frac{1}{nh}\sum_{i=1}^{n} k_+(X_i/h)m_1(X_i/h)m_1(X_i/h)'$

By Lemma S.A.1 in Calonico et al. (2014a) and following their notation we know that

$$\Gamma_+(h_n) = \tilde{\Gamma}_+(h_n) + O_p(\frac{1}{\sqrt{nh}})$$
$$\tilde{\Gamma}_+(h_n) = \underbrace{f(0)\Gamma_+}_{:=\tilde{Q}_+} + o(1) \qquad \text{(by continuity of the density)}$$

where $\Gamma_+ = \int_0^\infty k(u)m_1(u)m_1(u)'du$
that yields under assumption 3 on the bandwidths,

$$\Gamma_+(h_n) = f(0)\Gamma_+ + o_p(1)$$

Lemma S.A.1 in Calonico et al. (2019) together with lemma F.3 gives:
So we have $\tilde{Q}_{n,+} - \tilde{Q}_+ = o_p(1)$ and $\tilde{Q}_{n,+}^{-1} - \tilde{Q}_+^{-1} = o_p(1)$. The proof is analogous for the group-wise regression.

□

Lemma F.3 (Continuity of Inverse)**.** *Suppose* Γ_+ *is invertible and* $|\Gamma_+(h_n) - f(0)\Gamma_+| = o_p(1)$*, then* $\Gamma_+(h_n)$ *is invertible with probability approaching one as* $n \to \infty$ *and*

$$\Gamma_+^{-1}(h_n) = \frac{1}{f(0)}\Gamma_+^{-1} + o_p(1) \tag{64}$$

Proof. Proof follows from standard results (e.g. Horn & Johnson, Matrix Analysis, 2nd ed. (2013) chapter 5.8) □

F.V Interchanging of Deterministic Limiting Local Polynomial Weights

Lemma F.4 (Replacement with Deterministic Limiting Local Polynomial Weights)**.**

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} T = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}[\tilde{k}_{n,j,+}(X_i/h, Z_i) - \tilde{k}_{j,+}(X_i/h, Z_i)]e_i = o_p(1) \tag{65}$$

Proof. Plugging the definitions of the regression weights into equation (65) yields:

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}[\tilde{k}_{n,j,+}(X_i/h,Z_i)-\tilde{k}_{j,+}(X_i/h,Z_i)]e_i \overset{=}{} \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} e_1^\top[\tilde{Q}_{n,j,+}^{-1}-\tilde{Q}_{j,+}^{-1}]m(X_i/h)k_+(X_i/h)\mathbb{1}_{\{Z_i=j\}}e_i$$

$$=e_1^\top \underbrace{[\tilde{Q}_{n,j,+}^{-1}-\tilde{Q}_{j,+}^{-1}]}_{o_p(1)} \underbrace{\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} m(X_i/h)k_+(X_i/h)\mathbb{1}_{\{Z_i=j\}}e_i}_{O_p(1)}$$

$$=o_p(1)$$

Noting that the matrices $\tilde{Q}_{n,j,+}$ and $\tilde{Q}_{j,+}$ do not depend on i we can pull them out of the sum. The elements of the first factor are going to zero in probability by the Lemma F.2 and the element of the second factor can be bounded in probability using standard CLT arguments and by the fact that the kernel function is assumed to be bounded with bounded support (assumption 2).

□

F.VI Rearranging

As we have shown above it suffices to determine the limiting distribution of the following expression to determine the asymptotic distribution of $\hat{\tau}_S$ minus the leading terms of its bias. Note that by assumption on the bandwidth choices $\sqrt{nh}\frac{1}{nh_z} = \sqrt{nh}\frac{1}{nh\kappa} = \frac{1}{\kappa}\frac{1}{\sqrt{nh}}$

$$\sum_{i=1}^{n}\left[\underbrace{\frac{1}{\sqrt{nh}}\left(g_1\tilde{k}_1(X_i/h,Z_i)e_i+(1-g_1)\tilde{k}_0(X_i/h,Z_i)e_i+[\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)r_i\right)}_{Z_{n,i}}\right]$$

I define $Z_{n,i}$ here to construct a triangular array, because as the sample size n changes the bandwidth h_n changes, leading to different distribution of the random variables $Z_{n,i}$'s (the ones outside the bandwidth h_n will be set to 0.).

F.VII Lindeberg Conditions

To apply the Lindeberg CLT I need to verify **Lindeberg CLT Conditions:**

i **Mean Zero**

$E[Z_{n,i}] = 0$

Proof.

$$\begin{aligned}
E[Z_{n,i}] =& E\left[\frac{1}{\sqrt{nh}}\left(g_1\tilde{k}_1(X_i/h, Z_i)e_i + (1-g_1)\tilde{k}_0(X_i/h, Z_i)e_i + [\tau_1 - \tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h)r_i\right)\right] \\
=& E\left[\frac{1}{\sqrt{nh}}\left(g_1\tilde{k}_1(X_i/h, Z_i)E[e_i|\mathbf{X}, \mathbf{Z}] + (1-g_1)\tilde{k}_0(X_i/h, Z_i)E[e_i|\mathbf{X}, \mathbf{Z}]\right.\right. \\
&\left.\left. + [\tau_1 - \tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)E[r_i|\mathbf{X}]\right)\right] \qquad \text{(LIE)} \\
=& 0
\end{aligned}$$

□

ii **Asymptotic Variance**

$\sum_{i=1}^n Var(Z_{i,n}) = \sum_{i=1}^n E[Z_{i,n}^2] \xrightarrow[n\to\infty]{h\to 0} \sigma^2$ [8]

Note,

$$Z_{n,i}^2 := \frac{1}{nh}\left(g_1\tilde{k}_1(X_i/h, Z_i)e_i + (1-g_1)\tilde{k}_0(X_i/h, Z_i)e_i + [\tau_1 - \tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)r_i\right)^2$$

Lemma F.5 (Asymptotic Variance)**.**

$$\begin{aligned}
\sum_{i=1}^n E[Z_{n,i}^2] &= g_1^2\mathcal{V}_1(k)^2 + (1-g_1)^2\mathcal{V}_0(k)^2 + [\tau_1 - \tau_0]^2\frac{1}{\kappa}\mathcal{V}_g(k)^2 + o(1) \\
&\qquad \text{(by standard results)} \\
&= \mathcal{V}_O(k)^2 + [\tau_1 - \tau_0]^2\frac{1}{\kappa}\mathcal{V}_g(k)^2 + o(1) \quad \text{(by same argument as in lemma D.4)}
\end{aligned}$$

[8] Often assumed that $\sum_{i=1}^n Var(Z_{i,n}) = 1$ WLOG (see Billingsley (1995))

Proof.

$$
\begin{aligned}
E[Z_{n,i}^2] &= E\left[\frac{1}{nh}\left(g_1\tilde{k}_1(X_i/h,Z_i)e_i + (1-g_1)\tilde{k}_0(X_i/h,Z_i)e_i + [\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)r_i\right)^2\right] \\
&= E\left[\frac{1}{nh}\left([g_1\tilde{k}_1(X_i/h,Z_i)e_i]^2 + [(1-g_1)\tilde{k}_0(X_i/h,Z_i)e_i]^2 + [[\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)r_i]^2\right)\right. \\
&+ \frac{1}{nh}2g_1(1-g_1)\tilde{k}_0(X_i/h,Z_i)\tilde{k}_1(X_i/h,Z_i)e_i^2 \\
&+ \frac{1}{nh}2g_1\tilde{k}_1(X_i/h,Z_i)[\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)e_ir_i \\
&\left.+ \frac{1}{nh}2[(1-g_1)\tilde{k}_0(X_i/h,Z_i)[\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h_z)e_ir_i\right] \\
&= g_1^2E\left[\frac{1}{nh}[\tilde{k}_1(X_i/h,Z_i)e_i]^2\right] + (1-g_1)^2E\left[\frac{1}{nh}[\tilde{k}_0(X_i/h,Z_i)e_i]^2\right] \\
&+ [\tau_1-\tau_0]^2\frac{1}{\kappa^2}E\left[\frac{1}{nh}[\tilde{k}^z(X_i/h_z)r_i]^2\right] \\
&= g_1^2E\left[\frac{1}{nh}[\tilde{k}_1(X_i/h,Z_i)e_i]^2\right] + (1-g_1)^2E\left[\frac{1}{nh}[\tilde{k}_0(X_i/h,Z_i)e_i]^2\right] \\
&+ [\tau_1-\tau_0]^2\frac{1}{\kappa}E\left[\frac{1}{nh_z}[\tilde{k}^z(X_i/h_z)r_i]^2\right]
\end{aligned}
$$

The last three terms in equation 2 are zero due to the following reasons: Note that by definition that $\tilde{k}_0(u,z)\tilde{k}_1(u,z) = e_1'\left(\tilde{Q}_{0,+}\right)^{-1}m_q(u)k_+(u)e_1'\left(\tilde{Q}_{1,+}\right)^{-1}m_q(u)k_+(u)\underbrace{\mathbb{1}_{\{z=0\}}\mathbb{1}_{\{z=1\}}}_{=0} = 0$ due to the mutually exclusive dummies in the regression weights. An application of the LIE shows that $E[s(X_i)e_ir_i] = E[s(X_i)r_i\underbrace{E[e_i|X_i,Z_i]}_{=0}]$, for any function $s(X_i)$. Further, in the last equation we plug in $h = \frac{1}{\kappa}h_z$.

$$
\begin{aligned}
E\left[\frac{1}{nh}\sum_{i=1}^n[\tilde{k}_1(X_i/h,Z_i)e_i]^2\right] &= E\left[\frac{1}{nh}\sum_{i=1}^n\tilde{k}_1(X_i/h,Z_i)^2\sigma_L^2(X_i,Z_i)\right] \quad \text{by LIE} \\
&= E\left[\frac{1}{nh}\sum_{i=1}^n\tilde{k}_{1,+}(X_i/h,Z_i)^2\sigma_L^2(X_i,Z_i)\right] \\
&+ E\left[\frac{1}{nh}\sum_{i=1}^n\tilde{k}_{1,-}(X_i/h,Z_i)^2\sigma_L^2(X_i,Z_i)\right]
\end{aligned}
$$

The second line follows by plugging in $\tilde{k}_1(X_i/h, Z_i) = \tilde{k}_{1,+}(X_i/h, Z_i) - \tilde{k}_{1,-}(X_i/h, Z_i)$ and using the fact the linear estimators on the left and right side of the cutoff are disjoint and hence $\tilde{k}_{1,+}(X_i/h, Z_i)\tilde{k}_{1,-}(X_i/h, Z_i) = 0$.

Lastly, note that in the last line we have the standard regression weights of the groupwise-RD $\tilde{k}_j$ and local-linear estimator for an interior point $\tilde{k}^z$, multiplied by some constants. Standard results by Armstrong and Kolesár (2020) and extensions presented in appendix D apply, which gives the result. These typically involve a change of variable and subsequently evaluate the limit of the expressions.

To demonstrate the mechanics, note that Theorem B.4 in the online appendix of Armstrong and Kolesár (2020) implies that:

$$\begin{aligned} E\left[\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_{1,+}(X_i/h, Z_i)^2\sigma_L^2(X_i, Z_i)\right] &\xrightarrow[n\to\infty]{h\to 0} f_{X,Z}(0,1)\sigma_{L,+}^2(0,1)\int_0^\infty \tilde{k}_{1,+}(u,1)^2 du \\ &= \frac{1}{f_{X,Z}(0,1)}\sigma_{L,+}^2(0,1)\int_0^\infty k_q^*(u)^2 du \\ &= \frac{1}{f_{X,Z}(0,1)}\sigma_{L,+}^2(0,1)\overline{k_q^2} \end{aligned}$$

Finally, since $\tilde{k}_{1,+}(u,z) = \frac{1}{f_{X,Z(0,1)}}k_q^*(u)\mathbb{1}\{u \geq 0\}\mathbb{1}_{\{z=1\}}$, the density terms cancel. The results for the term involving $\tilde{k}_{1,-}(X_i/h, Z_i)$ follow likewise.

□

iii **Tail Negligibility (Lindeberg Condition)**

Lemma F.6 (Lindeberg condition)**.** *In the given setting the Lindeberg condition,* $L_n(\epsilon) := \sum_{i=1}^n E[Z_{n,i}^2\mathbb{1}_{\{|Z_{n,i}|>\epsilon\}}] \to 0$*, is satisfied.*

Proof. Note that for any $\eta > 0$ $\mathbb{1}_{\{|Z_{n,i}|>\epsilon\}} = \mathbb{1}_{\{|Z_{n,i}|^\eta>\epsilon^\eta\}} \leq \frac{|Z_{n,i}|^\eta}{\epsilon^\eta}$. That gives

$$
\begin{aligned}
Z_{n,i}^2 \mathbb{1}_{\{|Z_{n,i}|>\epsilon\}} &\leq \frac{|Z_{n,i}|^{2+\eta}}{\epsilon^\eta} \\
E[Z_{n,i}^2 \mathbb{1}_{\{|Z_{n,i}|>\epsilon\}}] &\leq \frac{E[|Z_{n,i}|^{2+\eta}]}{\epsilon^\eta} \\
\sum_{i=1}^n E[Z_{n,i}^2 \mathbb{1}_{\{|Z_{n,i}|>\epsilon\}}] &\leq \frac{1}{\epsilon^\eta} \sum_{i=1}^n E[|Z_{n,i}|^{2+\eta}]
\end{aligned}
$$

Bound $\sum_{i=1}^n E[|Z_{n,i}|^{2+\eta}]$ using the Lp bound in lemma L.1, with $J = 3$ and $p = 2 + \eta$:

$$
\begin{aligned}
&\sum_{i=1}^n E[|Z_{n,i}|^{2+\eta}] \\
\leq &\frac{1}{nh} \frac{1}{(nh)^{\eta/2}} 3^{(1+\eta)} \\
&\cdot \sum_{i=1}^n \left(E[|g_1 \tilde{k}_1(X_i/h, Z_i) e_i|^{2+\eta}] + E[|(1-g_1)\tilde{k}_0(X_i/h, Z_i) e_i|^{2+\eta}] + E[|[\tau_1 - \tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h) r_i|^{2+\eta}] \right) \\
\leq &\frac{1}{\eta} \frac{1}{(nh)^{\eta/2}} 3^{(1+\eta)} \\
&\cdot \left(\frac{1}{nh} \sum_{i=1}^n E[|g_1 \tilde{k}_1(X_i/h, Z_i)|^{2+\eta}] + \frac{1}{nh} \sum_{i=1}^n E[|(1-g_1)\tilde{k}_0(X_i/h, Z_i)|^{2+\eta}] + \frac{1}{nh} \sum_{i=1}^n E[|[\tau_1 - \tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h
\end{aligned}
$$

The above line follows after applying the fact that $|a \cdot b|^p = |a|^p \cdot |b|^p$ and assumption 7 and assumption 8 together with the law of iterated expectation:

$$
\begin{aligned}
E[|g_1 \tilde{k}_1(X_i/h, Z_i) e_i|^{2+\eta}] &= E[|g_1 \tilde{k}_1(X_i/h, Z_i)|^{2+\eta} \underbrace{E[|e_i|^{2+\eta} | X_i, Z_i]}_{\leq \frac{1}{\eta}}] \\
&\leq \frac{1}{\eta} E[|g_1 \tilde{k}_1(X_i/h, Z_i)|^{2+\eta}]
\end{aligned}
$$

And likewise for the other terms.

Note that since the densities are assumed to be bounded and equivalent kernel functions are bounded by compactness of the kernel it follows from the calculations in the proof of lemma D.2 that the following expression converges to a constant:

$$\begin{aligned} A_n :&= E\left[\frac{1}{nh}\sum_{i=1}^{n}\left(|g_1\tilde{k}_1(X_i/h, Z_i)|^{2+\eta} + |(1-g_1)\tilde{k}_0(X_i/h, Z_i)|^{2+\eta} + |[\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(X_i/h)|^{2+\eta}\right)\right] \\ &= \int_{\mathcal{X}} f_{X,Z}(c,1)|g_1\tilde{k}_1(u,1)|^{2+\eta} + f_{X,Z}(c,0)|(1-g_1)\tilde{k}_0(u,0)|^{2+\eta} + f_X(c)|[\tau_1-\tau_0]\frac{1}{\kappa}\tilde{k}^z(u)|^{2+\eta}du + o(1) \\ &:= A + o(1) \end{aligned}$$

where $g_1 := g_1(c) := \frac{f_{X,Z}(x,1)}{f_x(c)}$ is a shorthand notation.

Putting everything together one can conclude that,

$$\begin{aligned} L_n(\epsilon) &\leq \frac{1}{\epsilon^\eta}\sum_{i=1}^{n} E[|Z_{n,i}|^{2+\eta}] \\ &\leq \frac{1}{\epsilon^\eta}\frac{1}{\eta}\frac{1}{(nh)^{\eta/2}}A_n \\ &= \frac{1}{\epsilon^\eta}\frac{1}{\eta}\frac{1}{(nh)^{\eta/2}}(A + o(1)) \end{aligned}$$

It follows that the Lindeberg conditions holds, $L_n(\epsilon) \to 0$ for any $\epsilon > 0$, since $\frac{1}{(nh)^{\eta/2}} \to 0$ as long as $nh \to \infty$ for any $\eta > 0$, which allows the application of the CLT. □

Lemma F.7 (Lp-bound)**.** *Let $a, b \in \mathbb{R}$ and $p \geq 1$. Then*

$$|a+b|^p \leq 2^{p-1}\Big(|a|^p + |b|^p\Big).$$

And more generally let $\{x_1, ..., x_j, ..., x_J\} \in \mathbb{R}$ and $p \geq 1$, then

$$\left|\sum_{j=1}^{J} x_j\right|^p \leq J^{p-1}\sum_{j=1}^{J}|x_j|^p$$

Proof. By the triangle inequality,

$$|a+b| \le |a| + |b|$$
$$|a+b|^p \le \big(|a| + |b|\big)^p$$

Express the right-hand side as

$$(|a| + |b|)^p = 2^p \left(\frac{|a| + |b|}{2} \right)^p .$$

Since the function $x \mapsto x^p$ is convex on $[0, \infty)$ for $p \ge 1$, Jensen's inequality implies

$$\left(\frac{|a| + |b|}{2} \right)^p \le \frac{|a|^p + |b|^p}{2}.$$

Combining these, we obtain

$$|a+b|^p \le 2^p \cdot \frac{|a|^p + |b|^p}{2} = 2^{p-1} \Big(|a|^p + |b|^p \Big).$$

The general version follows by induction using the two-term inequality. □

G Asymptotic Worst Case Bias Comparisons

To derive worst-case bias expressions over the common parameter space I have to introduce additional notation for the constants that control smoothness of the regression functions and bound heterogeneity. To keep the notation compact I collect the constants in tuples. $\mathbf{M}$ collects smoothness constants on regression functions and $\mathbf{C}$ collects constants to bound heterogeneity. Further, depending on the order assumed on the p-th order Hölder smoothness class I will need more or less constants. In general $\mathbf{M}$ and $\mathbf{C}$ will contain $1+p$ constants. The, number of constants in an index is implicitly indexed through the subscript for the function class $\mathcal{F}_p^R$. Since the common parameter space is more complex additional notation is necessary. Let M^j denote the constant for the j-th order Hölder function class of the long regression function and M_z^j the constant for the j-th order Hölder function class of the group regression function. Further, $C_{\Delta m}^j$ denotes the constant for the j-th order Hölder function class of the difference of regression functions of the groups $[m(x,1)-m(x,0)]$.

For the case $p=1$:

$$\begin{aligned}\mathbf{M} &= \left\{M^1, M_z^1\right\} \\ \mathbf{C} &= \left\{C_{\Delta m}, C_{\Delta \tau}\right\}\end{aligned}$$

For the case $p=2$

$$\begin{aligned}\mathbf{M} &= \left\{M^2, M_z^2, M_z^1\right\} \\ \mathbf{C} &= \left\{C_{\Delta m}^1, C_{\Delta m}, C_{\Delta \tau}\right\}\end{aligned}$$

For the case $p=3$

$$\begin{aligned}\mathbf{M} &= \left\{M^3, M_z^3, M_z^2, M_z^1\right\} \\ \mathbf{C} &= \left\{C_{\Delta m}^2, C_{\Delta m}^1, C_{\Delta m}, C_{\Delta \tau}\right\}\end{aligned}$$

For the lower bound argument it will be useful to define a subclass that restricts heterogeneity in a way that the regression functions of the two groups are just vertically shifted:

For the case $p=1$, the class remains as is.

$$\mathbf{C}^* = \mathbf{C} \tag{66}$$

For the case $p = 2$

$$\mathbf{C}^* = \{C^1_{\Delta m} = 0, C_{\Delta m}, C_{\Delta \tau}\} \tag{67}$$

For the case $p = 3$

$$\mathbf{C}^* = \left\{C^2_{\Delta m} = 0, C^1_{\Delta m} = 0, C_{\Delta m}, C_{\Delta \tau}\right\} \tag{68}$$

G.I Base RD - Worst-Case Bias Lower Bound over the Common Parameter Space

Lemma G.1 (Lower Bound Argument for WCB of Base RD over Common Parameter Space)**.**

$$\sup_{m(x)\in\mathcal{F}^R_p(\mathbf{M},\mathbf{C}^*)} E[\hat{\tau} - \tau|\mathbf{X}] \leq \sup_{m(x)\in\mathcal{F}^R_p(\mathbf{M},\mathbf{C})} E[\hat{\tau} - \tau|\mathbf{X}]$$

Proof. By construction of $\mathbf{C}^*$ it follows that $\mathcal{F}^R_p(\mathbf{M}, \mathbf{C}^*) \subseteq \mathcal{F}^R_p(\mathbf{M}, \mathbf{C})$. Clearly, the supremum of a function over $\mathcal{F}^R_p(\mathbf{M}, \mathbf{C}^*)$ cannot exceed the supremum of the function over $\mathcal{F}^R_p(\mathbf{M}, \mathbf{C})$. □

Proposition 8 (Base RD - Worst-Case Bias over Subclass)**.**

$$\begin{aligned}\overline{\text{Bias}}^{LowerBound}(\hat{\tau}) :&= \sup_{m(x)\in\mathcal{F}^R_p(\mathbf{M},\mathbf{C}^*)} E[\hat{\tau} - \tau|\mathbf{X}] \\ &= h^p[2M^p_L + (2C_{\Delta m} + C_{\Delta\tau})M^p_z]\frac{1}{p!}B_{p,q}(k) + o(h^p)\end{aligned}$$

Proof. [**Available Upon Request**]

□

G.II Worst-Case Bias Comparison

Proposition 4 (Leading Term Worst-Case Bias Comparison)**.** *Assume assumption 2 to assumption 8. Additionally assume the smoothness constraints* $m(x), m(x) + \tau(x), m(x,z), m(x,z) + \tau(x,z), g_1(x), [\tau(x,1) - \tau(x,0)] \in \mathcal{F}^R_p(\mathbf{M}, \mathbf{C}, \mathbf{L})$ *Following assumption 3 one chooses* $h_z = \kappa h$.

Denote the leading bias terms of the asymptotic expansion as, $\overline{\text{Bias}}_{\text{lead}}(\hat{\tau}) := \lim_{n\to\infty} \frac{\overline{\text{Bias}}(\hat{\tau})}{h^p}$. *As long as* $\kappa > 0$, *is chosen smaller than a kernel specific constant*

$\kappa \leq U_{p,q}(k)$, that depends on p and polynomial order q, then the leading terms of the bias of the estimators can be ranked

$$\overline{\text{Bias}}_{\text{lead}}(\hat{\tau}_S) \leq \overline{\text{Bias}}_{\text{lead}}(\hat{\tau})$$

or equivalently,

$$\overline{\text{Bias}}(\hat{\tau}_S) \leq \overline{\text{Bias}}(\hat{\tau}) + o(h^p)$$

Proof. First, take the leading terms of the worst-case bias for the stratified RD and the lower bound on the worst-case bias of base RD, $\bar{B}_a^{LB}(\hat{\tau})$, from Proposition 3 and Proposition 8. To guarantee $\bar{B}_a(\hat{\tau}_S) \leq \bar{B}_a(\hat{\tau})$, it suffices to choose $\kappa \leq \left(\frac{B_{p,q}(k)}{B_{p,q}^{int}(k)}\right)^{\frac{1}{p}} := U_{p,q}(k)$. With this choice the following inequality holds:

$$\begin{aligned}
\overline{\text{Bias}}_{\text{lead}}(\hat{\tau}_S) &\leq \overline{\text{Bias}}_{\text{lead}}^{\text{LowerBound}}(\hat{\tau}) \leq \overline{\text{Bias}}_{\text{lead}}(\hat{\tau}) \\
\frac{2M_L}{p!}B_{p,q}(k) + \kappa^p C_{\Delta\tau}\frac{M_z}{p!}B_{p,q}^{int}(k) &\leq [2M_L + (2C_{\Delta m} + C_{\Delta\tau})M_z]\frac{1}{p!}B_{p,q}(k) \\
\kappa^p C_{\Delta\tau}\frac{M_z}{p!}B_{p,q}^{int}(k) &\leq [(2C_{\Delta m} + C_{\Delta\tau})M_z]\frac{1}{p!}B_{p,q}(k) \\
\kappa^p C_{\Delta\tau}M_z B_{p,q}^{int}(k) &\leq C_{\Delta\tau}M_z B_{p,q}(k) + 2C_{\Delta m}M_z B_{p,q}(k) \\
C_{\Delta\tau}M_z[\kappa^p B_{p,q}^{int}(k) - B_{p,q}(k)] &\leq 2C_{\Delta m}M_z B_{p,q}(k)
\end{aligned}$$

Where in the last line we plug-in assumption 3. Since all constants are greater or equal to zero, the right hand side, $2C_{\Delta m}M_z B_{p,q}(k)$, can be at lowest be zero. So to guarantee the inequality we need to pick κ such that that $[\kappa^p B_{p,q}^{int}(k) - B_{p,q}(k)] \leq 0$. This is guaranteed for any choice $\kappa \leq \left(\frac{B_{p,q}(k)}{B_{p,q}^{int}(k)}\right)^{\frac{1}{p}} := U_{p,q}(k)$

□

Valid choices of κ will depend on the kernel constants $B_{p,q}, B_{p,q}^{int}$, which are vary depending on the chosen kernel, Hölder class and polynomial order for the local polynomial regression. Here I provide constants for the case $p = 2, q = 1$ and for the uniform and triangular kernel based on table S.1 and S.2 in the online appendix of Armstrong and Kolesár (2020).

1. Uniform Kernel

$B_{2,1}(k) = \frac{1}{6}$ and $B_{2,1}^{int}(k) = \frac{1}{3}$

$$\kappa^2 \frac{1}{3} - \frac{1}{6} = 0$$
$$\kappa = \frac{1}{\sqrt{2}}$$

So for $\kappa \leq \frac{1}{\sqrt{2}}$ we get $\kappa^2 B_{2,1}^{int}(k) - B_{2,1}(k) \leq 0$.

2. Triangular Kernel

$B_{2,1}(k) = \frac{1}{10}$ and $B_{2,1}^{int}(k) = \frac{1}{6}$

$$\kappa^2 \frac{1}{6} - \frac{1}{10} = 0$$
$$\kappa = \sqrt{\frac{3}{5}}$$

So for $\kappa \leq \sqrt{\frac{3}{5}}$ we get $\kappa^2 B_{2,1}^{int}(k) - B_{2,1}(k) \leq 0$.

H Asymptotic Variance Comparison

Proposition 5 (Asymptotic Variance Comparison)**.** *Let $\mathcal{V}_B$ denote the asymptotic variance of the base RD estimator and $\mathcal{V}_S$ denote the asymptotic variance of the Stratified RD estimator respectively. For an appropriate scaling of the secondary bandwidth $h_z = \kappa h$ (assumption 3) with $\kappa \geq L_{p,q}(k)$ the following holds:*

$$\mathcal{V}_S \leq \mathcal{V}_B \tag{20}$$

with strict inequality if there is group heterogeneity.

Proof. The variance comparison is done for the local linear case (e.g. $q = 1$). However, the result can be easily generalized to a general q-th order local polynomial RD estimator.

From theorem 1 we know that the base-RD estimator has asymptotic variance:

$$\mathcal{V}_B = \frac{\overline{k_1^2}\left(\sigma^2_{P,+}(0) + \sigma^2_{P,-}(0)\right)}{f_X(c)}, \tag{69}$$

Now by plugging in lemma A.1 and noting that $\sigma^2_{P,-}(0)$ does not include a term on the difference in treatment effects, we get:

$$\begin{aligned}
\mathcal{V}_B &= \frac{\overline{k_1^2}\left((\Delta m(c)^2 + [\Delta m(c) + \Delta\tau(c)]^2)\sigma^2_Z(c) + \sigma^2_{L,+}(c) + \sigma^2_{L,-}(c)\right)}{f_X(c)} && (70)\\
&= \frac{\overline{k_1^2}\left([2\Delta m(c)^2 + 2\Delta m(c)\Delta\tau(c) + \Delta\tau(c)^2]\sigma^2_Z(c)\right)}{f_X(c)} + \frac{\overline{k_1^2}\left(\sigma^2_{L,+}(c) + \sigma^2_{L,-}(c)\right)}{f_X(c)} && (71)\\
&= \mathcal{V}_O + \frac{\overline{k_1^2}\left([2\Big(\Delta m(c) + \tfrac{1}{2}\Delta\tau(c)\Big)^2 + \tfrac{1}{2}\Delta\tau(c)^2]\sigma^2_Z(c)\right)}{f_X(c)} && (72)\\
&= \mathcal{V}_O + \frac{2\Big(\Delta m(c) + \tfrac{1}{2}\Delta\tau(c)\Big)^2\sigma^2_Z(c)\overline{k_1^2} + \tfrac{1}{2}\Delta\tau(c)^2\sigma^2_Z(c)\overline{k_1^2}}{f_X(c)} && (73)
\end{aligned}$$

where the second to last line follows by completing the square:

$$\Delta m(c)^2 + [\Delta m(c) + \Delta\tau(c)]^2 = 2\Delta m(c)^2 + 2\Delta m(c)\Delta\tau(c) + \Delta\tau(c)^2 = 2\Big(\Delta m(c) + \tfrac{1}{2}\Delta\tau(c)\Big)^2 + \tfrac{1}{2}\Delta\tau(c)^2$$

On the other hand from theorem 2 we get that:

$$
\begin{aligned}
\mathcal{V}_S &= \mathcal{V}_O + \frac{1}{\kappa}\Delta\tau(c)^2\mathcal{V}_W \\
&= \frac{\overline{k_1^2}\left(\sigma^2_{L,+}(0) + \sigma^2_{L,-}(0)\right)}{f_X(c)} + \frac{1}{\kappa}\frac{\widetilde{k_1^2}\Delta\tau(c)^2\sigma_Z^2(c)}{f_X(c)} \qquad \text{(by ??)} \\
&= \mathcal{V}_O + \frac{1}{\kappa}\frac{\widetilde{k_1^2}\Delta\tau(c)^2\sigma_Z^2(c)}{f_X(c)}
\end{aligned}
$$

where $\widetilde{k_1^2} := \int_{-1}^{1} k_1^*(u)^2 du$ is the kernel constant for an interior point estimation problem. For specific kernels (e.g. triangular) these constants are given in appendix D.3 (Armstrong and Kolesár 2020) and hence can be easily compared.

$$
\begin{aligned}
\mathcal{V}_B - \mathcal{V}_S &= \underbrace{\frac{2\Big(\Delta m(c) + \frac{1}{2}\Delta\tau(c)\Big)^2\sigma_Z^2(c)\overline{k_1^2}}{f_X(c)}}_{\geq 0} + \frac{\frac{1}{2}\Delta\tau(c)^2\sigma_Z^2(c)\overline{k_1^2}}{f_X(c)} - \frac{1}{\kappa}\frac{\widetilde{k_1^2}\Delta\tau(c)^2\sigma_Z^2(c)}{f_X(c)} \\
&\geq \Delta\tau^2\sigma_Z^2\frac{1}{f_X(c)}\left(\frac{1}{2}\overline{k_1^2} - \frac{1}{\kappa}\widetilde{k_1^2}\right) \qquad (74)
\end{aligned}
$$

We want to pick κ such that $\left(\frac{1}{2}\overline{k_1^2} - \frac{1}{\kappa}\widetilde{k_1^2}\right) \geq 0$ to guarantee that $\mathcal{V}_B - \mathcal{V}_S \geq 0$.

□

H.I Local Linear estimation with (q=1, p=2) using standard kernels

Now the tables S.1 and S.2 in Appendix D.3 of Armstrong and Kolesár (2020) give the kernel constants that let me determine the interval for κ for which the variance $V_B - V_S \geq 0$, by choosing κ such that $\frac{1}{2}\overline{k_1^2} - \frac{1}{\kappa}\widetilde{k_1^2} \geq 0$.

1. Uniform Kernel:

For example using the uniform kernel in a local linear regression implies kernel constant $\overline{k_1^2} = 4$ and $\widetilde{k_1^2} = \frac{1}{2}$ yields that the bandwidth $h_z = \kappa h$ has to be choosen with $\kappa \geq \frac{1}{4}$.

2. Triangular Kernel:

Using the triangular kernel in a local linear regression implies kernel constant $\overline{k_1^2} = \frac{24}{5}$ and $\widetilde{k_1^2} = \frac{2}{3}$ yields that the bandwidth $h_z = \kappa h$ has to be choosen with $\kappa \geq \frac{5}{18}$.

H.II MSE Comparison

Taken together the results in Proposition 4 and Proposition 5 imply the following theorem, where the respective bounds in the previous results form the interval $\kappa \in I_{p,q}(k) := [L_{p,q}(k), U_{p,q}(k)]$.

Theorem 6 (Asymptotic Worst-Case MSE Comparison)**.** *Assume assumption 2 to assumption 8. Additionally assume the smoothness constraints* $m(x), m(x)+\tau(x), m(x,z), m(x,z)+\tau(x,z), g_1(x), [\tau(x,1)-\tau(x,0)] \in \mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$*. The bandwidths are chosen according to assumption 3,* $h_z = \kappa h$*. Define the asymptotic worst-case MSE of the two estimators as*

$$\overline{AMSE}(\hat{\tau}(h)) := \left(h^p \overline{\text{Bias}}_{\text{lead}}(\hat{\tau}(h))\right)^2 + \frac{V_B}{nh},$$
$$\overline{AMSE}(\hat{\tau}_S(h, h_z)) := \left(h^p \overline{\text{Bias}}_{\text{lead}}(\hat{\tau}_S(h, h_z))\right)^2 + \frac{V_S}{nh}.$$

Then if κ *is chosen from a kernel specific interval* $\kappa \in I_{p,q}(k)$*, that depends on the used kernel, the order of smoothness* p *and polynomial order* q*, the asymptotic worst case MSE of the Stratified RD estimator is uniformly smaller than that of the base RD estimator over* $\mathcal{F}_p^R(\mathbf{M}, \mathbf{C}, \mathbf{L})$*.*

$$\overline{AMSE}(\hat{\tau}_S(h, h_z)) \leq \overline{AMSE}(\hat{\tau}(h))$$

With strict inequality whenever groups are heterogeneous.

I Rate Optimality

Corollary 7 (Rate Optimality)**.** *As long as $m(x,z) \in \mathcal{F}_p$ and $g(x) \in \mathcal{F}_p$ lie in a Hölder class of same order p and given bandwidth choices consistent with assumption 3, the estimators $\hat{\tau}(h)$ and $\hat{\tau}_s(h, h_z)$ achieve the mini-max optimal rate $n^{-\frac{2p}{2p+1}}$ over the smoothness class $\mathcal{F}^R(\vec{M})$.*

Proof. The proof follows directly by the proofs on the lower bound rate (theorem 9) and upper bound rate (theorem 2) □

Theorem 9 (Minimax Lower Bound rate for $\mathcal{F}_2^R$)**.** *The lower bound rate for $\mathcal{F}_2^R$ is $n^{\frac{2p}{2p+1}}$. Or put differently, there exist a constant c, such that*

$$c \leq \liminf_{n\to\infty} \underbrace{\left(n^{\frac{2p}{2p+1}}\right)}_{\psi_n^{-2}} \underbrace{\left[\inf_{\hat{\theta}} \sup_{m\in\mathcal{F}_2^R} \mathbb{E}\big[(\hat{\theta} - m(c))^2\big]\right]}_{\mathcal{R}_n^*(\mathcal{F}_2^R)} \tag{75}$$

Proof. Construct $\mathcal{F}_2^0$ such that $\mathcal{F}_2^0 \subseteq \mathcal{F}_2^R$ and so I can get a lower bound rate on $\mathcal{F}_2^0$ by existing results. By defining $\boldsymbol{K}_0 = \{K_m'', K_\tau'', K_{\Delta m} = 0, K_{\Delta m}' = 0, K_{\Delta\tau} = 0, K_{\Delta\tau}' = 0, K_Z' = 0, K_Z'' = 0\}$, we impose group homogeneity and create a subset $\mathcal{F}_2^0 \subseteq \mathcal{F}_2^R$. Within this subset the second derivative of $m(x)$ is:

$$m''(x) = \sum_{z\in\{0,1\}} \underbrace{m''(x,z)g_z(x)}_{\text{I}} + \underbrace{2m'(x,z)g_z'(x)}_{\text{II}} + \underbrace{m(x,z)g_z''(x)}_{\text{III}} \overset{\mathcal{F}_2^0}{=} m''(x,\cdot)$$

due to our assumption of the constants to be zero, which forces the regression functions in both groups to be identical (and the fact that conditional densities sum to 1).

Hence the function class is just a collection of functions $m(x)$ with a second derivative upper bounded by K_m'' uniformly over x. Standard lower bound derivations, on nonparametric regression with the p-th derivative bounded (c.f. Tsybakov (2009) p.95) give:

$$c \leq \liminf_{n\to\infty} \inf_{\hat{\theta}} \sup_{m\in\mathcal{F}_2^0} \mathbb{E}_m\Big[n^{\frac{2p}{2p+1}} \big(\hat{\theta} - m(c)\big)^2\Big], \tag{76}$$

$\mathcal{F}_2^0 \subseteq \mathcal{F}_2^R$ and properties of supremum imply for every estimator $\hat{\theta}$:

$$\sup_{m\in\mathcal{F}_2^0}\mathbb{E}\big[(\hat{\theta}-m(c))^2\big] \;\leq\; \sup_{m\in\mathcal{F}_2^R}\mathbb{E}\big[(\hat{\theta}-m(c))^2\big]$$

Since this holds for all $\hat{\theta}$ we get:

$$\begin{aligned}
\inf_{\hat{\theta}}\sup_{m\in\mathcal{F}_2^0}\mathbb{E}\big[n^{\frac{2p}{2p+1}}(\hat{\theta}-m(c))^2\big] &\leq \inf_{\hat{\theta}}\sup_{m\in\mathcal{F}_2^R}\mathbb{E}\big[n^{\frac{2p}{2p+1}}(\hat{\theta}-m(c))^2\big] \\
\liminf_{n\to\infty}\inf_{\hat{\theta}}\sup_{m\in\mathcal{F}_2^0}\mathbb{E}\big[n^{\frac{2p}{2p+1}}(\hat{\theta}-m(c))^2\big] &\leq \liminf_{n\to\infty}\inf_{\hat{\theta}}\sup_{m\in\mathcal{F}_2^R}\mathbb{E}\big[n^{\frac{2p}{2p+1}}(\hat{\theta}-m(c))^2\big]
\end{aligned}$$

Combined with equation (76) this gives the lower bound rate for $\mathcal{F}_2^R$.

$$c \leq \liminf_{n\to\infty}\underbrace{\left(n^{\frac{2p}{2p+1}}\right)}_{\psi_n^{-2}}\underbrace{\left[\inf_{\hat{\theta}}\sup_{m\in\mathcal{F}_2^0}\mathbb{E}\big[(\hat{\theta}-m(c))^2\big]\right]}_{\mathcal{R}_n^*(\mathcal{F}_2^0)} \;\leq\; \liminf_{n\to\infty}\underbrace{\left(n^{\frac{2p}{2p+1}}\right)}_{\psi_n^{-2}}\underbrace{\left[\inf_{\hat{\theta}}\sup_{m\in\mathcal{F}_2^R}\mathbb{E}\big[(\hat{\theta}-m(c))^2\big]\right]}_{\mathcal{R}_n^*(\mathcal{F}_2^R)} \tag{77}$$

□

J On Pointwise Asymptotics for Local Polynomial Estimators

[Work in Progress]

Higher Order Smoothness. The stylized example extends naturally to the case of higher-order smoothness, where the regression function of the binary group variable is assumed to satisfy, $g(x) \in \mathcal{F}_p(M_z)$. In this setting, the optimal polynomial order for the baseline RD estimator is $p-1$. The derivations proceed as above, and combining them with the results in Armstrong and Kolesár (2020) yields the worst-case bias.

$$\overline{\text{Bias}}(\hat{\tau}(h)) = h^p M_z [2C_{\Delta m} + C_{\Delta \tau}] \frac{1}{p!} B_{p,q}(k)(1 + o(1)) \tag{78}$$

Clearly, the worst-case bias converges at the standard nonparametric rate h^p. For example, choosing $p = 2$ leads to the canonical local linear base RD estimator, for which the worst-case bias converges at rate h^2.

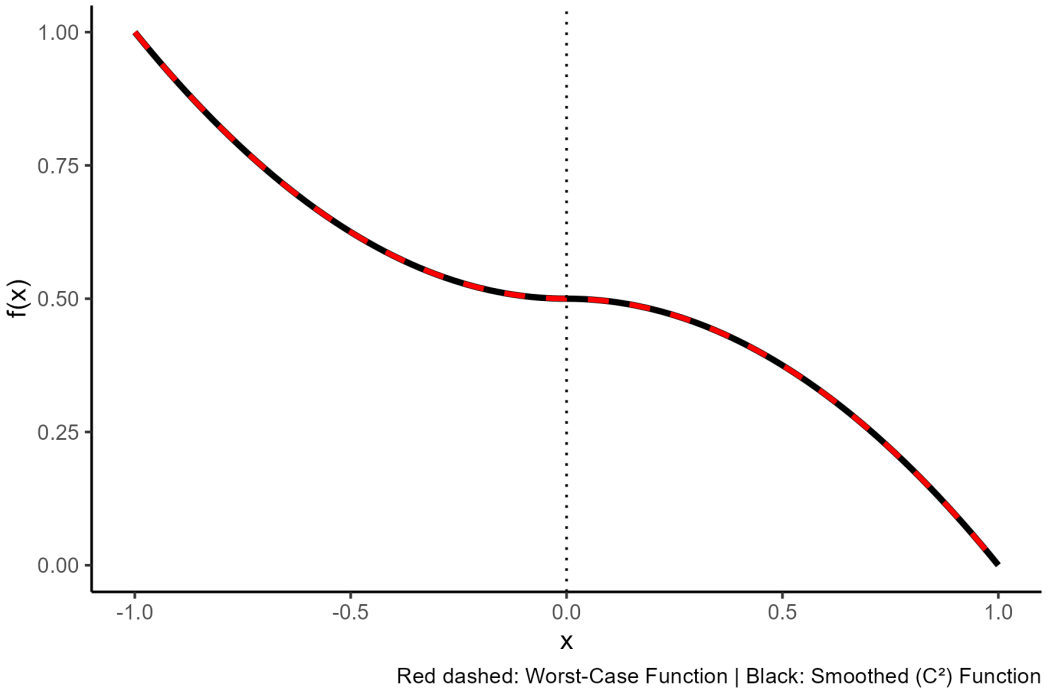


Figure 1: Approximation of the Worst-Case Function using a Polynomial Patch ($\epsilon = 0.01$)

J.I Simulation Example

To demonstrate the effect of pooling in the stylized example considered in the main text numerically, I carry out a simple simulation study using the following design. The regression functions are constant within groups shifted by a constant, $Y = 5+4Z_i+e_i$ with $e_i \sim N(0,1)$. The running variable is drawn from a uniform distribution $X_i \sim U[-1,1]$.

The distribution of the binary group variable is crucial for generating pooling bias. I model the group variable to be distributed according to a Bernoulli distribution with varying success probability, $Z_i \sim B(p(X_i))$. The success probability is modeled as a function of the running variable. Note that the success probability here is just the value of the regression function of the binary covariate, $p(X_i) = g(X_i)$. The varying success probability creates covariance imbalance across the cut-off. To illustrate pooling bias for the local constant and local linear base RD estimator, I model the group probability using two different functions $g(x)$ that lie in a first and second order Hölder class respectively, specified below.

Lastly I draw from the distributions n_k times with $n_k \in \{1000, 2000, ..., 20.000\}$ to create random samples $\{Y_i, X_i, Z_i\}_{i=1}^{n_k}$. Then I apply existing software implementations of the base RD estimator (rdrobust, RDHonest) and my implementation of the Stratified RD estimator on each sample. Bandwidths are chosen by rdrobust and RDHonest to be MSE-optimal, shrinking with the sample size at rate $h \asymp n^{-1/(2p+1)}$. I repeat the simulation $S = 1000$ times and average the estimates to empirically calculate the bias relative to a treatment effect of zero ($\tau = 0$) in this example.

Local Constant Estimation. The worst case function for the RD parameter under a Lipschitz constraint ($p = 1$) on the regression functions is given by $f(x) = Mx + c$. M denotes the derivative upper bound and the function can be arbitrarily shifted up an down by a constant c which does not affect the properties of the estimator. Since $g(x)$ is a conditional probability, I choose c such that the function still lies in $[0,1]$ over its domain. The parameterized function is $g(x) = 0.4x + 0.5$.

Local Linear Estimation. As in Armstrong and Kolesár (2018) the worst case function is $f(x) = -\frac{M}{2}x^2 \operatorname{sign}(x)$. Again, one can arbitrarily shift the function by a constant without affecting the properties of the estimator. Since, $g(x)$ is constrained to lie in $[0,1]$ I choose the constants M and c to satisfy this. Note that the worst-case function is not C^2, since the p-th order Hölder class is the closure of all p times continuously differentiable functions with upper bound M on the p-th derivative.

However, pointwise asymptotic theory assumes the regression functions to be sufficiently often continuously differentiable. To demonstrate the contradiction to the predictions of

pointwise asymptotic theory, I construct a function that satisfies all the assumptions in the pointwise asymptotic framework, but still has a bias that essentially behaves as under the worst-case function. To achieve this, I construct a polynomial patch $g_\epsilon(x)$ in an ϵ neighborhood around 0 that smoothes out the discontinuity in the second derivative of the worst-case function $\tilde{g}(x) = -\frac{1}{2}x^2 \operatorname{sign}(x) + 0.5$. The patch is constructed to match the function value, first, and second derivative at the intersection points $-\epsilon, \epsilon$. The second derivative of $g(x)$ remains globally bounded by 1 over the domain of the function. The specific patch used for the simulation is $g_\epsilon(x) = \frac{1}{16}\frac{1}{\epsilon^3}x^5 - \frac{3}{8}\frac{1}{\epsilon}x^3 - \frac{3}{16}\epsilon x + 0.5$.

The group regression function is then modeled as $g(x) = \tilde{g}(x)\mathbb{1}\{|x| \geq \epsilon\} + g_\epsilon(x)\mathbb{1}\{|x| < \epsilon\}$. For an illustration see fig. 1. The parametric choice for the group regression function is twice continuously differentiable satisfying the smoothness assumptions for the pointwise asymptotic theory. Higher order smoothness can be easily achieved by constructing a higher degree polynomial spline patch. By choosing $\epsilon \to 0$ one can make this function arbitrarily close in sup norm to the worst-case function $\tilde{g}(x)$. In the simulation, I choose $\epsilon = 0.01$.

Results. Figure 2 and Figure 4 depict the exact finite sample bias of the base RD estimator for the two data generating processes (DGPs). The graphs clearly show that pooling induces asymptotically non-negligible bias. Under DGP1 estimated with the local constant estimator the bias converges at rate h while as under DGP2 estimated with the local linear the bias converges at rate h^2, as predicted by the uniform framework. The Stratified RD estimator remains unbiased in this stylized setting.

This is clearly demonstrated in Figure 3 and Figure 5. The first panel in both figures shows that the two implementations of the base RD estimator (rdrobust, RDHonest) have a bias term that converges with increasing sample size, while the Stratified RD estimator is exactly unbiased. Panels two and three demonstrate that the base RD estimator also has higher MSE and higher variance than the Stratified RD estimator. The simulation results illustrate pooling bias and the MSE improvements achievable through stratification, which will be generalized in the following theorems.

DGP 1 - Local Constant Estimation

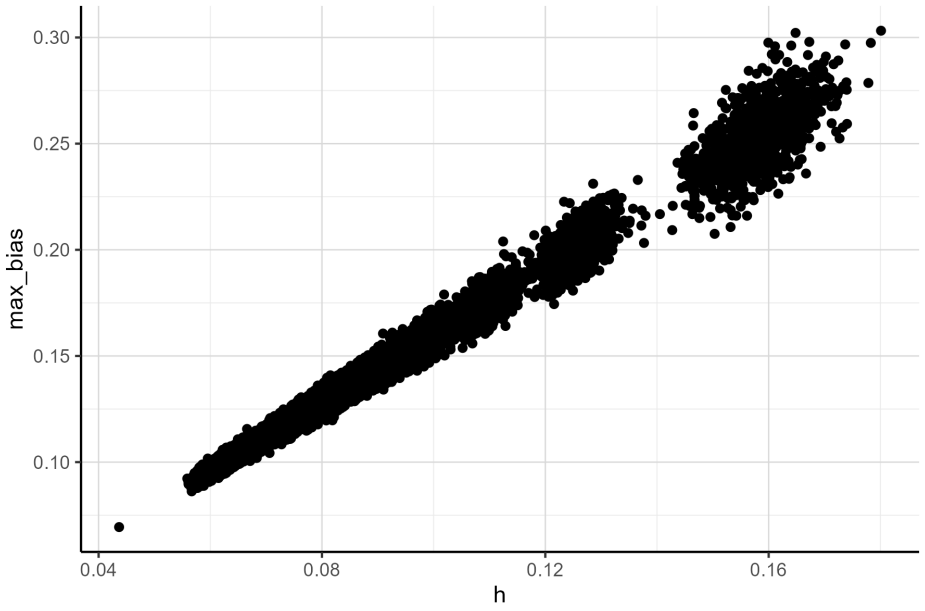


Figure 2: Finite Sample Theoretical Bias relative to Bandwidth for the base RD Estimator

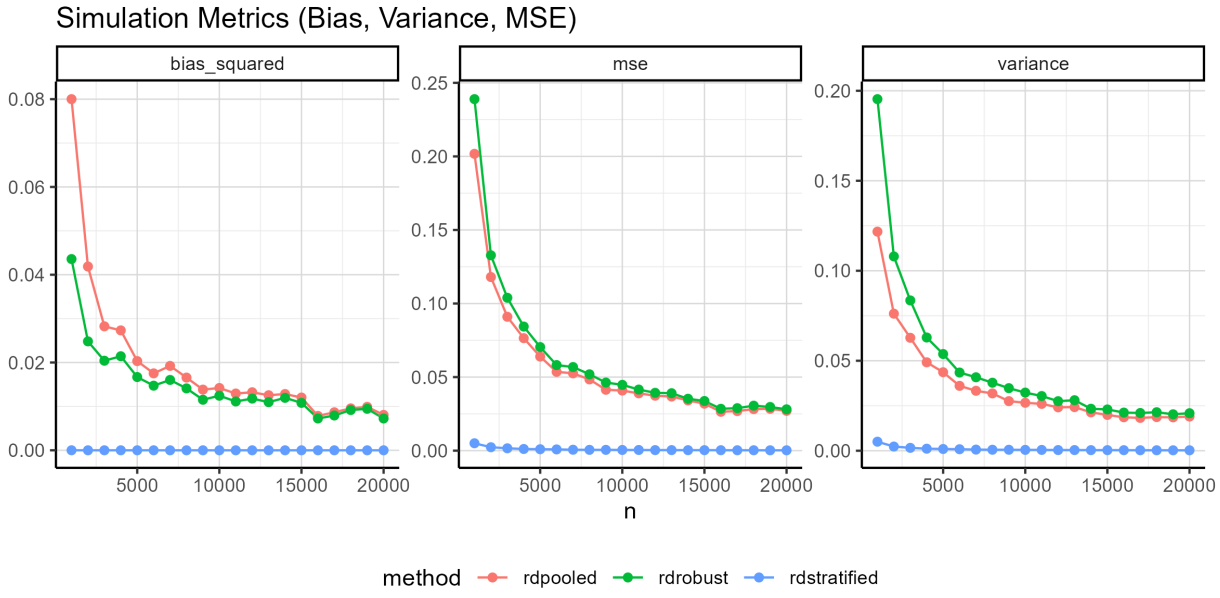


Figure 3: Squared Bias, Variance and MSE Relative to Sample Size

DGP 2 - Local Linear Estimation

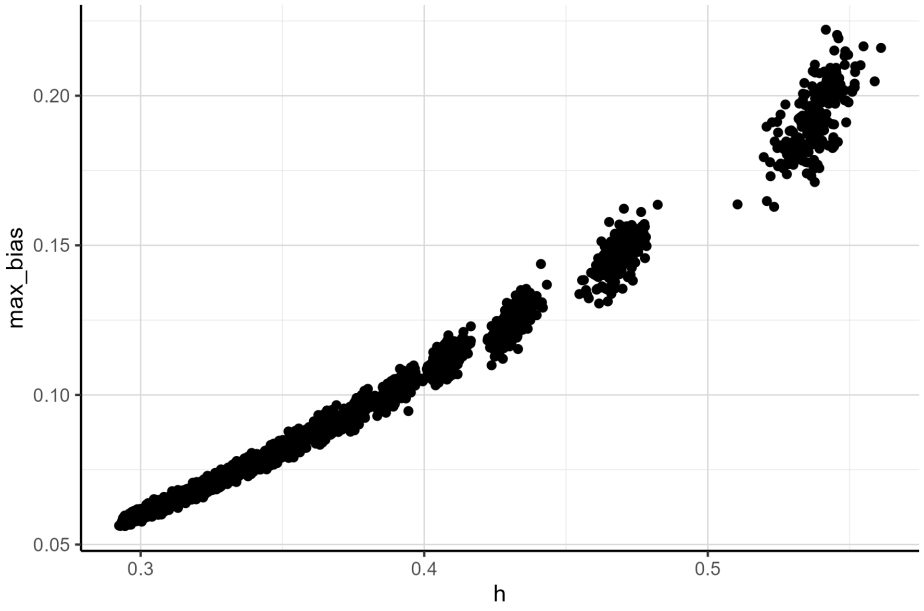


Figure 4: Finite Sample Theoretical Bias relative to Bandwidth for the base RD Estimator

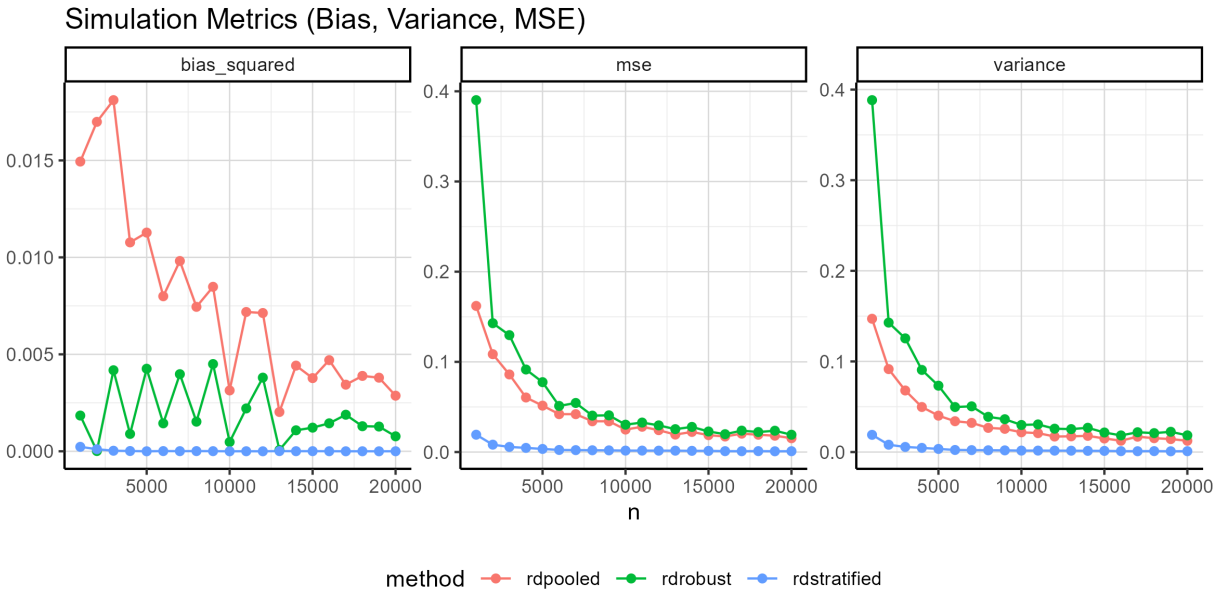


Figure 5: Squared Bias, Variance and MSE Relative to Sample Size

K Generalization to Multiple Groups

[Work in Progress]

L Additional Lemmas

Lemma L.1 (Lp-bound)**.** *Let $a, b \in \mathbb{R}$ and $p \geq 1$. Then*

$$|a+b|^p \leq 2^{p-1}\Big(|a|^p + |b|^p\Big).$$

And more generally let $\{x_1, ..., x_j, ..., x_J\} \in \mathbb{R}$ and $p \geq 1$, then

$$\left|\sum_{j=1}^{J} x_j\right|^p \leq J^{p-1} \sum_{j=1}^{J} |x_j|^p$$

Proof. By the triangle inequality,

$$|a+b| \leq |a| + |b|$$
$$|a+b|^p \leq \big(|a| + |b|\big)^p$$

Express the right-hand side as

$$(|a| + |b|)^p = 2^p \left(\frac{|a| + |b|}{2}\right)^p.$$

Since the function $x \mapsto x^p$ is convex on $[0, \infty)$ for $p \geq 1$, Jensen's inequality implies

$$\left(\frac{|a| + |b|}{2}\right)^p \leq \frac{|a|^p + |b|^p}{2}.$$

Combining these, we obtain

$$|a+b|^p \leq 2^p \cdot \frac{|a|^p + |b|^p}{2} = 2^{p-1}\Big(|a|^p + |b|^p\Big).$$

□