Composition Bias and Covariate Adjustment in Regression Discontinuity Designs

Tobias Großbölting*

November 15, 2026

Latest version available here

Abstract

In regression discontinuity designs covariates can be imbalanced in the samples used for estimation, even if their distribution is continuous in the running variable. I demonstrate that under such covariate imbalance the standard local polynomial RD estimator (base RD) implicitly compares units with different covariate values. This induces an additive term in the bias of the base RD estimator, which I refer to as composition bias. To reduce composition bias, I argue that researchers should adjust for pre-treatment covariates so that the estimator compares only units with similar covariate values. Specifically, I propose the Stratified RD estimator and prove that it has asymptotically weakly lower variance and worst-case bias than the base RD estimator, which leads to shorter bias-aware confidence intervals for the average treatment effect at the cutoff. In a simulation calibrated to replication data from a study of a merit-based financial aid program in Colombia, stratifying on parental education reduces the bias relative to the base RD estimator by 14–23%, and stratifying on both parental education and whether a student lives in a metropolitan area reduces it by 25–30%, depending on the bandwidth.

Keywords: Regression Discontinuity Designs, Heterogeneity, Covariate Imbalance, Covariate Adjustment, Stratification, Worst-Case Bias, Bias-Aware Inference, Local Polynomial Regression

JEL Codes: C14, C21

*University of Mannheim.

I thank Christoph Rothe and Paul Goldsmith-Pinkham for invaluable supervision. Further, I thank Soonwoo Kwon, Max Cytrynbaum, P. Aronow, Harrison Zhou, Jeffrey Wooldridge, Claudia Noack, Tomasz Olma, Mengshan Xu, Pablo C. Zarate, Bjarne Horst, Taylan Alpkaya and Emre Oral for useful feedback and comments.

1 Introduction

Regression discontinuity (RD) designs are observational studies in which units whose running variable exceeds a cutoff are assigned to treatment. To learn the causal effect of the treatment, researchers compare the outcomes of units just above and below this cutoff, supposing that "in the absence of treatment, persons close to the threshold are similar" (Hahn et al. 2001, p. 202). This premise is justified if the observed and unobserved covariates of the units are similar, because then so are their expected outcomes absent treatment. If the distribution of these covariates is continuous in the running variable, the average treatment effect is identified at the cutoff, because the covariates are balanced there (Lee 2008). However, estimation inevitably requires units away from the cutoff, where covariates can be imbalanced, even if continuity assumptions hold.

To illustrate the consequences, consider an observed binary covariate that splits the population into two groups, and suppose that the group composition changes smoothly with the running variable. A natural example is parental education in the setting of Londoño-Vélez et al. (2020), who study whether a scholarship for students whose test score exceeds a threshold affects college enrollment. Following standard practice, they estimate the treatment effect by fitting local polynomials in the running variable on each side of the cutoff using data from an estimation window around it (base RD). Within this window, students who have a parent with a higher education degree achieve higher test scores on average, so the share of this group increases with the test score. Under such within-window covariate imbalance, the base RD estimator implicitly compares students across the two groups, because it adjusts only for the running variable and ignores the covariate.

In this paper, I demonstrate that such comparisons across groups contribute a term, which I call composition bias, to the bias of the base RD estimator if expected outcomes given the running variable differ between groups. I therefore argue that researchers should adjust for pre-treatment covariates to reduce composition bias, unless they are willing to assume that covariates are either balanced throughout the estimation window or not predictive of the outcome given the running variable. Adjustment ensures that only units with similar covariate values are compared, the same mechanism by which covariate adjustment reduces bias in observational studies (Imbens and Rubin 2015). Specifically, I show that stratification, a form of covariate adjustment, eliminates composition bias for discrete covariates and shrinks it with the bin width for continuous covariates discretized into bins.

Although bias is the central concern for inference in RD designs, adjusting for covariates to reduce bias is not standard practice. Because the width of the estimation window, the bandwidth, is typically chosen to trade off bias against variance, the bias of the base RD estimator is of the same order as the standard deviation of the estimator even in large samples. This first-order bias distorts the coverage of conventional confidence intervals, and the recent RD literature develops methods to account for it (Calonico et al. 2014; Armstrong and Kolesár 2018; Armstrong and Kolesár 2020). Yet the literature views covariate adjustment mainly as a device for variance reduction (Armstrong and Kolesár 2018; Calonico et al. 2019; Kreiß and Rothe 2023; Ma and Yu 2024; Noack et al. 2026). Recent guidelines for empirical practice explicitly state: "The main and only goal of this practice

[covariate adjustment] is to improve efficiency and power [...], but the RD point estimate should remain unaffected" (Cattaneo et al. 2023, p. 164).

This emphasis on variance reduction reflects results derived under a pointwise asymptotic framework, in which the regression functions of the outcome and the covariate given the running variable[1] are fixed as the sample size tends to infinity. With the regression functions fixed, additively separable covariate adjustment leaves the first-order bias unchanged at odd polynomial orders, because the bias contributions of the adjustment terms cancel across the cutoff (Noack et al. 2026). For nonparametric estimation, however, Brown et al. (1997, p. 2610) caution that "fixed parameter asymptotics may be misleading" and are "not a reliable argument" for the desirability of an estimator in sufficiently large samples.

A reliable alternative is a uniform asymptotic framework, which evaluates the worst-case performance of estimators over the class of regression functions that satisfy the smoothness assumptions the researcher imposes. Such a uniform framework is well established in RD designs for conducting bias-aware inference, which accounts for the largest possible bias over that class (Armstrong and Kolesár 2018; Kolesár and Rothe 2018; Imbens and Wager 2019; Armstrong and Kolesár 2020; Noack and Rothe 2024). Whether covariate adjustment can reduce this worst-case bias remains unanswered in the literature.

The theoretical results of this paper show that covariate adjustment can reduce first-order worst-case bias in sharp RD designs within a uniform asymptotic framework.

First, I show that the worst-case bias of the base RD estimator contains a first-order composition bias term, jointly driven by covariate imbalance within the estimation window and heterogeneity in expected outcomes across covariate values. To address this bias, I propose the Stratified RD estimator, which estimates group-specific treatment effects using the base RD estimator within each group and then averages them using estimated group shares at the cutoff to obtain an estimate of the average treatment effect at the cutoff.

I then show that the Stratified RD estimator has weakly lower first-order worst-case bias and asymptotic variance than the base RD estimator. The reduction in worst-case bias is strict unless composition bias is ruled out by assumption, and the reduction in variance is strict unless the groups are homogeneous at the cutoff. The bias and variance reductions hold at any bandwidth chosen for the base RD estimator[2] and result in asymptotically shorter bias-aware confidence intervals, which account for the worst-case bias to improve coverage (Li 1989; Armstrong and Kolesár 2020).

Finally, I generalize the results to a continuous covariate by discretizing it into bins of equal width whose number should grow with the sample size, and show that, under smoothness assumptions

[1]Throughout the paper, I refer to the conditional expectation of V given W, $\mathbb{E}[V \mid W = w]$, as the regression function of V given W.

[2]The comparison uses the same bandwidth for the base RD estimator and for the first step of the Stratified RD estimator, which estimates the group-specific treatment effects. The results then only require that the second-step bandwidth for estimating the weights is proportional to the bandwidth used in the first step, with a proportionality constant that lies within an explicit interval. That interval depends only on the polynomial order and kernel used for estimation, and I show that it is nonempty for three commonly used kernels.

that bound the first derivatives, stratification likewise reduces worst-case bias and variance.[3]

To derive the results, I specify smoothness assumptions on the primitives: the distribution of the covariates and the long regression functions that condition on the running variable and covariates. The smoothness assumptions impose differentiability up to a given order, and the smoothness constants bound the derivative of that order and the heterogeneity across covariate values. Placing them on the primitives rather than on the pooled regression functions, which average over the covariates, keeps the worst-case composition bias distinguishable from the bias of approximating the long regression functions and reveals the reduction in worst-case bias through covariate adjustment.

The worst-case bias is the relevant quantity to assess the potential for bias in an empirical study, because at any sample size there exist regression functions that satisfy the assumptions of the pointwise framework and whose realized composition bias comes arbitrarily close to the worst case (Figure 1). Since the primitives are unknown, a researcher does not know how close the composition bias is to its worst-case bound in her study. The worst-case bias depends on the specified smoothness constants, so a researcher who expects that covariate adjustment has only a small effect on the bias in a given study can choose small ones. However, unless the researcher assumes no heterogeneity across covariate values or assumes the distribution of covariates to be constant in the running variable, by setting the respective bounds to zero, worst-case composition bias is of first order.

Within-window covariate imbalance is common in empirical RD studies, and it can be established by falsifying the assumption that the distribution of covariates is constant in the running variable within the estimation window. I test this assumption for binary covariates in 9 replication datasets also considered by Noack et al. (2026), using the bandwidths reported in the original papers.[4] Accounting for multiple hypothesis testing, balance is rejected for at least one covariate in 3 of the 9 studies at a 5% significance level, and in total 30 of the 95 tests reject. The rejections concentrate in studies with large samples where the tests have power (see Figure 5), so these numbers provide a conservative indication of how often within-window imbalance is present in these studies.

Researchers already worry about covariate imbalance, and balance checks, which apply the base RD estimator to a predetermined covariate to test for a discontinuity, are standard practice to assess the plausibility of continuity assumptions (Cattaneo et al. 2023). Such falsification exercises, however, cannot falsify continuity without additional assumptions (Kamat 2018; Bertanha and Moreira 2020). More fundamentally, they test balance at the cutoff, whereas composition bias is driven by imbalance within the estimation window, which current applied practice does not address. In the simple example of a binary covariate whose group composition drifts linearly through the cutoff, the estimated jump in the covariate regression function has expectation zero while the base RD estimator can carry first-order composition bias (see Example 1).

To illustrate the practical relevance of covariate adjustment and the proposed Stratified RD estimator, I conduct a calibrated simulation and a placebo exercise using the replication dataset

[3]The extension to higher-order smoothness assumptions is left for future research.

[4]For a binary covariate the regression function coincides with the conditional distribution, so that a difference in means on the two sides of the cutoff establishes that the distribution changes with the running variable. Of the 16 replication datasets considered in Noack et al. (2026), 9 include at least one binary covariate.

of Londoño-Vélez et al. (2020). In the replication dataset, both conditions for composition bias are present for parental education as a covariate. The share of students who have a parent with a higher education degree increases with the test score, which determines treatment. Given a test score, the binary indicator of whether at least one of the parents of a student has a higher education degree also predicts whether a student enrolls at a university after graduating from high school.

For the calibrated simulation, I follow Imbens and Kalyanaraman (2012) and fit quintic polynomials to estimate the regression functions in the large administrative dataset and treat them as the true model in the simulation. In the simulation, the Stratified RD estimator has 14–23% lower bias than the base RD estimator, depending on the bandwidth, while having similar variance. In a second calibrated simulation exercise, I consider both parental education and whether a student lives in a metropolitan area. In this simulation, the Stratified RD estimator has 25–30% lower bias than the base RD estimator, while having similar variance.

As a more agnostic exercise, I use a sample of students ineligible for the treatment to carry out a subsampling placebo exercise. I draw repeated subsamples from the sample of ineligible students and evaluate the bias of the Stratified and base RD estimators relative to the true treatment effect of zero. I find that the Stratified RD estimator has 13–29% lower bias while having similar variance.

1.1 Related Literature

The recent literature studies the effect of covariate adjustment on estimation under a pointwise asymptotic framework and does not establish a first-order bias reduction. Instead, the literature primarily stresses that linearly adjusting for covariates that are not affected by treatment can reduce variance (Calonico et al. 2019; Kramer and Kreiß 2023), with extensions to high-dimensional settings (Kreiß and Rothe 2023; Arai et al. 2025) and flexible additively separable adjustments (Noack et al. 2026). Ma and Yu (2024) consider balancing weights, which they show to be first-order equivalent to linear adjustment. Under the assumption that the covariate regression function has continuous derivatives at the cutoff, the adjusted and unadjusted estimators have the same first-order bias at odd polynomial orders in the pointwise framework (Noack et al. 2026). This equality in first-order bias, however, rests on a cancellation across the cutoff that does not hold uniformly over the class of functions that the standard smoothness assumptions admit (see Example 2) and is an instance of superefficiency in nonparametric function estimation (Brown et al. 1997).[5]

Besides the additively separable adjustments that most of the literature studies, Frölich and Huber (2019) propose an estimator that uses fully nonparametric covariate adjustment. Their estimator has the same two-step structure as the Stratified RD estimator, but differs in how it estimates the weights in the second step.[6] Further, Frölich and Huber (2019) conclude from their analysis in the pointwise asymptotic framework that a ranking of the estimators in terms of bias

[5] Drawbacks of pointwise asymptotics are well known in the literature on nonparametric kernel regression (Tsybakov 2009; Armstrong and Kolesár 2020) and in econometrics in general (for some notable examples, see Kamat 2018).

[6] Rather than using a local constant estimator with a specific boundary kernel, the Stratified RD estimator estimates the weights at an interior point, exploiting the assumed continuity of the covariate regression function, and tunes the bandwidth and polynomial order to the smoothness assumptions of the researcher.

seems impossible under covariate distributions that are continuous at the cutoff. The uniform framework of this paper resolves this problem by evaluating each term at its supremum, so that the elimination of composition bias unambiguously reduces worst-case bias.

Uniform asymptotics are not new to the RD literature, but the effect of covariate adjustment has not been studied under the assumptions on the primitives this paper proposes. Armstrong and Kolesár (2018) and Noack et al. (2026) use a uniform framework for bias-aware inference and consider covariate adjustment with smoothness assumptions imposed on the residualized outcome. Imbens and Wager (2019) study RD designs with multiple running variables under a smoothness assumption that couples the partial derivatives of the long regression functions. For continuous covariates, the covariate adjustment problem falls into their setup if the assignment boundary only depends on one of the considered variables with the others taking the role of covariates.

In a recent comment, Gelman and Imbens (2025, pp. 2–3) emphasize that in RD designs, as in any observational study, researchers need "to be concerned with balance on all pre-treatment predictors" and should "adjust for important pre-treatment variables". The results of this paper support that view by demonstrating that covariate-adjusted estimators can reduce bias in RD designs. Although covariates are not needed for identification, for estimation they matter as in observational studies.

Armstrong and Kolesár (2021) study estimation and inference in observational studies under general smoothness assumptions on the long regression functions. In their empirical example they impose Lipschitz conditions on the regression function in each covariate separately, which have the same structure as the assumptions proposed in this paper to make composition bias distinguishable. Their proposed estimator then minimizes worst-case finite-sample MSE by targeting both bias due to covariate imbalance and variance. Under treatment effect homogeneity their target parameter coincides with the average treatment effect at the cutoff, and their estimator with its software implementation for Lipschitz smoothness assumptions can be directly applied to the continuous covariate generalization in Section 6.2. Under treatment effect heterogeneity the target parameters differ, and I therefore propose and analyze regression-based estimators for the average treatment effect at the cutoff.[7] Developing finite-sample optimal procedures for this target parameter is left for future research.

2 Composition Bias

Consider a sharp RD in which units are treated if their running variable X_i exceeds a cutoff normalized to 0 without loss of generality, $T_i := \mathbb{1}\{X_i \geq 0\}$. The observed outcome Y_i relates to potential outcomes through $Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]T_i$, where $Y_i(0)$ is the potential outcome without treatment, which I refer to as the baseline outcome, and $Y_i(1) - Y_i(0)$ is the treatment effect. Further, assume the researcher observes a covariate Z_i that is unaffected by treatment, $Z_i = Z_i(1) = Z_i(0)$, and the data are an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$. To simplify the exposition, I

[7] Regression-based estimators have been shown to be close to optimal relative to optimal linear procedures in one-dimensional RD problems (Armstrong and Kolesár 2018).

present the case of a single binary covariate Z_i that splits the population into two groups. Section 6 generalizes the results to discrete covariates and to continuous covariates under Lipschitz assumptions on the primitives.

To formalize how a binary covariate affects estimation in RD designs, one has to consider the long and pooled regression functions. The long regression functions of baseline outcomes $m(x,z) := \mathbb{E}[Y_i(0) \,|\, X_i = x, Z_i = z]$ and of treatment effects $\tau(x,z) := \mathbb{E}[Y_i(1) - Y_i(0) \,|\, X_i = x, Z_i = z]$ condition on the running variable and the covariate. The pooled regression functions $\bar{m}(x) := \mathbb{E}[Y_i(0) \,|\, X_i = x]$ and $\bar{\tau}(x) := \mathbb{E}[Y_i(1) - Y_i(0) \,|\, X_i = x]$ condition only on the running variable and hence average over the groups, weighting them by their group shares at x, $\bar{m}(x) = \sum_{z \in \{0,1\}} m(x,z) f_{Z|X}(z \,|\, x)$. For group 1, that share is given by the covariate regression function, $g(x) := \mathbb{E}[Z_i \,|\, X_i = x] = f_{Z|X}(1 \,|\, x)$, which I assume to be continuous at the cutoff. The parameter of interest is the conditional average treatment effect at the cutoff, which is a weighted average of the group-specific treatment effects,

$$\tau := \bar{\tau}(0) = g\,\tau_1 + (1-g)\,\tau_0, \tag{1}$$

where $\tau_j := \tau(0,j)$ is a shorthand for the group-specific treatment effects and $g := g(0)$ and $(1-g)$ denote the group shares at the cutoff.

Base RD Estimator. This parameter is typically estimated using the base RD estimator, which fits local polynomial regressions on the treatment and control sides and reports the difference in the fitted values at the cutoff as an estimate of τ,

$$\hat{\tau}(h) := e_1^\top \arg\min_{\beta \in \mathbb{R}^{2q+2}} \sum_{i=1}^{n} k(X_i/h)(Y_i - S_q(X_i)^\top \beta)^2.$$

Here, k denotes a symmetric, nonnegative, and bounded kernel with bounded support, h the bandwidth, $S_q(X_i) = (T_i, 1, X_i, T_i X_i, \ldots, X_i^q, X_i^q T_i)^\top$ the q-th order polynomial in the running variable X_i fully interacted with T_i, and $e_1 := (1, 0, \ldots, 0)^\top$ selects the coefficient of T_i.

For illustration consider the local linear base RD estimator ($q = 1$), whose bias depends on the second derivatives of the pooled regression functions within the bandwidth. The estimator admits a weighted-sum representation $\hat{\tau}(h) = \sum_{X_i \geq 0} w_i^+(h) Y_i - \sum_{X_i < 0} w_i^-(h) Y_i$, with weights that eliminate the linear terms of a Taylor expansion of the pooled regression functions. The weights depend on the data only through the realizations of the running variable and taking the expectation conditional on $\mathbf{X} := \{X_i\}_{i=1}^n$ gives the conditional finite-sample bias:

$$\mathbb{E}[\hat{\tau}(h) \,|\, \mathbf{X}] - \tau = \sum_{X_i \geq 0} w_i^+(h) \int_0^{X_i} [\bar{m}''(s) + \bar{\tau}''(s)](X_i - s)ds - \sum_{X_i < 0} w_i^-(h) \int_0^{X_i} \bar{m}''(s)(X_i - s)ds. \tag{2}$$

Since the pooled regression functions average the long regression functions with the group weights, their derivatives include interactions between changes in group composition and group heterogeneity in baseline outcomes, $\Delta m(x) := m(x,1) - m(x,0)$, and treatment effects, $\Delta\tau(x) := \tau(x,1) - \tau(x,0)$.

The second derivatives of the pooled regression functions therefore decompose into

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

$$\bar{\tau}''(x) = \underbrace{\tau''(x,1)g(x) + \tau''(x,0)(1-g(x))}_{A} + \underbrace{2\Delta\tau'(x)g'(x)}_{B} + \underbrace{\Delta\tau(x)g''(x)}_{C}. \tag{4}$$

The conditional bias therefore consists of three parts that correspond to the terms in the decomposition of the derivatives: (A) induces what I call approximation bias of the long regression functions, (B) composition bias, and (C) weight-estimation bias, which arises because the base RD estimator implicitly estimates the group shares at the cutoff when estimating τ.

To approximate the conditional bias in large samples with a closed-form expression, I derive its first-order term. Following Armstrong and Kolesár (2020), I match the polynomial order q of the estimator to the order p of the derivative that the smoothness assumptions restrict and I impose a bandwidth that shrinks with the sample size at the rate that balances bias and variance. I additionally assume a continuous and positive density of the running variable at the cutoff.

Assumption 1 (Bandwidths)**.** The polynomial order is $q = p - 1$, and the bandwidth sequence satisfies

$$0 < \liminf_n h\, n^{1/(2p+1)} \leq \limsup_n h\, n^{1/(2p+1)} < \infty.$$

Assumption 2 (Marginal Density)**.** The marginal density of the running variable f_X is continuous and positive at the cutoff.

Under Assumptions 1 and 2 the estimator weights any unit proportionally to the equivalent kernel evaluated at its scaled running variable in large samples (Lemma E.4):

$$w_i^{\pm}(h) = \frac{1}{nh}\frac{k_q^*(|X_i|/h)}{f_X(0)} + o_p((nh)^{-1}), \tag{5}$$

uniformly in i, where the equivalent kernel is $k_q^*(u) := e_1^\top \left(\int_0^\infty m_q(t)\, m_q(t)^\top k(t)\, dt\right)^{-1} m_q(u)\, k(u)$ with $m_q(u) := (1, u, \ldots, u^q)^\top$. The first-order bias is then derived by replacing the finite-sample weights in equation (2) with their limit. This gives a closed-form expression that depends on the derivatives of the regression functions in equation (3) and on the integrals of the equivalent kernel, the kernel constants.

2.1 Stylized Examples

To illustrate how a changing group composition and group heterogeneity induce first-order composition bias, consider the base RD estimator with a uniform kernel and a running variable supported on $[-1, 1]$ in the following two stylized examples.

Example 1 (Composition Bias in a Linear Setting)**.**
Suppose that the group composition $g(x) = g + \theta x$ changes linearly in the running variable,[8] that both groups share the same linear baseline regression function $m(x, z) = \alpha + \beta x$, and that only group 1 is affected by the treatment, $\tau(x, z) = (\tau_1 + \delta x)z$. Averaging over groups gives $\bar{m}(x) = \alpha + \beta x$ and $\bar{\tau}(x) = \tau_1 g + (\tau_1\theta + \delta g)\, x + \delta\theta\, x^2$. The pooled treatment effect function is quadratic, even though all primitives are linear.

This composition-induced curvature biases the base RD estimator. Plugging the derivatives $\bar{m}''(x) = 0$ and $\bar{\tau}''(x) = 2\delta\theta$ into equation (2) gives its finite-sample bias, which contains only composition bias, and standard calculations yield its first-order term:

$$\mathbb{E}[\hat{\tau}(h) \,|\, \mathbf{X}] - \tau = 2\delta\theta \sum_{X_i \geq 0} w_i^+(h)\, \frac{X_i^2}{2} = -\delta\theta\, h^2 \frac{1}{6} + o_p(h^2), \tag{6}$$

where $1/6$ is the kernel constant of the uniform kernel. Clearly, the covariate regression function is continuous at the cutoff and the standard falsification exercise, the base RD estimator applied to Z_i, is unbiased for zero. Conventional practice would leave practitioners unconcerned about covariates while the base RD estimator carries first-order composition bias.

Example 1 rests on a treatment effect function whose slope differs between the two groups. However, in many empirical studies there is also substantial heterogeneity in the baseline regression functions. For example, in the dataset studied by Londoño-Vélez et al. (2020), the university enrollment rate of students is predicted by the educational background of their parents irrespective of receiving treatment (see Figure 2a).

Example 2 (Composition Bias through Baseline Heterogeneity)**.**
Suppose that the long baseline regression function is constant in x, $m(x, z) = \mu_0(1 - z) + \mu_1 z$, and heterogeneous across groups, $\mu_0 \neq \mu_1$, and that the treatment effect $\tau(x, z) = \tau$ is homogeneous. Further, suppose that the group composition $g(x)$ is a nonparametric function, restricted only by the standard smoothness assumption that g is twice continuously differentiable with $|g''(x)|$ globally bounded by a constant $L > 0$. The pooled regression functions are $\bar{m}(x) = \mu_0 + \Delta m\, g(x)$ and $\bar{\tau}(x) = \tau$, where $\Delta m := \mu_1 - \mu_0$ denotes the heterogeneity in baseline outcomes between groups. The conditional finite-sample bias in equation (2) is then entirely driven by group heterogeneity and changes in the group composition:

$$\mathbb{E}[\hat{\tau}(h) \,|\, \mathbf{X}] - \tau = \Delta m \left(\sum_{X_i \geq 0} w_i^+(h) \int_0^{X_i} g''(s)(X_i - s)ds - \sum_{X_i < 0} w_i^-(h) \int_0^{X_i} g''(s)(X_i - s)ds \right). \tag{7}$$

One way to approximate this bias in large samples is to evaluate its first-order term holding the

[8] To ensure that the shares of the groups are nonnegative throughout the support of the running variable, I restrict $|\theta| \leq \min\{g, 1 - g\}$.

unknown covariate regression function $g(x)$ fixed (pointwise asymptotics):

$$\mathbb{E}[\hat{\tau}(h) \mid \mathbf{X}] - \tau = -\frac{h^2}{12}\Big(\Delta m g''(0) - \Delta m\, g''(0)\Big) + o_p(h^2) = o_p(h^2). \tag{8}$$

Because $g(x)$ is assumed to be twice continuously differentiable, the second derivatives on both sides of the cutoff converge to the same value and their bias contributions cancel across the cutoff, so that the first-order term capturing composition bias is zero. It is tempting to conclude that the conditional bias converges to zero faster than the squared bandwidth h^2 and that under standard bandwidth choices (Assumption 1) it hence will be small relative to the standard deviation in large samples.

However, this cancellation only holds in the asymptotic limit when holding the group composition fixed. At any positive bandwidth, composition bias depends on the shape of the unknown function $g(x)$ within the estimation window. Specifically, the size of the composition bias depends on the integral of g'' over the estimation window, and within that window g'' can differ between the treated and control sides, even when it is continuous. To gauge the potential for composition bias in an empirical study one can instead consider the worst-case conditional finite-sample bias over the class of all functions allowed under the smoothness assumptions and then study its behavior in large samples (uniform asymptotics). Taking the supremum over all g which are twice continuously differentiable and whose second derivative is globally bounded $|g''|_\infty \leq L$ gives the worst-case conditional bias:

$$\sup_{g(x)} \mathbb{E}[\hat{\tau}(h) \mid \mathbf{X}] - \tau = -\frac{L|\Delta m|}{2}\Big[\sum_{X_i \geq 0} w_i^+(h)\, X_i^2 + \sum_{X_i < 0} w_i^-(h)\, X_i^2\Big] \tag{9}$$

$$= h^2 L|\Delta m|\frac{1}{6} + o_p(h^2), \tag{10}$$

which has a first-order composition bias term that depends on the product of group heterogeneity in baseline outcomes and the assumed bound on the curvature of the group composition.

The two examples demonstrate how a group composition that varies with the running variable can bias the base RD estimator if the groups are heterogeneous, even if the long regression functions are linear or constant.

2.2 Uniform versus Pointwise Asymptotics

The displays in equation (8) and equation (10) of Example 2 are derived under the same assumptions but in two different asymptotic frameworks and give different conclusions on whether composition bias is a first-order problem. Which of the two frameworks should guide empirical practice?

Suppose a researcher asks how large composition bias is in her study. The pointwise framework asserts that composition bias will be small relative to the standard deviation and hence negligible in large enough samples. However, it does not provide a sample size at which this would hold true and thus does not give clear guidance to the researcher. In contrast, the worst-case bias of the

uniform framework gives a valid upper bound on the composition bias in any finite sample, as long as the smoothness assumptions hold. The bound is also informative for empirical practice, because at any sample size there are functions g that satisfy the smoothness assumptions for which the conditional bias of the base RD estimator comes arbitrarily close to the worst-case composition bias in Example 2.

To make this concrete, note that the worst-case bias in Example 2 is attained by the function $\tilde{g}(x) = -\frac{L}{2}\, x^2 \operatorname{sign}(x) + \frac{1}{2}$, which is a valid conditional density within a small bandwidth around the cutoff and whose second derivative jumps from L to $-L$ at the cutoff, making it not twice continuously differentiable (Armstrong and Kolesár 2020). A natural objection is that the worst case is only attained by a function that does not satisfy the differentiability assumptions. However, a simple construction interpolates the jump over an interval of width 2ϵ around the cutoff (Figure 1), which gives a function that is twice continuously differentiable with a second derivative that is bounded by L, and for which the conditional bias comes arbitrarily close to the worst-case bias.[9]

$$g_\epsilon(x) = \begin{cases} \tilde{g}(x) - \frac{L\epsilon^2}{6} \operatorname{sign}(x) & |x| \geq \epsilon, \\ -\frac{L}{6\epsilon}\, x^3 - \frac{L\epsilon}{2}\, x + \frac{1}{2} & |x| < \epsilon. \end{cases} \tag{11}$$

Since g is unknown, the researcher does not know whether composition bias is small or large in her study, and the worst-case bound is the only guarantee on the size of the bias she has.

The disagreement between the two asymptotic frameworks is not specific to Example 2. When the pooled regression functions are assumed twice continuously differentiable at the cutoff, as is standard in the RD literature, the terms involving $\bar{m}''(x)$ in equation (2) likewise cancel, so that the first-order bias is driven only by curvature in the pooled treatment effect function at the cutoff $\bar{\tau}''(0)$ under pointwise asymptotics.[10] It is for the same reason that linear covariate adjustment or flexible versions thereof do not affect the first-order bias of the base RD estimator in the pointwise asymptotic framework when the polynomial order is odd (Noack et al. 2026).[11] Yet the pointwise cancellation follows the same mechanism as the superefficiency example for nonparametric density estimation in the paper by Brown et al. (1997), in which they caution against judging the performance of an estimator in large samples based on pointwise asymptotic statements.

For these reasons I consider the worst-case bias of estimators when analyzing the effect of covariate adjustment in RD designs.

[9] Since the composition bias term in equation (7) is linear in g, and since $\sup_x |g_\epsilon(x) - \tilde{g}(x)| = L\epsilon^2/6$, its value at g_ϵ converges to the worst-case bound as $\epsilon \to 0$, at any fixed sample size.

[10] Based on this, one could conclude that, in settings where the pooled treatment effect function has a zero second derivative, for example in placebo regressions or in the partially linear model considered by Ghosh et al. (2025), the local linear base RD estimator has no first-order bias, so that conventional confidence intervals are asymptotically valid under standard bandwidth choices (Assumption 1). In these settings bias-corrected inference methods are nevertheless recommended and used in practice (Cattaneo et al. 2020).

[11] For even-order polynomials, the leading bias of the linearly adjusted estimator can differ from that of the base RD estimator even in the pointwise framework.

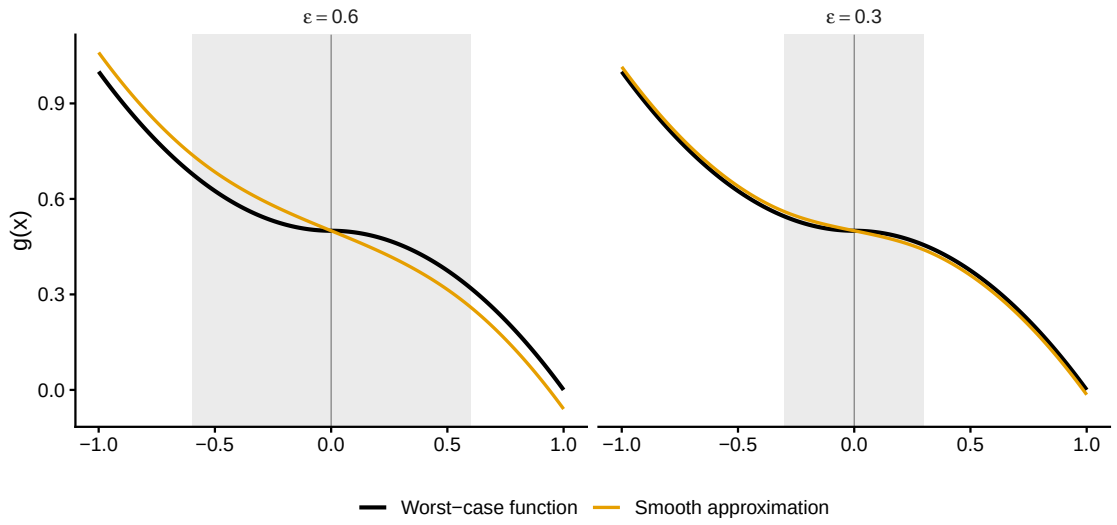


Figure 1: Worst-Case Function and Smooth Approximations

2.3 Worst-Case Composition Bias

To understand how composition bias contributes to the worst-case bias of the base RD estimator in a general setting, note that the conditional bias of the base RD estimator with polynomial order $q = p - 1$ depends on the p-th derivative of the pooled regression functions. Those derivatives in turn depend on the primitives, the long regression functions and the group composition:

$$\bar{m}^{(p)}(x) = m^{(p)}(x, 1)\, g(x) + m^{(p)}(x, 0)\, (1 - g(x)) + R_p(x), \tag{12}$$

with $R_p(x) := \sum_{k=0}^{p-1} \binom{p}{k} \Delta m^{(k)}(x)\, g^{(p-k)}(x)$ and $\Delta m^{(k)}(x) := m^{(k)}(x, 1) - m^{(k)}(x, 0)$. The first two terms form a convex combination of within-group derivatives, weighted by the group composition $g(x)$. Composition bias is induced by the terms in $R_p(x)$: cross-products of group heterogeneity in the long regression functions and the derivatives of the group composition. The same decomposition holds for the pooled treatment effect function $\bar{\tau}$, except that the term for $k = 0$, $\Delta\tau(x)\, g^{(p)}(x)$, induces weight-estimation bias.

To control the conditional bias one needs to impose smoothness assumptions on the different terms in equation (12). As in the examples, a natural way of imposing smoothness on the regression functions is to assume that they are p times continuously differentiable and to impose a bound on the p-th derivative that holds uniformly across the domain of the running variable. In the rest of the paper I follow Armstrong and Kolesár (2020) and work with the closure of this class of functions, which for $p \in \mathbb{N}^+$ gives the (integer) Hölder class:

$$\mathcal{F}_p(L) := \left\{ f : |f^{(p-1)}(x) - f^{(p-1)}(y)| \leq L|x - y| \quad \forall x, y \in \mathbb{R} \right\}. \tag{13}$$

Additionally, one needs to bound the heterogeneity between groups to control the terms in

equation (12) that depend on the difference in their regression functions and how it changes with the running variable.

Assumption 3 (Smoothness of the Primitives)**.** For $p \in \mathbb{N}^+$, $\underline{g} \in (0, 1/2]$ and nonnegative constants $M, L, C_{\Delta m}, C_{\Delta\tau}, L_j, C_{\Delta,j}$, the following holds for all $z \in \{0,1\}$ and $x \in \mathbb{R}$.

(i) $m(\cdot, z),\ m(\cdot, z) + \tau(\cdot, z) \in \mathcal{F}_p(M)$.

(ii) $g \in \mathcal{F}_p(L)$ and $g(0) \in \left[\underline{g}, 1 - \underline{g}\right]$.

(iii) $|\Delta\tau(x)| \le C_{\Delta\tau}$.

(iv) $|\Delta m(x)| \le C_{\Delta m}$.

(v) For all $j \in \{1, \dots, p-1\}$, $|\Delta m^{(j)}(x) + \Delta\tau^{(j)}(x)| \le C_{\Delta,j}$, $|\Delta m^{(j)}(x)| \le C_{\Delta,j}$ and $|g^{(j)}(x)| \le L_j$.

The smoothness constants restrict the primitives jointly and hence should be compatible.[12] The primitives (m, τ, g) that satisfy Assumption 3 form the class $\mathcal{F}_p^R(M, \mathbf{C}, \mathbf{L})$, with heterogeneity bounds $\mathbf{C} := (C_{\Delta m}, C_{\Delta\tau}, C_{\Delta,1}, \dots, C_{\Delta,p-1})$ and composition bounds $\mathbf{L} := (L, L_1, \dots, L_{p-1})$.[13]

Under the smoothness assumptions on the primitives, one can bound the p-th derivative of the pooled regression functions by applying the triangle inequality.[14] The resulting upper bound on the worst-case bias is valid but conservative, as there exists no function that attains all the bounds in Assumption 3 at once. Further, an upper bound only limits the size of worst-case composition bias but does not show that it is first-order. I therefore construct an attainable lower bound on the worst-case bias of the base RD estimator, by setting all the constants $C_{\Delta,j} = 0$ for $j \in \{1, \dots, p-1\}$. This restricts $\Delta m(x)$ and $\Delta\tau(x)$ to be constant in x, so the regression functions of the two groups differ only by a vertical shift, and the bound implied by the triangle-inequality argument becomes attainable.

Proposition 2.1 (Base RD Worst-Case Bias Lower Bound)**.** *Assume Assumptions 1 to 3. Define* $\mathbf{C}^* := (C_{\Delta m}, C_{\Delta\tau}, 0, \dots, 0)$*. Then*

$$\begin{aligned}
\overline{\text{Bias}}(\hat{\tau}(h)) &:= \sup_{(m,\tau,g)\in\mathcal{F}_p^R(M,\mathbf{C},\mathbf{L})} |\mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}]| \\
&\ge \sup_{(m,\tau,g)\in\mathcal{F}_p^R(M,\mathbf{C}^*,\mathbf{L})} |\mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}]| \\
&= h^p\big(b_B^{\text{low}} + o_p(1)\big),
\end{aligned}$$

[12] For instance, assuming $L_j = 0$ for some $j \in \{1, \dots, p-1\}$ implies that $g^{(l)} = 0$ for $l \ge j$, so that the bound L is not binding. Whenever $L > 0$ I therefore assume $L_j > 0$ for all $j \in \{1, \dots, p-1\}$.

[13] The class also depends on $\underline{g}$, but the researcher does not need to know or choose it. I suppress it in the class notation because neither the worst-case bias statements nor the asymptotic variances depend on it.

[14] An alternative approach is to directly impose a single bound $|\bar{m}^{(p)}(x)| \le \bar{M}$, which places a joint constraint on the within-group p-th derivatives and the cross-terms causing composition bias in $R_p(x)$. Besides imposing implicit restrictions on the relative signs of the terms, such an assumption also makes it impossible to quantify the extent of composition bias.

where the lower bound on the first-order term is

$$b_B^{\text{low}} := \frac{1}{p!}\Big[2M\, B_{p,q}(k) + 2C_{\Delta m} L\, B_{p,q}(k) + C_{\Delta\tau} L\, B_{p,q}(k)\Big], \tag{14}$$

and $B_{p,q}(k) := p \int_0^\infty \left| \int_t^\infty k_q^*(u)\,(u-t)^{p-1}\, du \right| dt$ *is the boundary worst-case bias kernel constant as in Armstrong and Kolesár (2020, Thm. 3.1).*

The first-order term of this lower bound is the sum of the three sources of bias. The term $2M$ captures bias due to approximating the long regression functions, $2C_{\Delta m}L$ captures composition bias the estimator incurs by implicitly comparing units from different groups, and $C_{\Delta\tau}L$ captures the weight-estimation bias the estimator incurs for implicitly estimating the group shares of the weighted average of heterogeneous treatment effects. All three terms are first-order and the share of composition bias crucially depends on the assumptions of the researcher.[15]

Given Proposition 2.1, a researcher can deal with composition bias in three ways. The first is to rule out composition bias by assumption, imposing that the groups are homogeneous or that the group composition does not change with the running variable. In the study of Londoño-Vélez et al. (2020), neither assumption is plausible, since parental education predicts university enrollment of students and the share of students with college-educated parents increases with the test score (Figure 2). The second is to account for composition bias in inference, by constructing bias-aware confidence intervals for the base RD estimator that inflate the interval to account for the worst-case bias in Proposition 2.1. These intervals are valid, but their length grows with the composition bias term. The third is to reduce composition bias by adjusting for the covariate.

3 Covariate Adjustment

In this section I consider two forms of covariate adjustment that differ in how flexibly they adjust for the covariate, and hence in how much composition bias they can remove. The existing literature has recommended adding covariates linearly to the local polynomial regression of the base RD estimator to reduce variance (Calonico et al. 2019). However, as demonstrated below, linear covariate adjustment does not eliminate composition bias in Example 1. As an alternative form of covariate adjustment, I propose the Stratified RD estimator and show that it fully eliminates composition bias for discrete covariates. The reason is that it adjusts more flexibly for the covariate than linear adjustment, fitting separate regression functions within each group, and thereby effectively exploits the smoothness that Assumption 3 imposes.

[15] Example 2 is the special case of the proposition with $p = 2$, $M = 0$, $C_{\Delta\tau} = 0$, and $C_{\Delta,1} = 0$, and using a uniform kernel for which the bound reduces to equation (10).

Linear Adjustment RD Estimator. The linear adjustment RD estimator proposed by Calonico et al. (2019) is defined by the following weighted least squares problem:

$$\hat{\tau}_{\text{lin}}(h) = e_1^\top \arg\min_{\beta \in \mathbb{R}^{2q+2},\, \gamma \in \mathbb{R}} \sum_{i=1}^{n} k(X_i/h)(Y_i - S_q(X_i)^\top \beta - \gamma Z_i)^2. \tag{15}$$

Stratified RD Estimator. The Stratified RD estimator first estimates the treatment effect within each group separately, ensuring that only units from the same group are compared. It then averages the group-specific estimates, weighted by the estimated group shares at the cutoff, to obtain an estimate of τ. By stratifying on the covariate, it directly targets the long regression functions, which removes group differences in expected outcomes from the error term and thereby can reduce not only bias but also variance.

Formally, let $\hat{\tau}_j(h_j)$ denote the base RD estimator estimated within group j with bandwidth h_j, and $\hat{g}(h_z)$ the estimate of the group weight at the cutoff, $g(0) = f_{Z|X}(1 \,|\, 0)$, from a local polynomial regression for an interior point with bandwidth h_z,

$$\hat{\tau}_j(h_j) = e_1^\top \arg\min_{\beta \in \mathbb{R}^{2q+2}} \sum_{i=1}^{n} k(X_i/h_j)(Y_i - S_q(X_i)^\top \beta)^2 \mathbb{1}_{\{Z_i = j\}}, \tag{16a}$$

$$\hat{g}(h_z) = e_1^\top \arg\min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^{n} k(X_i/h_z)(Z_i - \tilde{S}_q(X_i)^\top \beta)^2, \tag{16b}$$

where $\tilde{S}_q(X_i) = (1, X_i, \ldots, X_i^q)^\top$. For the theoretical analysis I set the bandwidth for estimating group-specific treatment effects to that of the base RD estimator, $h_j = h$ for all groups throughout the paper.[16] The Stratified RD estimator in the two-group case is then given by

$$\hat{\tau}_S(h, h_z) = \hat{\tau}_1(h)\hat{g}(h_z) + \hat{\tau}_0(h)(1 - \hat{g}(h_z)). \tag{17}$$

Since Assumption 3 imposes smoothness assumptions of the same order p for $m(x, z)$, $\tau(x, z)$, and $g(x)$, the MSE-optimal rate at which the bandwidths shrink with the sample size is the same for the first and second estimation steps. I impose this through a proportionality assumption on the bandwidths h and h_z.

Assumption 4 (Weight-Estimation Bandwidths)**.** The bandwidth h_z for estimating the weights is proportional to the bandwidth for estimating the treatment effects, $h_z = \kappa h$, for some $\kappa > 0$.

Covariate adjustment requires positive density of the running variable in both groups of the binary covariate at the cutoff. Further, for the asymptotic results I impose regularity conditions on the error term $e_i := Y_i - m(X_i, Z_i) - \tau(X_i, Z_i)T_i$ of the long regression functions. I denote the conditional variance of the error term by $\sigma_L^2(x, z) := \text{Var}(e_i \,|\, X_i = x, Z_i = z)$, and its average over groups by $\sigma_L^2(x) := \mathbb{E}[\sigma_L^2(x, Z_i) \,|\, X_i = x]$.

[16] Generally, each group can have its own bandwidth h_j, which can be chosen separately for each group to minimize the MSE numerically. The theoretical analysis proves something stronger by showing that the Stratified RD estimator outperforms the base RD estimator even at the suboptimal bandwidth choice $h_j = h$ for all groups.

Assumption 5 (Regularity Conditions on the Long Error Terms)**.** The error terms $\{e_i\}_{i=1}^n$ have bounded conditional moments $\mathbb{E}[|e_i|^{2+\eta} \,|\, X_i = x, Z_i = z] \leq 1/\eta$ for all x and $z \in \{0,1\}$, for some $\eta > 0$. The one-sided limits of the conditional variance $\sigma^2_{L,+}(0,z) := \lim_{x \downarrow 0} \sigma^2_L(x,z)$ and $\sigma^2_{L,-}(0,z) := \lim_{x \uparrow 0} \sigma^2_L(x,z)$ exist and are strictly positive for all $z \in \{0,1\}$.

The required regularity conditions on the joint density $f_{X,Z}$, the covariate error term $r_i := Z_i - g(X_i)$ of the covariate regression function and the pooled error terms $u_i := Y_i - \bar{m}(X_i) - \bar{\tau}(X_i)T_i$ follow from the previous assumptions (see Lemma D.1). I denote the conditional variance of the respective error terms as $\sigma^2_Z(x) := \text{Var}(r_i \,|\, X_i = x)$ and $\sigma^2_P(x) := \text{Var}(u_i \,|\, X_i = x)$.

3.1 Covariate Adjustment in the Stylized Examples

Since the long regression functions are linear within group in the stylized examples, it seems intuitive that stratifying and estimating a local linear base RD estimator within each group should work well. The continued examples below formalize that intuition for the local linear implementation of the Stratified RD estimator, $q = 1$, and contrast the effect of stratification on bias with that of linear covariate adjustment. The leading bias of the Stratified RD estimator, established in Proposition 3.1, is:

$$LB := g\, \mathbb{E}[\hat{\tau}_1(h) - \tau_1 \,|\, \mathbf{X}, \mathbf{Z}] + (1-g)\, \mathbb{E}[\hat{\tau}_0(h) - \tau_0 \,|\, \mathbf{X}, \mathbf{Z}] + [\tau_1 - \tau_0]\, \mathbb{E}[\hat{g}(h_z) - g \,|\, \mathbf{X}]. \tag{18}$$

The first two terms are the conditional finite-sample biases of the treatment effect estimation within each group, which are weighted by the group shares at the cutoff, and depend on term (A) of the derivative in equation (3). The third term is the bias of estimating the group shares conditional on the realizations of the running variable multiplied by the treatment effect heterogeneity and depends on term (C). Note that the leading bias of the Stratified RD estimator does not include term (B), because the estimator only compares units within the same group and thereby avoids composition bias.

Standard least squares algebra shows that the linear adjustment RD estimator can be rewritten as $\hat{\tau}_{\text{lin}}(h) = \hat{\tau}(h) - \hat{\gamma}(h)\hat{\tau}_Z(h)$, where $\hat{\gamma}(h)$ is the estimated coefficient on Z_i and $\hat{\tau}_Z(h)$ is the base RD estimator applied to Z_i. Since $\hat{\gamma}(h)$ converges in probability to the population coefficient γ (Calonico et al. 2019, Lemma SA-6), the term $[\hat{\gamma}(h) - \gamma]\,\hat{\tau}_Z(h)$ is of smaller order than the standard deviation. As for the Stratified RD estimator, I derive the leading bias of the linear adjustment RD estimator from its linearization,

$$LB_{\text{lin}} := \mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}] - \gamma\, \mathbb{E}[\hat{\tau}_Z(h) \,|\, \mathbf{X}]. \tag{19}$$

Example 1 (continued).
Because the long regression functions are assumed linear in Example 1, the within-group estimates of the treatment effects in the first step of the estimator are unbiased, $\mathbb{E}[\hat{\tau}_1(h) - \tau_1 \,|\, \mathbf{X}, \mathbf{Z}] = \mathbb{E}[\hat{\tau}_0(h) - \tau_0 \,|\, \mathbf{X}, \mathbf{Z}] = 0$. Also, the group composition is assumed to be linear in x, so that the

weight estimation is likewise unbiased, $\mathbb{E}[\hat{g}(h_z) - g \,|\, \mathbf{X}] = 0$. All three terms of LB are zero for any bandwidths h and h_z. Under the conditions of Proposition 3.1, the Stratified RD estimator has no first-order bias in this example, while the base RD estimator carries first-order composition bias, as shown in equation (6).

Because the group composition is assumed to be linear, the conditional bias of the base RD estimator applied to the covariate is zero, $\mathbb{E}[\hat{\tau}_Z(h) \,|\, \mathbf{X}] = 0$. By equation (19), the linear adjustment RD estimator therefore has the same first-order composition bias as the base RD estimator. Intuitively, linear adjustment does not include separate slope coefficients for each group, but differences in the slopes of the regression functions of the two groups cause the composition bias in this example.

Example 2 (continued).
Because the long regression functions are constant and the treatment effect is homogeneous across groups, the Stratified RD estimator is unbiased for τ conditional on the realizations of the running variable and covariate at any bandwidths, $\mathbb{E}[\hat{\tau}_S(h, h_z) \,|\, \mathbf{X}, \mathbf{Z}] = \hat{g}(h_z)\tau + (1 - \hat{g}(h_z))\tau = \tau$. Likewise, the linear adjustment RD estimator is unbiased for τ conditional on the realizations of the running variable and covariate at any bandwidth. The constant group difference Δm is captured by the coefficient on Z_i, so that the linear model underlying the least squares problem in equation (15) is correctly specified. Both covariate-adjusted estimators remove all conditional bias, while the base RD estimator retains first-order worst-case bias.[17]

Because the Stratified RD estimator eliminates composition bias in both Example 1 and Example 2, I focus on establishing its asymptotic distribution and worst-case bias in the remainder of this section and compare its performance to that of the base RD estimator in Section 4. An analysis of the worst-case bias of the linear adjustment RD estimator is left for future research.

3.2 Asymptotic Distribution and Worst-Case Bias of Stratified RD

Linearizing the Stratified RD estimator around the true population values $\tau_j := \tau(0, j)$ for $j \in \{0, 1\}$ and $g := g(0)$ shows that its estimation error depends at first order on the error from estimating the group-specific treatment effects at known weights and the error from estimating the weights, scaled by the treatment effect heterogeneity at the cutoff:

$$\hat{\tau}_S(h, h_z) - \tau = g\,[\hat{\tau}_1(h) - \tau_1] + (1 - g)\,[\hat{\tau}_0(h) - \tau_0] + [\tau_1 - \tau_0]\,[\hat{g}(h_z) - g] + o_p((nh)^{-1/2}). \tag{20}$$

These first-order terms appear in the asymptotic variance and leading bias of the Stratified RD estimator.

[17] More generally, if one assumes $M = C_{\Delta\tau} = 0$, but $C_{\Delta m} > 0$ and $L > 0$, the treatment effect at the cutoff can be estimated at the parametric rate using the Stratified RD estimator and only at a nonparametric rate using the base RD estimator. While this result only holds under stylized assumptions and might not be useful for empirical practice, it clearly shows the benefits of stratification.

Proposition 3.1 (Asymptotic Normality and Worst-Case Bias of Stratified RD)**.** *Assume Assumptions 1 to 5. Then*

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

where $LB := g\,\mathbb{E}[\hat{\tau}_1(h) - \tau_1 \,|\, \mathbf{X}, \mathbf{Z}] + (1-g)\,\mathbb{E}[\hat{\tau}_0(h) - \tau_0 \,|\, \mathbf{X}, \mathbf{Z}] + [\tau_1 - \tau_0]\,\mathbb{E}[\hat{g}(h_z) - g \,|\, \mathbf{X}]$ *denotes the leading bias. The worst-case leading bias over the primitives is*

$$\overline{\text{Bias}}(\hat{\tau}_S(h,h_z)) := \sup_{(m,\tau,g)\in\mathcal{F}_p^R(M,\mathbf{C},\mathbf{L})} |LB| = h^p(b_S + o_p(1)), \tag{21}$$

with the first-order term

$$b_S := \frac{1}{p!}\Big[2M\,B_{p,q}(k) + \kappa^p\,C_{\Delta\tau}L\,B^{\text{int}}_{p,q}(k)\Big], \tag{22}$$

where $B^{\text{int}}_{p,q}(k) := p\int_0^\infty \Big|\int_{|u|\geq t} k_q^{\text{int}}(u)\,(|u|-t)^{p-1}\,du\Big|\,dt$ *is the interior-point worst-case bias kernel constant as in Armstrong and Kolesár (2020, Thm. 3.1), with the interior equivalent kernel* $k_q^{\text{int}}(u) := e_1^\top\big(\int_{-\infty}^\infty m_q(t)\,m_q(t)^\top k(t)\,dt\big)^{-1} m_q(u)\,k(u)$.

The asymptotic variance can be written as

$$\mathcal{V}_S = \mathcal{V}_O + [\tau_1 - \tau_0]^2\frac{1}{\kappa}\mathcal{V}_g,$$
$$\text{with}\quad \mathcal{V}_O = \frac{\overline{k_q^2}(k)\left(\sigma^2_{L,+}(0) + \sigma^2_{L,-}(0)\right)}{f_X(0)} \quad \text{and} \quad \mathcal{V}_g = \frac{\widetilde{k_q^2}(k)\,g(1-g)}{f_X(0)},$$

where $\sigma^2_{L,\pm}(0)$ *denote the one-sided limits of* $\sigma^2_L(x)$ *at the cutoff,* $\overline{k_q^2}(k) := \int_0^\infty k_q^*(u)^2\,du$ *is the boundary variance kernel constant, and* $\widetilde{k_q^2}(k) := \int_{-\infty}^\infty k_q^{\text{int}}(u)^2\,du$ *the interior-point variance kernel constant.*

The terms of the leading bias LB condition on the sample realizations of different random variables. The Stratified RD estimator runs the base RD estimator within each group and its weights depend on the running variable and the group indicator. Conditional on $(\mathbf{X}, \mathbf{Z})$ those weights are fixed and the results in Armstrong and Kolesár (2020) apply. Further, with the binary covariate there is a nonzero, but asymptotically vanishing, probability that all units lie in the same group. Conditioning on the realizations of both the running variable and the binary covariate $(\mathbf{X}, \mathbf{Z})$ ensures that the conditional expectation is nonetheless well defined. When considering the estimated group shares, I only condition on the realizations of the running variable $\mathbf{X}$ to accurately capture the effect of the sampling variability of Z_i on the Stratified RD estimator.[18] Asymptotically the realizations in the conditioning set do not matter because the worst-case leading bias has a deterministic first-order term.

The leading bias LB depends on the unknown regression functions, but taking the supremum over the class of functions admissible under Assumption 3 gives its first-order worst-case bias. The

[18] Conditional on $(\mathbf{X}, \mathbf{Z})$ the estimator of the group shares is not random.

worst-case bias inherits the additive structure characterized in the linearization and the first two terms collapse to a single expression, because the same bandwidth h is used for estimation in both groups and Assumption 3 imposes that the long regression functions in both groups share the same smoothness constant. The first term in equation (22) therefore reflects the worst-case bias from estimating the group-specific treatment effects and depends on the smoothness constant M of the long regression functions. The second term is the additional weight-estimation bias originating from estimating the group shares. It depends on the product of the assumed bound on treatment effect heterogeneity and the smoothness constant of the covariate regression function $C_{\Delta\tau}L$, and it grows with the bandwidth ratio κ. Notably, stratification eliminates all composition bias and therefore the worst-case bias does not contain bounds on the baseline heterogeneity.[19]

The asymptotic variance $\mathcal{V}_S$ contains the first-order terms from the linearization: $\mathcal{V}_O$ reflects the variance of estimating the group-specific treatment effects at known population weights and $[\tau_1 - \tau_0]^2\kappa^{-1}\mathcal{V}_g$ captures the additional variance from estimating the weights. It scales with the squared treatment effect heterogeneity and is zero under treatment effect homogeneity. The bandwidth ratio $\kappa = h_z/h$ governs the precision of weight estimation relative to that of the treatment effect estimation. A larger κ widens the window used for weight estimation, reducing $\kappa^{-1}\mathcal{V}_g$ relative to $\mathcal{V}_O$ at the cost of increased worst-case weight-estimation bias.

4 Estimator Comparison

Having established the asymptotic properties of the Stratified RD estimator and the base RD estimator, I compare their first-order worst-case bias over the class $\mathcal{F}_p^R(M, \mathbf{C}, \mathbf{L})$ and their asymptotic variance. Both estimators use the same bandwidth h to estimate the treatment effects to isolate the effect of stratification in the comparison. The Stratified RD estimator additionally uses the second bandwidth $h_z = \kappa h$ (Assumption 4) to estimate the group shares at the cutoff.

To show the reduction in first-order worst-case bias, I compare the first-order terms in Propositions 2.1 and 3.1 that approximate the worst-case bias of the Stratified RD estimator and bound that of the base RD estimator from below in sufficiently large samples. In general, the worst-case biases of the two estimators satisfy

$$\overline{\text{Bias}}(\hat{\tau}(h)) - \overline{\text{Bias}}(\hat{\tau}_S(h, h_z)) \geq h^p(b_B^{\text{low}} - b_S + o_p(1)). \tag{23}$$

Showing that $b_B^{\text{low}} - b_S > 0$ establishes that the Stratified RD estimator has smaller worst-case bias in sufficiently large samples. Plugging in the first-order terms yields

$$b_B^{\text{low}} - b_S = \frac{L}{p!}\Big[2C_{\Delta m}\, B_{p,q}(k) + C_{\Delta\tau}(B_{p,q}(k) - \kappa^p B_{p,q}^{\text{int}}(k))\Big]. \tag{24}$$

[19] Since stratification eliminates composition bias in the case of discrete covariates, conditions (iv) and (v) of Assumption 3 are technically not needed in Proposition 3.1. The supremum over the larger class of functions satisfying only conditions (i) to (iii) would yield the same first-order worst-case bias.

The first term is the composition bias of the base RD estimator, which stratification removes. The second term compares the weight-estimation biases of the two estimators. The dominance result in worst-case bias across all possible constants requires a restriction on κ that depends only on the kernel constants, $\kappa \leq U_{p,q}(k) := \left(B_{p,q}(k)/B^{\text{int}}_{p,q}(k)\right)^{1/p}$. Intuitively, this condition ensures that the bandwidth h_z is small enough relative to h, so that the worst-case weight-estimation bias of the Stratified RD estimator is guaranteed to be weakly smaller than that of the base RD estimator.

Considering variance, the results in Armstrong and Kolesár (2020) show that the base RD estimator has asymptotic variance $\mathcal{V}_B$ that satisfies $\text{Var}[\hat{\tau}(h) \,|\, \mathbf{X}] = \mathcal{V}_B/(nh)\,(1 + o_p(1))$ with

$$\mathcal{V}_B = \frac{\overline{k_q^2}(k)\,(\sigma^2_{P,+}(0) + \sigma^2_{P,-}(0))}{f_X(0)}, \tag{25}$$

where $\sigma^2_{P,\pm}(0)$ denote the one-sided limits of the pooled error variance at the cutoff. The variance of the pooled error u_i relates to that of the long error e_i through the law of total variance (Lemma C.1). By the continuity of the regression functions, the one-sided limits of the pooled error variance can be rewritten as

$$\sigma^2_{P,-}(0) = \sigma^2_{L,-}(0) + g(1-g)\,\Delta m(0)^2, \qquad \sigma^2_{P,+}(0) = \sigma^2_{L,+}(0) + g(1-g)\,(\Delta m(0) + \Delta\tau(0))^2. \tag{26}$$

The difference between $\mathcal{V}_B$ and $\mathcal{V}_S$ then is

$$\mathcal{V}_B - \mathcal{V}_S = \frac{g(1-g)}{f_X(0)}\left[2\,\overline{k_q^2}(k)\left(\Delta m(0) + \frac{\Delta\tau(0)}{2}\right)^2 + \left(\frac{\overline{k_q^2}(k)}{2} - \frac{\widetilde{k_q^2}(k)}{\kappa}\right)\Delta\tau(0)^2\right]. \tag{27}$$

The first term is nonnegative and the sign of the second term depends on the bandwidth ratio κ. It is nonnegative for all values of $\Delta\tau(0)$ if and only if $\kappa \geq L_{p,q}(k) := 2\,\widetilde{k_q^2}(k)/\overline{k_q^2}(k)$, which requires h_z to be large enough relative to h.

Any $\kappa \in I_{p,q}(k) := [L_{p,q}(k), U_{p,q}(k)]$ satisfies both conditions on κ and therefore guarantees that the Stratified RD estimator has simultaneously weakly lower asymptotic variance and first-order worst-case bias. For such a choice of κ the asymptotic worst-case MSE of the Stratified RD estimator is weakly lower than that of the base RD estimator. I define the asymptotic worst-case MSE of the Stratified RD estimator and the corresponding lower bound for the base RD estimator as

$$\overline{\text{AMSE}}(\hat{\tau}_S(h, h_z)) := \left(h^p\, b_S\right)^2 + \frac{\mathcal{V}_S}{nh} \qquad \text{and} \qquad \overline{\text{AMSE}}^{\text{low}}(\hat{\tau}(h)) := \left(h^p\, b_B^{\text{low}}\right)^2 + \frac{\mathcal{V}_B}{nh}, \tag{28}$$

where the asymptotic variances are evaluated at the true regression functions.

Theorem 4.1 (Comparison of the Stratified RD and Base RD Estimators)**.** *Assume Assumptions 1 to 5.*

(i) If $\kappa \leq U_{p,q}(k)$, then $b_S \leq b_B^{\text{low}}$, with strict inequality if $C_{\Delta m} L > 0$.

(ii) If $\kappa \geq L_{p,q}(k)$, then $\mathcal{V}_S \leq \mathcal{V}_B$, with strict inequality if $\Delta m(0) + \Delta\tau(0)/2 \neq 0$.

(iii) If $\kappa \in I_{p,q}(k)$*, then*

$$\overline{\text{AMSE}}(\hat{\tau}_S(h, h_z)) \leq \overline{\text{AMSE}}^{\text{low}}(\hat{\tau}(h)), \tag{29}$$

with strict inequality if $C_{\Delta m} L > 0$ *or* $\Delta m(0) + \Delta\tau(0)/2 \neq 0$*.*

Theorem 4.1 establishes general kernel-specific intervals for the bandwidth ratio. Choosing κ within those intervals is then sufficient to guarantee that the weight-estimation bandwidth h_z is neither too large nor too small relative to h, and that the Stratified RD estimator yields simultaneously weakly lower worst-case bias and weakly lower variance than the base RD estimator.

The question remains whether this interval is nonempty and hence gives a constructive way to choose the bandwidth ratio. Table 1 tabulates the kernel constants, conditions for κ and the interval $I_{p,q}(k)$ rounded to two decimal places for three common kernels and smoothness orders $p \in \{1, 2, 3\}$. For all three kernels and smoothness orders the interval is nonempty. For $p = 1$ the interval is a singleton for all three kernels, but the width of the interval increases for smoothness order $p = 2$ and $p = 3$. Given a bandwidth h that a researcher would have chosen to estimate the base RD estimator, she can simply set $h_z = \kappa h$ with κ within the intervals provided in Table 1 to implement the Stratified RD estimator. Any such choice satisfies the assumptions of Theorem 4.1 and ensures that its results hold.

Table 1: Kernel Constants and Interval Endpoints, $q = p - 1$

Kernel	p	$\overline{k_q^2}$	$\widetilde{k_q^2}$	$B_{p,q}$	$B_{p,q}^{\text{int}}$	$L_{p,q}$	$U_{p,q}$	$I_{p,q}$
Uniform	1	1	1/2	1/2	1/2	1	1	$\{1\}$
	2	4	1/2	1/6	1/3	1/4	$\sqrt{1/2}$	$[0.25, 0.71]$
	3	9	9/8	1/20	1/16	1/4	$(4/5)^{1/3}$	$[0.25, 0.93]$
Triangular	1	4/3	2/3	1/3	1/3	1	1	$\{1\}$
	2	24/5	2/3	1/10	1/6	5/18	$\sqrt{3/5}$	$[0.28, 0.77]$
	3	72/7	456/343	1/35	8/245	38/147	$(7/8)^{1/3}$	$[0.26, 0.96]$
Epanechnikov	1	6/5	3/5	3/8	3/8	1	1	$\{1\}$
	2	4.4980	3/5	11/95	1/5	0.2668	$\sqrt{11/19}$	$[0.27, 0.76]$
	3	9.8165	5/4	15/448	5/128	0.2547	$(6/7)^{1/3}$	$[0.25, 0.95]$

Since Theorem 4.1 holds for any bandwidth sequence h that satisfies Assumption 1, it also holds for the MSE-optimal bandwidths that standard algorithms estimate (Imbens and Kalyanaraman 2012; Armstrong and Kolesár 2020). Rather than setting the first-step bandwidth in the Stratified RD estimator to h, one could optimize over the bandwidths h_j for estimating the treatment effects within each group and the bandwidth for estimating the group shares h_z to minimize the asymptotic worst-case MSE of the Stratified RD estimator. The results in Theorem 4.1 imply that the Stratified RD estimator also has asymptotically weakly lower first-order worst-case MSE if both estimators are evaluated at their respective worst-case asymptotic MSE-optimal bandwidth choices.

In empirical practice, researchers should construct confidence intervals that account for the bias of the estimator to prevent large coverage distortions due to the first-order bias of RD estimators.

One approach is to construct bias-aware confidence intervals whose width is inflated according to the ratio of the worst-case bias to the standard error of the estimator (Armstrong and Kolesár 2020). In large samples this ratio can be approximated for an estimator with first-order worst-case bias b and asymptotic variance $\mathcal{V}$ by the first-order bias-sd ratio, $t(b, \mathcal{V}) := \frac{h^p b}{\sqrt{\mathcal{V}/(nh)}}$. Choosing the critical value when constructing the confidence interval as the $1-\alpha$ quantile of $|\mathcal{N}(t, 1)|$ accounts for the first-order worst-case bias (Armstrong and Kolesár 2020). A corollary of parts (i) and (ii) in Theorem 4.1 is that bias-aware confidence intervals of the Stratified RD estimator will be asymptotically weakly shorter than those of the base RD estimator.

Corollary 4.2 (Asymptotic CI Half-Length Comparison)**.** *Assume Assumptions 1 to 5. If $\kappa \in I_{p,q}(k)$, then the asymptotic CI half-length $\ell_a(b, \mathcal{V}) := \mathrm{cv}_{1-\alpha}(t(b, \mathcal{V})) \cdot \sqrt{\mathcal{V}/(nh)}$ satisfies*

$$\ell_a(b_S, \mathcal{V}_S) \leq \ell_a(b_B^{\mathrm{low}}, \mathcal{V}_B).$$

Remark 1 (Feasible CI). Under any variance estimator $\hat{V}$ that is consistent for $\mathcal{V}$, for example the HC2 or nearest-neighbor estimators discussed in Kolesár and Rothe (2018), the feasible half-length $\ell(b, \hat{V}) := \mathrm{cv}_{1-\alpha}(t(b, \hat{V})) \cdot \sqrt{\hat{V}/(nh)}$ satisfies $\ell(b, \hat{V})/\ell_a(b, \mathcal{V}) \xrightarrow{p} 1$. The ordering in Corollary 4.2 therefore guarantees that in sufficiently large samples the Stratified RD estimator has shorter CIs than the base RD estimator, that is $\ell(b_S, \hat{V}_S)/\ell(b_B^{\mathrm{low}}, \hat{V}_B) \xrightarrow{p} \ell_a(b_S, \mathcal{V}_S)/\ell_a(b_B^{\mathrm{low}}, \mathcal{V}_B) < 1$ if the improvement in worst-case bias is strict.

The intuition behind Corollary 4.2 is that the bias-aware confidence interval of the base RD estimator must be inflated to account for composition bias, which enters its worst-case bias through the product $2C_{\Delta m}L$, whereas the Stratified RD estimator eliminates composition bias. For $\kappa \in I_{p,q}(k)$, the confidence intervals of the Stratified RD estimator are therefore strictly shorter in sufficiently large samples whenever the researcher cannot rule out composition bias a priori, that is, whenever $C_{\Delta m}L > 0$.

5 Empirical Application

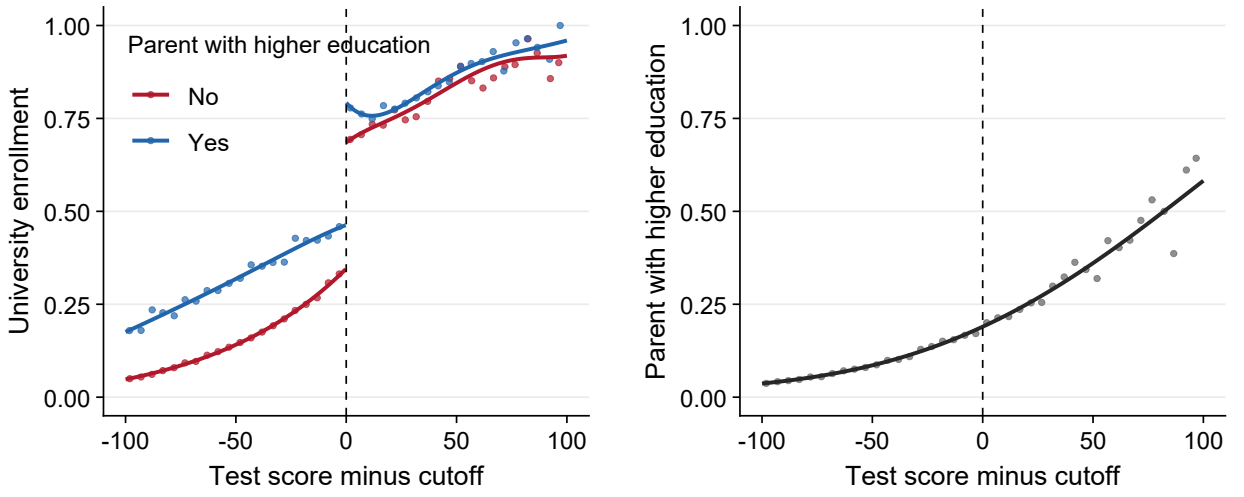
I illustrate the potential for bias reduction through covariate adjustment using the replication dataset from Londoño-Vélez et al. (2020). The paper studies the effect of a merit-based financial aid program on university enrollment after high school graduation. The program covers university tuition for students who score in the top decile in their high school exit exam (SABER 11). The data include information on whether the parents of a student have a higher education degree. It seems intuitive that the educational background of a student's parents affects her performance in school and her university enrollment decisions.

I therefore consider the indicator of whether either of the parents of a student has a higher education degree as a binary covariate to adjust for in two empirical exercises. First, to illustrate the effect of covariate adjustment on bias I run a calibrated simulation study with regression functions and covariate distributions estimated in the large administrative dataset used by Londoño-Vélez et al.

(2020). Second, I follow the approach of Noack et al. (2026), who use a placebo cutoff in the original sample of the paper and perform a subsampling Monte Carlo study to evaluate covariate-adjusted and unadjusted estimators relative to the true jump of zero at the placebo cutoff. In my exercise I instead use a placebo sample and keep the cutoff as in the original paper.

5.1 Calibrated Simulation

Demonstrating the effect of covariate adjustment on bias is difficult, because in any study the true parameter and underlying regression functions are unobserved, so that the bias cannot be computed. Following Imbens and Kalyanaraman (2012), I therefore conduct a calibrated simulation in which I first fit quintic polynomials to the large administrative dataset used by Londoño-Vélez et al. (2020) to estimate the underlying regression functions and then treat those estimates as the true model. For the simulation, I draw the running variable from a normal distribution chosen to match the shares of students within ten test score points below and above the cutoff in the empirical distribution of the running variable (see Figure 6). With the data-generating process fully specified, I then evaluate the bias, standard deviation, and length of bias-aware confidence intervals of the base RD and Stratified RD estimators at growing sample sizes.



(a) University enrollment rate by group

(b) Share of students who have a parent with a higher education degree

Figure 2: University enrollment and group composition of eligible students who took the SABER 11 test in fall 2014 for whom information on parental education is available ($N = 274{,}260$). The group composition depicts the share of students with a parent who has a higher education degree. Dots are means over bins of five score points. In panel (a) the lines are quintic polynomials fitted separately on each side of the cutoff for each group. In panel (b) the line is a quintic polynomial fitted across the cutoff.

Figure 2 shows the quintic polynomial estimates of the long and covariate regression functions, which I treat as the true model in the subsequent simulation study. Panel (a) shows the long regression functions, which in this setting describe the probability that a student enrolls in university

after high school graduation given her test score and her parents' educational background. It clearly shows that students with higher test scores are more likely to enroll in a university, but also that the parental background markedly increases the probability of a student enrolling in university. Panel (b) depicts the share of students with at least one parent holding a higher education degree. The share monotonically increases with the test score, indicating that students with at least one parent with a university degree score higher on average.

This change in student composition with the test score and the heterogeneity between the two groups depicted in Figure 2 induce composition bias. Figure 3 shows the resulting specification bias, also considered in Kolesár and Rothe (2018, p. 2287), which measures the distance of the population projection of the estimators to the target parameter when the bandwidth is held fixed.[20] Panel (a) shows that the bias of the base RD, linear adjustment RD and Stratified RD estimators converges to zero as the bandwidth shrinks, as predicted by the theory. The base RD and linear adjustment RD estimators are close to each other in terms of bias, while the Stratified RD estimator has the lowest specification bias due to removing the composition bias (grey shaded area). To better depict the first-order bias as the bandwidth gets close to zero I rescale the bias by $1/h^2$. Panel (b) clearly shows that composition bias makes up a substantial share of the bias of the base RD estimator even for small bandwidths. This indicates that composition bias is first order and makes up 26% of the leading constant of the bias of the base RD estimator.

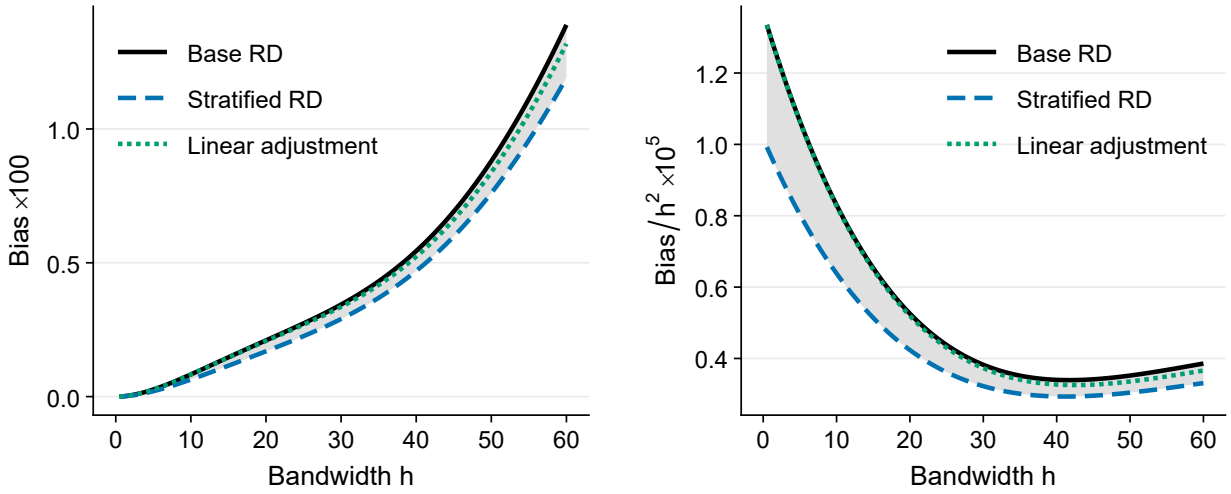
By removing composition bias, the Stratified RD estimator has the lowest bias among the three estimators, which leads to improved inference. Choosing the smoothness constants based on the calibrated model, I construct bias-aware confidence intervals for the target parameter based on the base RD and Stratified RD estimates. Table 2 shows that the Stratified RD estimator reduces the finite-sample bias of the base RD estimator by 14% to 23% while having almost identical standard deviation. The bias-aware confidence intervals of the Stratified RD estimator are 6% to 31% shorter than those of the base RD estimator, depending on the method used to choose the bandwidth. Panel A of the table rescales the bandwidth according to the sample size in the simulation draw relative to a bandwidth chosen with rdrobust in the entire sample, as done in the original paper.[21] Panel B uses the RMSE-optimal bandwidth in Armstrong and Kolesár (2020, eq. 19) with the known theoretical quantities from the calibrated simulation. Panel C uses the same rule but plugs in a smoothness constant chosen using the rule of thumb suggested by Armstrong and Kolesár (2020). According to this rule of thumb one fits a quartic polynomial on each side of the cutoff in the entire sample and reads off the largest second derivative within the support of the data. Since the bias-aware confidence intervals account for the worst-case bias, nominal coverage is achieved in all panels with at least 95.9% coverage (not reported). The reduction in confidence interval length

[20] Formally the specification bias at a fixed bandwidth h is $\operatorname{plim}_{n\to\infty} \hat{\tau}(h) - \tau$ for the base RD estimator and $\operatorname{plim}_{n\to\infty} \hat{\tau}_S(h, h_z) - \tau$ for the Stratified RD estimator. Conditional on the event where the estimators are well defined, they are in expectation close to their population projection and converge to them with increasing sample size. Figure 7 shows that the Monte Carlo finite-sample bias measured in the calibrated simulation is close to the population-based specification bias.

[21] This leads to the bandwidth in the simulation $h_s = h_p(n_p/n_s)^{(1/5)}$, where h_p is the bandwidth used in the paper, n_p the sample size in the paper and n_s the sample size of the simulation draw.

varies across the panels, because depending on the chosen bandwidth the worst-case bias to standard error ratio of the estimators changes. The wide bandwidths used in Panel A lead to proportionally more worst-case bias and stratification thus leads to a bigger reduction in interval length.

Appendix B.I extends the calibrated simulation by stratifying on both whether a student has a parent with higher education and whether she lives in a metropolitan area. The bias reduction in Table 4 when stratifying on both binary variables is 25–30% relative to the base RD estimator, depending on the bandwidth, while both estimators have similar variance.



(a) Specification bias ×100 of the three estimators

(b) Specification bias $\times 10^5$ of the three estimators scaled by $1/h^2$

Figure 3: Specification bias at fixed bandwidth of the base RD, the linear adjustment RD, and the Stratified RD estimators. All three estimators use the same bandwidth h and the Stratified RD estimator additionally uses $h_z = \kappa h$ with $\kappa = \sqrt{1/2}$ within the interval for the uniform kernel from Table 1 to estimate the group shares. The shaded area is the difference between the biases of base RD and Stratified RD, the composition bias that stratification removes. Panel (a) shows the bias of the three estimators in levels, which converges to zero at rate h^2 as predicted by the theory. Panel (b) shows the bias of the three estimators scaled by $1/h^2$.

Table 2: Calibrated Simulation

Bandwidth selection		Finite-sample bias ×100			SD ×100			CI length
n	h	Base	Strat.	Ratio	Base	Strat.	Ratio	Ratio
Panel A. CCT rdrobust								
10,000	58.9	1.330	1.135	0.85	3.63	3.63	1.00	0.69
80,000	38.9	0.516	0.445	0.86	1.59	1.60	1.00	0.69
320,000	29.5	0.336	0.283	0.84	0.91	0.91	1.00	0.69
Panel B. AK optimal bandwidth								
10,000	18.6	0.193	0.152	0.79	6.62	6.59	1.00	0.94
80,000	12.3	0.112	0.087	0.78	2.80	2.81	1.00	0.94
320,000	9.3	0.074	0.057	0.77	1.60	1.60	1.00	0.94
Panel C. AK rule of thumb								
10,000	30.2	0.349	0.290	0.83	5.04	5.03	1.00	0.82
80,000	19.9	0.209	0.169	0.81	2.20	2.19	1.00	0.81
320,000	15.1	0.148	0.117	0.79	1.27	1.27	1.00	0.81

Notes: Samples of size n drawn from the calibrated data-generating process of Figure 2, with 1,000 samples for each n. Base RD and Stratified RD use the same bandwidth h, and Stratified RD estimates the group share with $h_z = \kappa h$. The panels show results for different bandwidth selectors. CCT rdrobust extrapolates the MSE-optimal bandwidth of rdrobust (Calonico et al. 2014) on the full sample ($h = 29.9$ for $N = 299{,}475$) with the rate $n^{-1/5}$. AK optimal bandwidth is the bandwidth of Armstrong and Kolesár (2020) for base RD with the curvature bound of the calibration, and AK rule of thumb uses their rule-of-thumb bound instead. The finite-sample bias column reports the mean over samples of the finite-sample bias conditional on the realizations of the running variable and the SD column is the Monte Carlo standard deviation of the estimates, both in percentage points. Ratio is Stratified RD divided by base RD. CI length is the average length of the bias-aware 95% confidence interval with HC1 standard errors, reported as this ratio.

5.2 Placebo Monte Carlo Study

As a second exercise I run a placebo subsampling study following the approach in Noack et al. (2026). The advantage of such a placebo study is that one does not need to specify the underlying data-generating process, but can simply evaluate the estimators against the true effect of zero in the placebo sample.

For this exercise, I use the sample of students who were ineligible for the financial aid program as the placebo group. To be eligible, students had to clear a second threshold that excluded students from wealthier households from the financial aid program. This rule based on the SISBEN score of a household splits the sample into students who were in principle eligible for treatment and a clear placebo group that was not. Exploiting the fact that in the placebo group the cutoff for merit-based financial aid did not affect the probability of enrolling in a college (Figure 4), I evaluate the deviation of the two estimators relative to the true effect of zero in this placebo group of ineligible students.

In the placebo study, I use three different methods to choose the bandwidths: the bandwidth selector of the rdrobust package (Calonico et al. 2014), and the MSE-optimal bandwidth of Armstrong and Kolesár (2020) under the rule-of-thumb smoothness bounds suggested in Armstrong and Kolesár (2020) and Imbens and Wager (2019). I run each of these methods to select a bandwidth for the

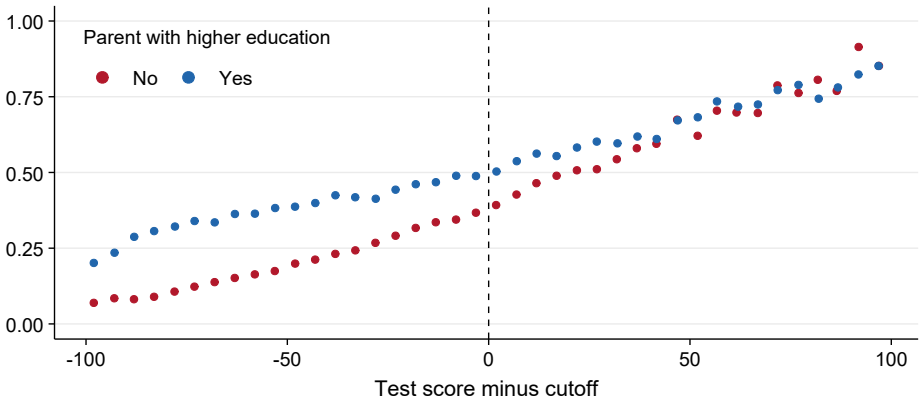


Figure 4: University enrollment of the SISBEN-ineligible students who took the SABER 11 test in fall 2014 ($N = 248{,}415$). Dots are means over bins of five score points.

entire placebo population from which I then subsample 5,000, 10,000 or 20,000 students. For each of the subsamples I then rescale the bandwidth according to the MSE-optimal rate and then compute the base RD and Stratified RD estimators at the shared bandwidth h. The Stratified RD estimator additionally uses the bandwidth $h_z = \sqrt{1/2}h$ for estimating the group weights. For both estimators I also construct conventional confidence intervals that do not account for the bias.

The results in Table 3 show that the Stratified RD estimator has lower bias than the base RD estimator across all sample sizes and bandwidth selection rules. The bias reduction is between 13% and 29%, while the standard deviation is similar for both estimators. At the largest bandwidth in the rdrobust panel with a sample size of 5,000 the standard deviation of the Stratified RD estimator is slightly larger than that of the base RD estimator. I also report the bias to standard deviation ratio and the rejection rate of conventional confidence intervals that are distorted here due to the first-order bias (Armstrong and Kolesár 2020).

Panel A uses the largest bandwidths and hence has the largest bias to standard deviation ratios that lead to rejection rates of up to 7.6%, above the nominal level of 5%. The bias reduction through stratification alleviates this problem and consequently the Stratified RD estimator has a rejection rate of 6.6%. The rejection rate is still above the nominal level, since, while it reduces its bias relative to the base RD estimator, the Stratified RD estimator still carries its own first-order bias. Panel B uses much smaller bandwidths, so that the bias to standard deviation ratio is much smaller and the conventional confidence intervals reject slightly below the nominal level. Panel C sits in between the results in Panels A and B.

The results demonstrate that at bandwidth selection rules that trade off bias and variance the size distortions due to bias are real but modest. The theoretical MSE-optimal bandwidth results in a bias to standard deviation ratio of 0.5, which leads to a rejection rate of 7.9% for a conventional test at 5% nominal level (Armstrong and Kolesár 2020, p. 27). Stratification reduces bias while having a similar standard deviation, and thus reduces the bias to standard deviation ratio and size distortions when using conventional inference. Since both estimators carry first-order bias,

researchers should use bias-aware inference even after adjusting for covariates to reduce bias.

Table 3: Placebo Monte Carlo Study

Bandwidth selection		Bias ×100			SD ×100			Bias-SD ratio		Rejection (%)	
n	h	Base	Strat.	Ratio	Base	Strat.	Ratio	Base	Strat.	Base	Strat.
Panel A. CCT rdrobust											
5,000	67.3	1.54	1.30	0.84	3.87	3.98	1.03	0.40	0.33	6.8	6.1
10,000	58.6	1.00	0.84	0.84	2.87	2.91	1.01	0.35	0.29	5.7	5.4
20,000	51.0	1.26	1.07	0.85	2.11	2.14	1.01	0.60	0.50	7.6	6.6
Panel B. AK rule of thumb											
5,000	35.5	0.73	0.58	0.79	5.13	5.16	1.01	0.14	0.11	4.6	4.8
10,000	30.9	0.47	0.33	0.71	3.87	3.85	1.00	0.12	0.09	4.6	4.4
20,000	26.9	0.61	0.44	0.72	2.91	2.91	1.00	0.21	0.15	4.5	4.3
Panel C. IW rule of thumb											
5,000	48.6	1.07	0.91	0.85	4.44	4.50	1.01	0.24	0.20	5.6	5.6
10,000	42.3	1.30	1.14	0.87	3.32	3.33	1.00	0.39	0.34	6.0	5.6
20,000	36.8	0.90	0.78	0.87	2.49	2.50	1.01	0.36	0.31	5.6	5.3

Notes: Subsamples of size n drawn without replacement from the SISBEN-ineligible students who took the SABER 11 test in fall 2014 ($N = 248{,}415$), with 10,000 subsamples for each n. The panels differ in the bandwidth selector, each applied once to the full population to obtain h_N and scaled to the subsample size using $h = h_N(N/n)^{1/5}$. CCT rdrobust is the MSE-optimal bandwidth of rdrobust (Calonico et al. 2014) with a uniform kernel. The AK rule of thumb minimizes the finite-sample RMSE of Armstrong and Kolesár (2020, eq. 21) with their rule-of-thumb bound M, the largest second derivative of a global quartic fitted on each side of the cutoff. The IW rule of thumb minimizes the same criterion with the bound of Imbens and Wager (2019), which is twice the largest second derivative of a global quadratic fitted on each side of the cutoff. All estimators use the same bandwidth h, and Stratified RD estimates the group share with $h_z = \kappa h$. Bias is the Monte Carlo mean of the estimates, and SD their standard deviation. The Bias-SD ratio is the bias of each estimator divided by its SD. The rejection rate is computed for a conventional 5% t-test against the hypothesis of an effect of 0 with HC1 standard errors.

6 Extensions to Discrete and Continuous Covariates

This section extends the results of Sections 2 to 4 to a discrete covariate, and to a continuous covariate under Lipschitz conditions on the primitives ($p = 1$).

6.1 Generalization to a Discrete Covariate

Let the covariate be discrete, so that it takes $J + 1$ values, $Z_i \in \{0, 1, \ldots, J\}$, with group shares $g_j(x) := \Pr(Z_i = j \,|\, X_i = x)$. To analyze the worst-case bias one has to adapt the smoothness assumptions to several groups by imposing the conditions of Assumption 3 on the regression functions and the share of every group, and the heterogeneity bounds on every possible pair of groups. As before, the Stratified RD estimator estimates the treatment effect within each group and the share of each group at the cutoff, and averages them, $\hat{\tau}_S(h, h_z) := \sum_{j=0}^{J} \hat{\tau}_j(h)\, \hat{g}_j(h_z)$.

One can show that the worst-case biases then have the same structure in this setting as in

Propositions 2.1 and 3.1:

$$\overline{\text{Bias}}(\hat{\tau}(h)) \geq \frac{h^p}{p!}\Big[2M + (2C_{\Delta m} + C_{\Delta\tau})\, K_J L\Big] B_{p,q}(k) + o_p(h^p),$$
$$\overline{\text{Bias}}(\hat{\tau}_S(h, h_z)) = \frac{h^p}{p!}\Big[2M\, B_{p,q}(k) + \kappa^p\, C_{\Delta\tau} K_J L\, B^{\text{int}}_{p,q}(k)\Big] + o_p(h^p),$$

where $K_J := \lfloor (J+1)/2 \rfloor$. The first-order terms are those in equations (14) and (22) with $K_J L$ in place of L.

The structure is the same because the shares of the $J+1$ groups sum to one. The worst case then maximizes the change in group composition by reducing the share of some groups and increasing that of others as much as the smoothness constant L allows as the running variable increases. Further, the worst case aligns the heterogeneity in the regression functions with the increase or decrease of their shares. It pushes the regression functions of the different groups as far apart as the heterogeneity bounds allow. The result is that all groups belong either to a high or to a low meta-group. For these two meta-groups the problem has the structure of the two-group case. The only difference lies in the smoothness of the shares of the meta-groups. Since all shares sum to one, the increase in the share of the high meta-group must be offset by the decrease in the share of the low meta-group, whose groups can each fall at most at the rate L. At most half of the groups can therefore increase their shares at the rate L, so the share of the high meta-group lies in $\mathcal{F}_p(K_J L)$, and this bound is attained when half of the group shares are shifted in the same direction at the bound L.

Worst-case composition bias of the base RD estimator thus grows linearly in the number of groups, as the composition of groups can change more rapidly, while the Stratified RD estimator still eliminates it for any J. For the Stratified RD estimator the factor K_J enters the weight-estimation bias, because more groups allow the composition to change faster. Since $K_J L$ scales both terms in the difference of the first-order worst-case biases in equation (24), the upper threshold $U_{p,q}(k)$ does not depend on J. Further, the asymptotic variances do not involve L, and in the difference of the variances in equation (27) the terms $g(1-g)(\Delta m(0) + \Delta\tau(0)/2)^2$ and $g(1-g)\,\Delta\tau(0)^2$ become $\text{Var}[m(0, Z_i) + \tau(0, Z_i)/2 \,|\, X_i = 0]$ and $\text{Var}[\tau(0, Z_i) \,|\, X_i = 0]$, so the lower threshold $L_{p,q}(k)$ remains unchanged. Consequently, Propositions 2.1 and 3.1, Theorem 4.1, and Corollary 4.2 hold with the constant L replaced by $K_J L$ and with a generalized variance expression, for a discrete covariate with any number of support points, with the interval for the bandwidth ratio $I_{p,q}(k)$ unchanged.

6.2 Generalization to a Continuous Covariate

This section extends the worst-case bias analysis to a continuous covariate. The extension is developed for Lipschitz smoothness assumptions and local constant estimation, so that $p = 1$ and $q = 0$. Let Z_i now be a continuous random variable taking values in the bounded domain $\mathcal{Z} := [-\zeta, \zeta]$ and let the running variable X_i also have a bounded domain $\mathcal{X}$. Local constant estimation with a symmetric and nonnegative kernel implies that the weights of the base RD estimator are nonnegative

on each side of the cutoff. Together with the bandwidth restriction and a kernel supported on $[-1,1]$, all units with positive weight lie in the compact set $[-h,h]\times\mathcal{Z}\subset\mathbb{R}^2$. As before, the target parameter is the conditional average treatment effect at the cutoff, which for a continuous covariate is a weighted integral of the treatment effect function over $\mathcal{Z}$

$$\tau := \bar{\tau}(0) = \int_{\mathcal{Z}} \tau(0,z)\, f_{Z|X}(z\,|\,0)\, dz.$$

For a continuous covariate I impose smoothness assumptions on the long regression functions and the conditional density.

Assumption 6 (Outcome Surface Smoothness)**.** For constants $M, C_m, C_\tau \geq 0$ the following holds:

(i) $m(\cdot,z),\ m(\cdot,z)+\tau(\cdot,z)\in\mathcal{F}_1(M)$ for all $z\in\mathcal{Z}$.

(ii) $m(x,\cdot)\in\mathcal{F}_1(C_m)$ for all $x\in\mathcal{X}$.

(iii) $\tau(x,\cdot)\in\mathcal{F}_1(C_\tau)$ for all $x\in\mathcal{X}$.

Condition (i) is the counterpart of m, $m+\tau\in\mathcal{F}_p(M)$ and bounds the bias in approximating the long regression functions. Conditions (ii) and (iii) bound heterogeneity along the covariate direction and are analogous to $C_{\Delta m}$ and $C_{\Delta\tau}$ in Assumption 3. However, here they are measured per unit of covariate distance rather than as level gaps between two groups. The long regression functions (m,τ) satisfying Assumption 6 are collected in the class $\mathcal{F}_1^C(M,C_m,C_\tau)$.

Assumption 7 (Conditional Density Smoothness)**.** For constants $L_x, L_z \geq 0$, with $L_x = 0$ whenever $L_z = 0$, and $\underline{f} > 0$ the following holds:

(i) $f_{Z|X}(z\,|\,\cdot)\in\mathcal{F}_1(L_x)$ for all $z\in\mathcal{Z}$.

(ii) $f_{Z|X}(\cdot\,|\,x)\in\mathcal{F}_1(L_z)$ for all $x\in\mathcal{X}$.

(iii) $f_{Z|X}(z\,|\,x)\geq\underline{f}$ for all $(x,z)\in\mathcal{X}\times\mathcal{Z}$.

This assumption is the counterpart of $g(x)\in\mathcal{F}_p(L)$. Condition (i) restricts how fast the covariate composition can drift along the running variable. Condition (ii) restricts the slope in the covariate direction with the constant L_z. Condition (iii) ensures that each covariate value remains represented at every x, and it makes the within-bin estimators of the stratified analysis well defined for any density in the class.[22] The conditional densities satisfying Assumption 7 are collected in the class $\mathcal{F}_1^D(L_x,L_z,\underline{f})$.

These assumptions define the smoothness classes $\mathcal{F}_1^C(M,C_m,C_\tau)$ and $\mathcal{F}_1^D(L_x,L_z,\underline{f})$ over which the worst case is evaluated, and their product is analogous to the class $\mathcal{F}_p^R(M,\mathbf{C},\mathbf{L})$ defined by

[22]The uniform density on $\mathcal{Z}$ has the constant level $1/(2\zeta)$, and the lower bound on the density can be thought of as a fraction of the uniform density, $\underline{f}=\rho/(2\zeta)$ with $\rho\in(0,1]$. Note that $\rho=1$ implies that the conditional density is uniform across the support of the covariate, which the researcher can explicitly impose by assuming $L_x=L_z=0$. Further, if $L_z=0$ is assumed, the properties of the conditional density imply that $f_{Z|X}(\cdot\,|\,x)$ is the uniform density for every x. Setting $L_x=0$ in that case is therefore without loss of generality.

Assumption 3. In addition to the two smoothness classes I impose regularity assumptions on the joint density, which in the binary case follow from Assumptions 2 and 3.

Assumption 8 (Joint Density with a Continuous Covariate)**.** The joint density $f_{X,Z}(x,z)$ is continuous in x for every $z \in \mathcal{Z}$ and bounded away from 0 and infinity on $\mathcal{X} \times \mathcal{Z}$.

Finally, I impose the regularity conditions on the long error terms of Assumption 5 with the covariate ranging over $\mathcal{Z}$.

Assumption 9 (Regularity of the Long Error Terms with a Continuous Covariate)**.** The error terms $\{e_i\}_{i=1}^n$ satisfy $\mathbb{E}[|e_i|^{2+\eta} \,|\, X_i = x, Z_i = z] \leq 1/\eta$ for all $(x,z) \in \mathcal{X} \times \mathcal{Z}$ and some $\eta > 0$. The one-sided limits $\sigma^2_{L,\pm}(0,z)$ of the conditional variance $\sigma^2_L(x,z) = \mathrm{Var}(e_i \,|\, X_i = x, Z_i = z)$ at the cutoff exist for every $z \in \mathcal{Z}$, and $\inf_{z\in\mathcal{Z}} \sigma^2_{L,\pm}(0,z) > 0$.

The variance averaged across the covariate, $\sigma^2_L(x) = \mathbb{E}[\sigma^2_L(x, Z_i) \,|\, X_i = x]$, then also has one-sided limits $\sigma^2_{L,\pm}(0)$ at the cutoff.

Base RD Estimator with a Continuous Covariate. Since the base RD estimator ignores the covariate, it is the estimator defined in Section 2, but with $p = 1$ and $q = 0$ fixed for the continuous covariate extension.

Proposition 6.1 (Asymptotic Worst-Case Bias with a Continuous Covariate)**.** *Assume Assumption 1 and Assumptions 6 to 8. With the suprema taken over $(m,\tau) \in \mathcal{F}_1^C(M, C_m, C_\tau)$ and $f_{Z|X} \in \mathcal{F}_1^D(L_x, L_z, \underline{f})$, the asymptotic worst-case bias of $\hat{\tau}(h)$ is*

$$\overline{\mathrm{Bias}}(\hat{\tau}(h)) := \sup_{f_{Z|X}} \sup_{(m,\tau)} |\mathbb{E}[\hat{\tau}(h) \,|\, \mathbf{X}] - \tau| = h(b_B^{\mathrm{cont}} + o_p(1)),$$

where the first-order term is

$$b_B^{\mathrm{cont}} := B_{1,0}(k)\left[2M + (2C_m + C_\tau)\, L_x \zeta^2\right]. \tag{30}$$

The product $L_x\zeta^2$ takes the role of the composition constant L of the binary case, and C_m and C_τ bound the heterogeneity in the covariate values as in the binary case. Note that the asymptotic worst-case bias depends on the domain of Z. The larger it is, the further covariate values can lie apart, leading to comparisons of units distant in the covariate space. Since the base RD estimator ignores the continuous covariate, the results of Armstrong and Kolesár (2020) apply and its asymptotic variance is $\mathcal{V}_B$ in equation (25) with $q = 0$.

Stratification for a Continuous Covariate. To apply the Stratified RD estimator to a continuous covariate I partition the support $\mathcal{Z}$ of the covariate into J equal-width bins $\{I_j\}_{j=1}^J$ with bin width $w = 2\zeta/J$. Then for each unit I consider the indicator of whether its covariate Z_i falls into bin j, $\mathbb{1}_{\{Z_i \in I_j\}}$. Across the entire sample and all J bins this gives the realization of group membership, $\mathbf{Z}^{(J)} := \{\mathbb{1}_{\{Z_i \in I_j\}}\}_{i\leq n,\, j \leq J}$. For such a partition I then define the

regression function of the bin indicators $g_j(x) := \Pr(Z_i \in I_j \,|\, X_i = x)$, the baseline regression function $m_j(x) := \mathbb{E}[m(x, Z_i) \,|\, X_i = x,\, Z_i \in I_j]$ and the treatment effect regression function $\tau_j(x) := \mathbb{E}[\tau(x, Z_i) \,|\, X_i = x,\, Z_i \in I_j]$ conditional on the running variable and the bin indicator. In this setup, the target parameter can then be rewritten as

$$\tau = \sum_{j=1}^{J} g_j \tau_j, \tag{31}$$

with $\tau_j := \tau_j(0)$ and $g_j := g_j(0)$.

Based on this identity, the Stratified RD estimator then estimates this parameter by applying the two-step procedure defined in Section 3 with Z_i replaced by the indicator $\mathbb{1}_{\{Z_i \in I_j\}}$,

$$\hat{\tau}_j(h) = e_1^\top \arg\min_{\beta \in \mathbb{R}^{2q+2}} \sum_{i=1}^{n} k(X_i/h)(Y_i - S_q(X_i)^\top \beta)^2 \mathbb{1}_{\{Z_i \in I_j\}}, \tag{32a}$$

$$\hat{g}_j(h_z) = e_1^\top \arg\min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^{n} k(X_i/h_z)(\mathbb{1}_{\{Z_i \in I_j\}} - \tilde{S}_q(X_i)^\top \beta)^2, \tag{32b}$$

and then averages over the J bins,

$$\hat{\tau}_S(h, h_z) := \sum_{j=1}^{J} \hat{\tau}_j(h)\, \hat{g}_j(h_z).$$

The number of bins should increase with the number of units in the sample, because composition bias shrinks with the bin width. However, in small samples too many bins can lead to invertibility and degrees-of-freedom problems. Therefore, I assume that the number of bins grows slowly with the sample size, so that the partition becomes finer with smaller bin widths while each bin still contains enough units.

Assumption 10 (Growing Number of Bins)**.** The number of bins grows with the sample size, $J = J_n \to \infty$ with $J_n = o(\sqrt{nh})$.

The rate $J_n = o(\sqrt{nh})$ ensures that in large samples there are enough units in all bins, so that all within-bin estimators are well defined simultaneously. This condition implies that the effective sample size within each bin grows asymptotically, $wnh \to \infty$. Under Assumption 1 with $p = 1$ this further implies that the bin width shrinks slower than the bandwidth, $w/h \to \infty$. With the bin width shrinking at an appropriate rate, I show that the Stratified RD estimator is asymptotically normally distributed and bound its worst-case bias.

Proposition 6.2 (Asymptotic Normality and Worst-Case Bias of the Stratified RD Estimator with a Continuous Covariate)**.** *Assume Assumptions 1 and 4 and Assumptions 6 to 10. Then*

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

where $LB := \sum_{j=1}^{J_n} g_j\, \mathbb{E}[\hat{\tau}_j(h) - \tau_j \,|\, \mathbf{X}, \mathbf{Z}^{(J)}] + \sum_{j=1}^{J_n} \tau_j\, \mathbb{E}[\hat{g}_j(h_z) - g_j \,|\, \mathbf{X}]$ *is the leading bias and its*

worst case over $(m,\tau) \in \mathcal{F}_1^C(M, C_m, C_\tau)$ *and* $f_{Z|X} \in \mathcal{F}_1^D(L_x, L_z, \underline{f})$ *satisfies*

$$\overline{\text{Bias}}(\hat{\tau}_S(h, h_z)) := \sup_{f_{Z|X}} \sup_{(m,\tau)} |LB| \leq h(b_S^{\text{cont}} + o_p(1)),$$

where the first-order term is

$$b_S^{\text{cont}} := 2M\, B_{1,0}(k) + \kappa\, C_\tau L_x \zeta^2\, B_{1,0}^{\text{int}}(k). \tag{33}$$

The asymptotic variance can be written as

$$\mathcal{V}_S = \mathcal{V}_O + \frac{1}{\kappa} \frac{\widetilde{k_0^2}(k)\, \sigma_\tau^2(0)}{f_X(0)}, \qquad \mathcal{V}_O = \frac{\overline{k_0^2}(k)\, (\sigma_{L,+}^2(0) + \sigma_{L,-}^2(0))}{f_X(0)},$$

with $\sigma_\tau^2(x) := \text{Var}(\tau(0, Z_i) \,|\, X_i = x)$ *the variance of the treatment effect function across covariate values,* $\sigma_\tau^2(0)$ *its value evaluated at the cutoff and* $\sigma_{L,\pm}^2(0)$ *the one-sided limits of* $\sigma_L^2(x)$*.*

Importantly, as in Proposition 3.1, the first-order worst-case bias of the Stratified RD estimator contains no composition bias term that depends on the bound on heterogeneity C_m in the baseline regression function and the bound L_x on the slope of the conditional density function in the running variable. The elimination of this composition bias term results in the Stratified RD estimator having weakly lower first-order worst-case bias under suitably chosen bandwidths.

Given Propositions 6.1 and 6.2 and the base RD variance in equation (25), I compare the two estimators as in Section 4. The difference of the first-order worst-case bias terms is

$$b_B^{\text{cont}} - b_S^{\text{cont}} = L_x \zeta^2 \Big[2C_m\, B_{1,0}(k) + C_\tau (B_{1,0}(k) - \kappa\, B_{1,0}^{\text{int}}(k)) \Big], \tag{34}$$

and the difference of the asymptotic variances takes the form of equation (27) with $\text{Var}[m(0, Z_i) + \tau(0, Z_i)/2 \,|\, X_i = 0]$ and $\text{Var}[\tau(0, Z_i) \,|\, X_i = 0]$ in place of $g(1-g)(\Delta m(0) + \Delta\tau(0)/2)^2$ and $g(1-g)\,\Delta\tau(0)^2$. For local constant estimation with a symmetric kernel, $B_{1,0}^{\text{int}}(k) = B_{1,0}(k)$ and $2\,\widetilde{k_0^2}(k) = \overline{k_0^2}(k)$, so both thresholds equal one, $U_{1,0}(k) = L_{1,0}(k) = 1$. The asymptotic worst-case MSEs are defined as in equation (28) with $p = 1$ and b_S^{cont} and b_B^{cont}.

Theorem 6.3 (Estimator Comparison with a Continuous Covariate)**.** *Assume Assumptions 1 and 4 and Assumptions 6 to 10.*

(i) If $\kappa \leq 1$*, then* $b_S^{\text{cont}} \leq b_B^{\text{cont}}$*, with strict inequality if* $C_m L_x > 0$*.*

(ii) If $\kappa \geq 1$*, then* $\mathcal{V}_S \leq \mathcal{V}_B$*, with strict inequality if* $\text{Var}[m(0, Z_i) + \tau(0, Z_i)/2 \,|\, X_i = 0] > 0$*.*

(iii) If $\kappa = 1$*, then*

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

with strict inequality if $C_m L_x > 0$ *or* $\text{Var}[m(0, Z_i) + \tau(0, Z_i)/2 \,|\, X_i = 0] > 0$*.*

Part (iii) requires setting the bandwidth for estimating the weights equal to the bandwidth used for estimating the bin-specific treatment effects ($h_z = h$). The constant b_S^{cont} in Proposition 6.2 is an upper bound on the first-order worst-case bias of the Stratified RD estimator, and I conjecture this upper bound to be sharp. The conclusion of Theorem 6.3 does not depend on this conjecture, because the bound can only overstate the first-order worst-case bias of the Stratified RD estimator.

The result in Theorem 6.3 demonstrates that the Stratified RD estimator applied to a discretized continuous covariate achieves weakly lower worst-case AMSE than the base RD estimator if the discretization splits the covariate support into bins of equal width that shrink with the sample size. It follows from proving that the Stratified RD estimator has weakly lower asymptotic variance and weakly lower asymptotic worst-case bias if the weight-estimation bandwidth is chosen as $h_z = h$. The reduction in worst-case AMSE is strict if composition bias is not ruled out by the smoothness assumptions.

7 Conclusion

In many empirical RD studies covariates are observed and can be imbalanced in the samples researchers use for estimation, even if their distributions are continuous in the running variable. The results of this paper show that, under such covariate imbalance, the base RD estimator can suffer from composition bias if the expected outcome given the running variable varies across covariate values. The analysis of replication datasets of 9 published RD studies shows that within-window covariate imbalance is common. In 3 out of 9 studies the null hypothesis of covariate balance within the estimation window is rejected after correcting for multiple testing.

Based on these findings, I recommend that researchers nonparametrically adjust for covariates to address composition bias, unless they are willing to assume that covariates are balanced within the estimation window or not predictive of the outcome given the running variable. Specifically, I propose the Stratified RD estimator as one form of covariate adjustment. By stratifying on the covariate, the estimator ensures that only units with similar covariate values are compared, which eliminates composition bias for discrete covariates and reduces it for continuous covariates. The theoretical results of the paper show that the Stratified RD estimator has asymptotically weakly lower variance and worst-case bias than the base RD estimator. This leads to asymptotically shorter bias-aware confidence intervals for the average treatment effect at the cutoff, which makes the estimator relevant for empirical practice.

I demonstrate the favorable properties of the Stratified RD estimator in a calibrated simulation and a placebo exercise using the replication dataset of Londoño-Vélez et al. (2020). In both exercises stratification leads to a reduction in bias and thereby improves inference. Implementation of the Stratified RD estimator is straightforward for discrete covariates with a small number of support points relative to the sample size and for a single continuous covariate. Important questions of model selection and regularization in high-dimensional covariate settings, as well as studying other forms of covariate adjustment, are left for future research.

References

Arai, Yoichi, Taisuke Otsu, and Myung Hwan Seo (2025). "Regression Discontinuity Design with Potentially Many Covariates". *Econometric Theory* 41.6, pp. 1416–1451.

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

— (2020). "Simple and Honest Confidence Intervals in Nonparametric Regression". *Quantitative Economics* 11.1, pp. 1–39.

— (2021). "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects under Unconfoundedness". *Econometrica* 89.3, pp. 1141–1177.

Bertanha, Marinho and Marcelo J. Moreira (2020). "Impossible Inference in Econometrics: Theory and Applications". *Journal of Econometrics* 218.2, pp. 247–270.

Brown, Lawrence D., Mark G. Low, and Linda H. Zhao (1997). "Superefficiency in Nonparametric Function Estimation". *The Annals of Statistics* 25.6, pp. 2607–2625.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik (2019). "Regression Discontinuity Designs Using Covariates". *The Review of Economics and Statistics* 101.3, pp. 442–451.

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

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. 1st ed. Cambridge University Press.

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*. Ed. by José R. Zubizarreta, Elizabeth A. Stuart, Dylan S. Small, and Paul R. Rosenbaum. Boca Raton: Chapman and Hall/CRC, pp. 153–168.

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

Gelman, Andrew and Guido Imbens (2025). "Stable Adjustments in Regression Discontinuity. A Reply to Albada (Journal of Comments and Replications in Economics, 2025)". *Journal of Comments and Replications in Economics* 4.

Ghosh, Aditya, Guido Imbens, and Stefan Wager (2025). *PLRD: Partially Linear Regression Discontinuity Inference*. arXiv:2503.09907.

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

Imbens, Guido and Karthik Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". *The Review of Economic Studies* 79.3, pp. 933–959.

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

Imbens, Guido and Stefan Wager (2019). "Optimized Regression Discontinuity Designs". *The Review of Economics and Statistics* 101.2, pp. 264–278.

Kamat, Vishal (2018). "On Nonparametric Inference in the Regression Discontinuity Design". *Econometric Theory* 34.3, pp. 694–703.

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

Kramer, Patrick and Alexander Kreiß (2023). *Estimating Average Treatment Effects in Regression Discontinuity Designs with Covariates under Minimal Assumptions*. arXiv:2310.08976.

Kreiß, Alexander and Christoph Rothe (2023). "Inference in Regression Discontinuity Designs with High-Dimensional Covariates". *The Econometrics Journal* 26.2, pp. 105–123.

Lee, David S. (2008). "Randomized Experiments from Non-Random Selection in U.S. House Elections". *Journal of Econometrics* 142.2, pp. 675–697.

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

Londoño-Vélez, Juliana, Catherine Rodríguez, and Fabio Sánchez (2020). "Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia". *American Economic Journal: Economic Policy* 12.2, pp. 193–227.

Ma, Jun and Zhengfei Yu (2024). *Empirical Likelihood Covariate Adjustment for Regression Discontinuity Designs*. arXiv:2008.09263.

Noack, Claudia, Tomasz Olma, and Christoph Rothe (2026). "Flexible Covariate Adjustments in Regression Discontinuity Designs". *Journal of Econometrics* 257, p. 106298.

Noack, Claudia and Christoph Rothe (2024). "Bias-Aware Inference in Fuzzy Regression Discontinuity Designs". *Econometrica* 92.3, pp. 687–711.

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

Appendix

A Testing Covariate Balance in Replication Datasets

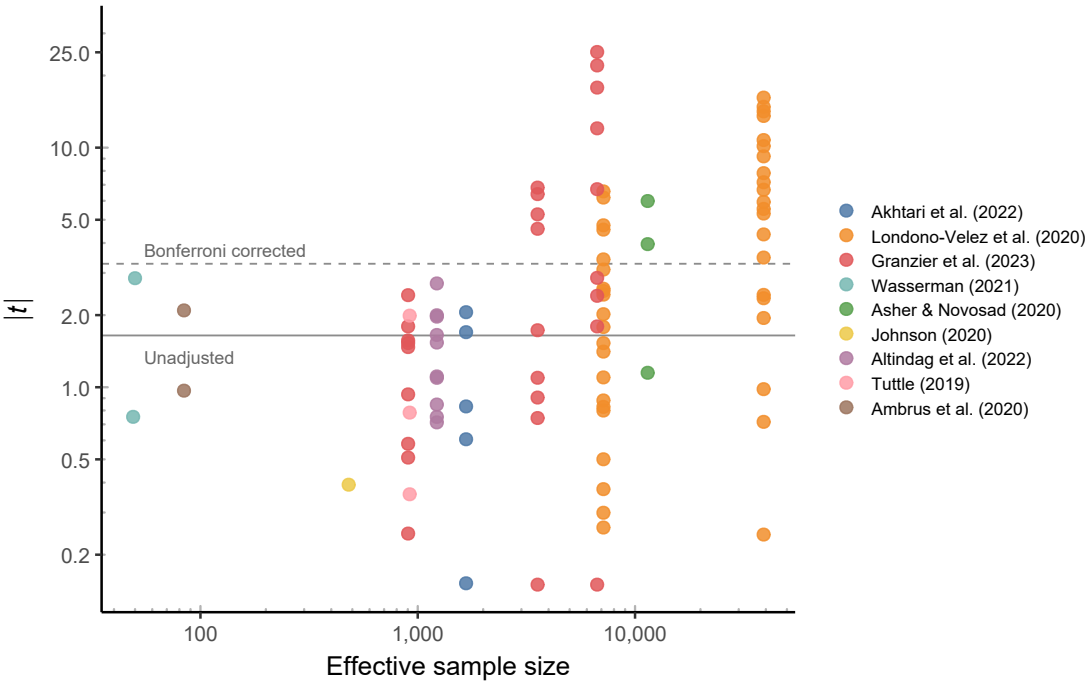


Figure 5: Within-Window Balance Tests and Effective Sample Size

Notes: Each point is one binary covariate in a replication study. The vertical axis shows the absolute Welch t statistic comparing covariate means above and below the cutoff, computed within the bandwidth reported in the original paper. Following the original papers, cluster-robust standard errors are used where assignment varies at a coarser level than the observation. The horizontal axis is the effective sample size, the number of clusters where the statistic is cluster-robust and the number of in-window observations otherwise. Some papers have two separate RD settings with separate running variables, so they appear at different effective sample sizes in the plot. Both axes are on a logarithmic scale. The solid horizontal line is the unadjusted two-sided 5% critical value, the dashed line the Bonferroni-adjusted critical value across all tests.

B Additional Results for the Empirical Application

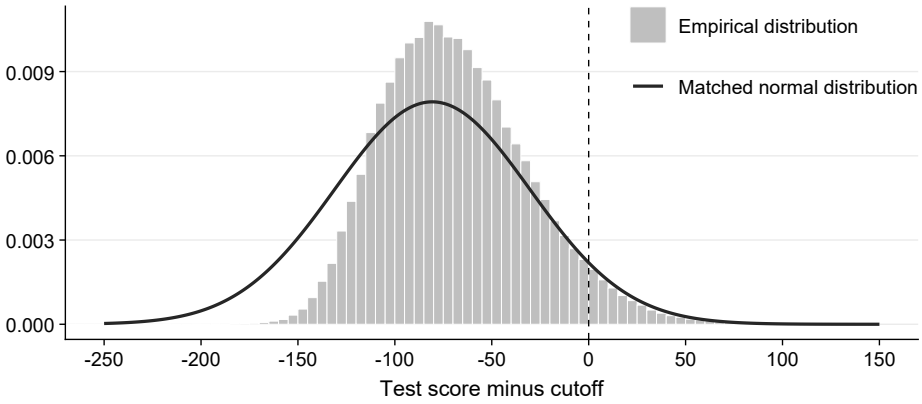


Figure 6: Empirical distribution of the test scores in bins of five points and the matched normal distribution $\mathcal{N}(-80.7, 50.4^2)$ used in the calibrated simulation.

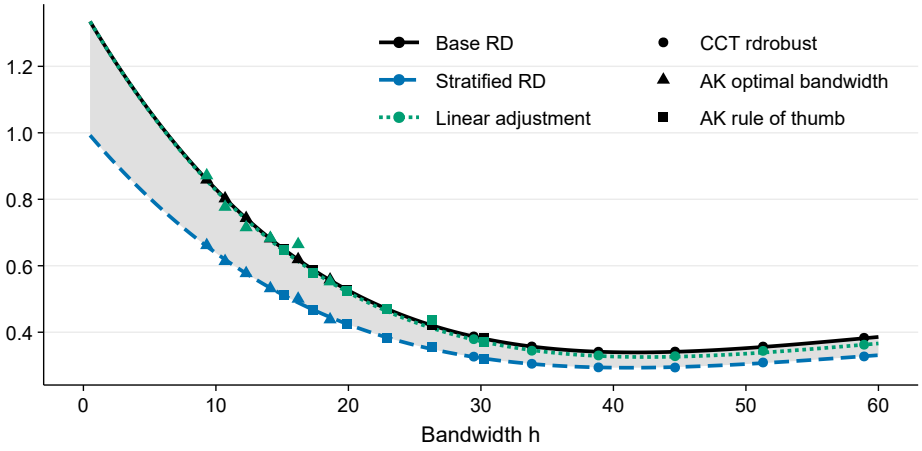
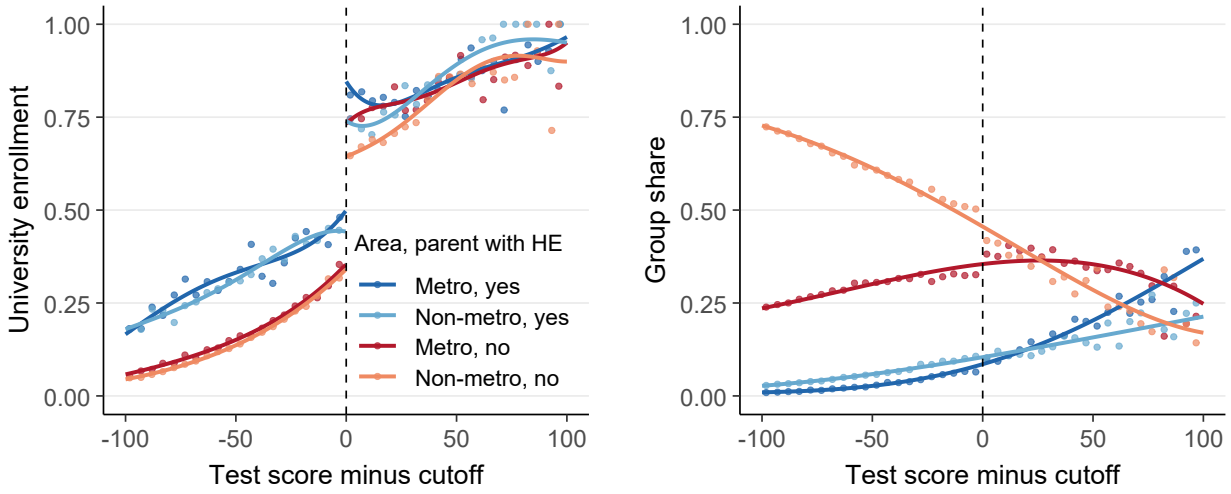


Figure 7: Specification bias and finite-sample bias in the calibrated simulation. The lines show the specification bias, scaled as in Figure 3b. The markers show the average finite-sample bias in the calibrated simulation, the mean over 1,000 samples of the conditional finite-sample bias divided by h^2, at the bandwidths of the three bandwidth selectors of Table 2.

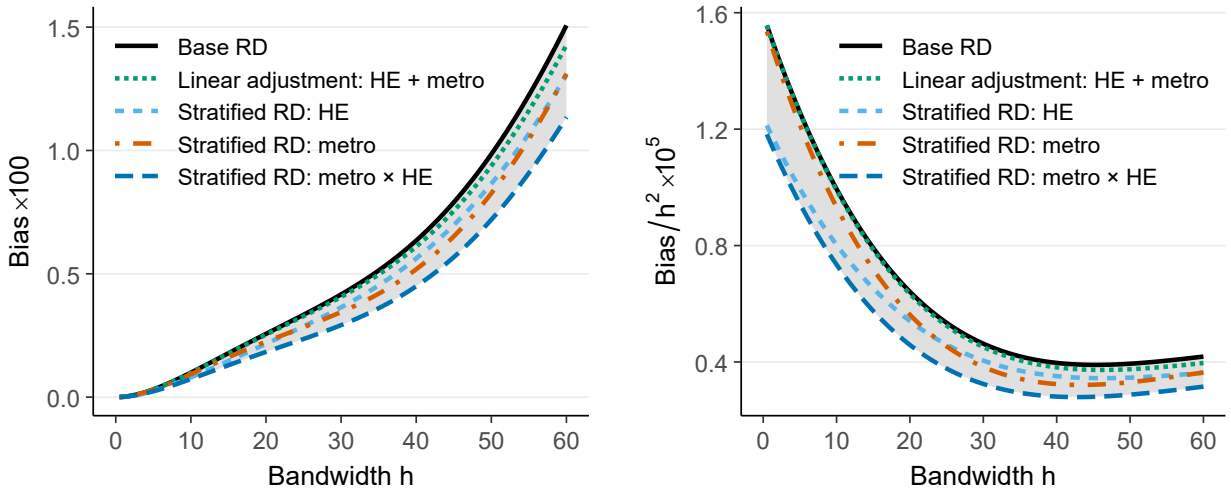
B.I Calibrated Simulation with Two Binary Covariates



(a) University enrollment rate by group

(b) Group shares

Figure 8: University enrollment and group composition of eligible students who took the SABER 11 test in fall 2014 ($N = 274{,}260$), with four groups defined by the SISBEN area (metropolitan or not) and by whether a parent has a higher education degree (HE). Dots are means over bins of five score points. In panel (a) the lines are quintic polynomials fitted separately on each side of the cutoff for each group. In panel (b) the lines are quintic polynomials fitted across the cutoff, and the colors are those of panel (a).



(a) Specification bias ×100

(b) Specification bias $\times 10^5$ scaled by $1/h^2$

Figure 9: Specification bias at fixed bandwidth in the calibration of Figure 8. The linear adjustment RD estimator adjusts for indicators of HE and the metropolitan area. The Stratified RD estimators stratify on HE, on the metropolitan area, or on both (metro × HE). All estimators use the same bandwidth h and the Stratified RD estimators estimate the group shares with $h_z = \kappa h$ as in Figure 3. The shaded area is the difference between the biases of base RD and Stratified RD on metro × HE.

Table 4: Calibrated Simulation with Four Groups

Bandwidth selection		Finite-sample bias ×100			SD ×100			CI length
n	h	Base	Strat.	Ratio	Base	Strat.	Ratio	Ratio
Panel A. CCT rdrobust								
10,000	58.9	1.446	1.085	0.75	3.78	3.90	1.03	0.58
80,000	38.9	0.605	0.427	0.71	1.56	1.56	1.00	0.59
320,000	29.5	0.407	0.284	0.70	0.92	0.91	0.99	0.59
Panel B. AK optimal bandwidth								
10,000	16.3	0.200	0.147	0.74	6.82	6.93	1.02	0.93
80,000	10.8	0.111	0.083	0.75	2.95	2.93	0.99	0.93
320,000	8.2	0.072	0.052	0.72	1.70	1.69	0.99	0.93
Panel C. AK rule of thumb								
10,000	30.2	0.422	0.295	0.70	5.16	5.27	1.02	0.73
80,000	19.9	0.256	0.185	0.72	2.14	2.13	1.00	0.73
320,000	15.1	0.180	0.129	0.72	1.28	1.26	0.99	0.73

Notes: Samples of size n drawn from the calibrated data-generating process of Figure 8, with 1,000 samples for each n. Strat. is the Stratified RD estimator with the four groups metro × HE. The bandwidth selectors and columns are as in Table 2, with the smoothness bounds of this calibration for the AK optimal bandwidth.

C Pooled and Long Regression Functions

This appendix presents the pooled and long regression models for a covariate Z_i that is unaffected by treatment and may be discrete or continuous. Where the discrete and continuous cases differ I state them separately. A discrete covariate with two support points gives the setup used in the main text.

C.I Pooled Regression Functions

$$Y_i = \bar{m}(X_i) + \bar{\tau}(X_i)\, T_i + u_i \tag{35}$$

with $\bar{m}(x) := \mathbb{E}[Y_i(0) \,|\, X_i = x]$, $\bar{\tau}(x) := \mathbb{E}[Y_i(1) - Y_i(0) \,|\, X_i = x]$, and $\mathbb{E}[u_i \,|\, X_i] = 0$ by the definition of the conditional expectation. Denote $\sigma_P^2(x) := \mathrm{Var}(u_i \,|\, X_i = x)$.

C.II Long Regression Functions

$$Y_i = m(X_i, Z_i) + \tau(X_i, Z_i)\, T_i + e_i \tag{36}$$

with $m(x,z) := \mathbb{E}[Y_i(0) \,|\, X_i = x, Z_i = z]$, $\tau(x,z) := \mathbb{E}[Y_i(1) - Y_i(0) \,|\, X_i = x, Z_i = z]$, and $\mathbb{E}[e_i \,|\, X_i, Z_i] = 0$ by the definition of the conditional expectation. Denote $\sigma_L^2(x,z) := \mathrm{Var}(e_i \,|\, X_i = x, Z_i = z)$ and its average over covariate values by $\sigma_L^2(x) := \mathbb{E}\big[\sigma_L^2(x, Z_i) \,|\, X_i = x\big]$. For discrete covariates this evaluates to $\mathbb{E}\big[\sigma_L^2(x, Z_i) \,|\, X_i = x\big] = \sum_z \sigma_L^2(x,z)\, f_{Z|X}(z \,|\, x)$, and for continuous covariates to $\mathbb{E}\big[\sigma_L^2(x, Z_i) \,|\, X_i = x\big] = \int \sigma_L^2(x,z)\, f_{Z|X}(z \,|\, x)\, dz$.

The pooled regression functions are linked to the long regression functions through the law of iterated expectations,

$$\bar{m}(x) = \mathbb{E}\big[m(x, Z_i) \,|\, X_i = x\big], \qquad \bar{\tau}(x) = \mathbb{E}\big[\tau(x, Z_i) \,|\, X_i = x\big], \tag{37}$$

so the pooled functions average the long functions with weights given by the conditional distribution $f_{Z|X}(z \,|\, x)$.

C.III Relating Error Terms of the Pooled and Long Regression Functions

Adding and subtracting the pooled regression functions in equation (36) gives

$$Y_i = \bar{m}(X_i) + \bar{\tau}(X_i)\, T_i + \underbrace{\eta_i + e_i}_{u_i},$$

where $\eta_i := \big[m(X_i, Z_i) - \bar{m}(X_i)\big] + \big[\tau(X_i, Z_i) - \bar{\tau}(X_i)\big] T_i$ is the difference between the long and pooled regression functions at (X_i, Z_i) and satisfies $\mathbb{E}[\eta_i \,|\, X_i] = 0$ by equation (37). The pooled error can be rewritten as $u_i = \eta_i + e_i$. For discrete Z_i the difference takes the form $\eta_i = \sum_j \big[m(X_i, j) + \tau(X_i, j) T_i\big] r_{ij}$ with the group-indicator errors $r_{ij} := \mathbb{1}_{\{Z_i = j\}} - f_{Z|X}(j \,|\, X_i)$. For binary Z_i it takes the form $\eta_i = \big[\Delta m(X_i) + \Delta\tau(X_i) T_i\big] r_i$ with $r_i := Z_i - g(X_i)$.

C.IV Variance Comparison

Lemma C.1 (Pooled and Long Variance Comparison)**.**

$$\sigma_P^2(X_i) = \mathrm{Var}\big(m(X_i, Z_i) + \tau(X_i, Z_i)\, T_i \,|\, X_i\big) + \sigma_L^2(X_i).$$

Proof. Note that $T_i = \mathbb{1}_{\{X_i \geq 0\}}$ is a function of X_i alone, so conditional on X_i it is a constant. Using the decomposition $u_i = \eta_i + e_i$,

$$\begin{aligned}
\sigma_P^2(X_i) &= \mathrm{Var}(u_i \,|\, X_i) \\
&= \mathbb{E}[u_i^2 \,|\, X_i] \\
&= \mathbb{E}[\eta_i^2 \,|\, X_i] + 2\,\mathbb{E}[\eta_i e_i \,|\, X_i] + \mathbb{E}[e_i^2 \,|\, X_i] \\
&= \mathbb{E}[\eta_i^2 \,|\, X_i] + \mathbb{E}[e_i^2 \,|\, X_i] \\
&= \mathrm{Var}\big(m(X_i, Z_i) + \tau(X_i, Z_i)\, T_i \,|\, X_i\big) + \sigma_L^2(X_i).
\end{aligned}$$

The cross term vanishes because η_i is a function of (X_i, Z_i), so $\mathbb{E}[\eta_i e_i \,|\, X_i] = \mathbb{E}\big[\eta_i\, \mathbb{E}[e_i \,|\, X_i, Z_i] \,|\, X_i\big] = 0$. The last line uses $\mathbb{E}[\eta_i \,|\, X_i] = 0$, so $\mathbb{E}[\eta_i^2 \,|\, X_i]$ is the conditional variance of $m(X_i, Z_i) + \tau(X_i, Z_i)T_i$ given X_i, and $\mathbb{E}[e_i^2 \,|\, X_i] = \mathbb{E}\big[\sigma_L^2(X_i, Z_i) \,|\, X_i\big] = \sigma_L^2(X_i)$. □

Conditional on $X_i = x$, the treatment indicator equals $\mathbb{1}_{\{x \geq 0\}}$, so the between-group variance is $\mathrm{Var}(m(X_i, Z_i) \,|\, X_i = x)$ for $x < 0$ and $\mathrm{Var}(m(X_i, Z_i) + \tau(X_i, Z_i) \,|\, X_i = x)$ for $x \geq 0$. For binary Z_i it equals $[\Delta m(x) + \Delta\tau(x)\mathbb{1}_{\{x \geq 0\}}]^2\, \sigma_Z^2(x)$.

C.V Covariate Regression Function

As usual in the covariate adjustment RD literature, I assume that the covariate is unaffected by treatment (as is the case for predetermined covariates), so no treatment effect function appears in the covariate regression functions below.

Discrete Covariate. For a discrete $Z_i \in \{0, 1, \ldots, J\}$ each support point induces a group indicator $Z_{ij} := \mathbb{1}_{\{Z_i = j\}}$. These $J + 1$ binary variables then have the regression model

$$Z_{ij} = g_j(X_i) + r_{ij}, \qquad g_j(x) := \mathbb{E}[Z_{ij} \,|\, X_i = x], \tag{38}$$

with $g_j(x) = f_{Z|X}(j \,|\, x)$ and $\mathbb{E}[r_{ij} \,|\, X_i] = 0$ by the definition of the conditional expectation.

For binary Z_i the indicators can be written as $Z_{i1} = Z_i$ and $Z_{i0} = 1 - Z_i$. Then equation (38) reduces to $Z_i = g(X_i) + r_i$, with the covariate regression function g referenced throughout the main text.

Continuous Covariate. For continuous Z_i the composition at x is described by the conditional distribution given the running variable. The conditional distribution function is the regression

function of the indicator $\mathbb{1}_{\{Z_i \le t\}}$ for each t,

$$F_{Z|X}(t \,|\, x) = \mathbb{E}[\mathbb{1}_{\{Z_i \le t\}} \,|\, X_i = x],$$

and the corresponding conditional density is $f_{Z|X}(z \,|\, x)$.

The Stratified RD estimator of Section 6.2 creates groups by partitioning $\mathcal{Z}$ into bins I_j. The bin indicators $Z_{ij} = \mathbb{1}_{\{Z_i \in I_j\}}$ then follow equation (38) with shares $g_j(x) = \mathbb{E}[\mathbb{1}_{\{Z_i \in I_j\}} \,|\, X_i = x] = \int_{I_j} f_{Z|X}(z \,|\, x)\, dz$, so the discrete case applies to the discretized covariate.

D Implied Regularity Conditions

This appendix shows that the regularity conditions on the marginal density (Assumption 2) and on the long error terms (Assumption 5), together with the restrictions on the primitives (Assumption 3), imply the analogous conditions for the joint density, the error terms of the covariate regression and the error terms of the pooled regression. Both estimators are therefore analyzed under a common set of assumptions. Recall the covariate error $r_i := Z_i - g(X_i)$ and the pooled error $u_i := Y_i - \bar{m}(X_i) - \bar{\tau}(X_i)T_i$, and their conditional variances $\sigma_Z^2(x) := \mathrm{Var}(r_i \mid X_i = x)$ and $\sigma_P^2(x) := \mathrm{Var}(u_i \mid X_i = x)$.

Lemma D.1 (Implied Regularity Conditions)**.** *Assume Assumptions 2, 3 and 5. Then the following holds for all* $(m, \tau, g) \in \mathcal{F}_p^R(M, \mathbf{C}, \mathbf{L})$.

(i) The densities $f_{X,Z}(\cdot, z)$, $z \in \{0, 1\}$, *are continuous at the cutoff and satisfy* $f_{X,Z}(0, z) \geq \underline{g}\, f_X(0) > 0$.

(ii) The covariate error terms satisfy $\mathbb{E}[|r_i|^{2+\eta} \mid X_i = x] \leq 1$ *for all* x, *and* σ_Z^2 *is continuous at the cutoff with* $\sigma_Z^2(0) > 0$.

(iii) The pooled error terms satisfy $\mathbb{E}[|u_i|^{2+\eta} \mid X_i = x] \leq 1/\eta'$ *for all* x, *for a constant* η' *that depends only on* η, $C_{\Delta m}$ *and* $C_{\Delta\tau}$, *and the one-sided limits* $\sigma_{P,+}^2(0) := \lim_{x \downarrow 0} \sigma_P^2(x)$ *and* $\sigma_{P,-}^2(0) := \lim_{x \uparrow 0} \sigma_P^2(x)$ *exist and are strictly positive.*

Proof. The functions g, $m(\cdot, z)$ and $m(\cdot, z) + \tau(\cdot, z)$ are continuous by Assumption 3(i) and (ii).

(i) Since Z_i is binary, $f_{X,Z}(x, 1) = f_X(x)\, g(x)$ and $f_{X,Z}(x, 0) = f_X(x)\,(1 - g(x))$. Both are continuous at the cutoff, and $g(0) \in [\underline{g}, 1 - \underline{g}]$ gives $f_{X,Z}(0, z) \geq \underline{g}\, f_X(0)$, which is positive by Assumption 2.

(ii) Since Z_i and $g(X_i)$ lie in $[0, 1]$, $|r_i| \leq 1$, which gives the moment bound. The conditional variance $\sigma_Z^2(x) = g(x)(1 - g(x))$ is continuous, and $\sigma_Z^2(0) \geq \underline{g}(1 - \underline{g}) > 0$.

(iii) Since $m(x, z) - \bar{m}(x) = \Delta m(x)\,(z - g(x))$ and analogously for τ, the pooled error is $u_i = [\Delta m(X_i) + \Delta\tau(X_i)\, T_i]\, r_i + e_i$. The coefficient on r_i is bounded by $C_{\Delta m} + C_{\Delta\tau}$ by Assumption 3(iii) and (iv), so Lemma J.1, the moment bound from part (ii) and Assumption 5 give

$$\begin{aligned}\mathbb{E}[|u_i|^{2+\eta} \mid X_i = x] &\leq 2^{1+\eta}\Big(|\Delta m(x) + \Delta\tau(x)\, \mathbb{1}_{\{x \geq 0\}}|^{2+\eta}\ \mathbb{E}[|r_i|^{2+\eta} \mid X_i = x] + \mathbb{E}[|e_i|^{2+\eta} \mid X_i = x]\Big) \\ &\leq 2^{1+\eta}\big((C_{\Delta m} + C_{\Delta\tau})^{2+\eta} + 1/\eta\big) =: 1/\eta'.\end{aligned}$$

The pooled variance can be decomposed as $\sigma_P^2(x) = [\Delta m(x) + \Delta\tau(x)\, \mathbb{1}_{\{x \geq 0\}}]^2\, \sigma_Z^2(x) + \sigma_L^2(x)$, because $\mathbb{E}[e_i \mid X_i, Z_i] = 0$. Both terms have one-sided limits at the cutoff by continuity of Δm, $\Delta\tau$ and g and by Assumption 5, and $\sigma_{P,\pm}^2(0) \geq g(0)\, \sigma_{L,\pm}^2(0, 1) + (1 - g(0))\, \sigma_{L,\pm}^2(0, 0) > 0$. □

E Preliminaries

E.I Weighted Sum Representation of Estimators

Let k be a symmetric and nonnegative kernel that is bounded with bounded support. 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 equivalently using $k_-(v) := k(v)\mathbb{1}_{\{v<0\}}$. Let $m_q(u) := (1, u, \ldots, u^q)^\top$. For simplicity, I restrict the analysis to local polynomial estimators, with equal bandwidth choices on both sides of the cutoff, and a kernel that is bounded and has bounded support (as done in Armstrong and Kolesár (2020)).

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

$$\begin{aligned}\tilde{k}_{n,+}(u) =& e_1^\top \left(\frac{1}{nh}\sum_{i=1}^n k_+(X_i/h)m_q(X_i/h)m_q(X_i/h)^\top\right)^{-1} m_q(u)k_+(u) &(39)\\ :=& e_1^\top(\tilde{Q}_{n,+})^{-1}m_q(u)k_+(u) &(40)\end{aligned}$$

and $\tilde{k}_{n,-}$ is defined analogously. By Lemma E.4, the weights converge uniformly in probability to

$$\begin{aligned}\tilde{k}_+(u) =& e_1^\top \left(f_X(0)\int_0^\infty k_+(t)m_q(t)m_q(t)^\top dt\right)^{-1} m_q(u)k_+(u) &(41)\\ :=& e_1^\top(\tilde{Q}_+)^{-1}m_q(u)k_+(u) &(42)\\ =& \frac{1}{f_X(0)}k_q^*(u)\mathbb{1}_{\{u\geq 0\}} &(43)\end{aligned}$$

and

$$\begin{aligned}\tilde{k}_-(u) =& e_1^\top \left(f_X(0)\int_{-\infty}^0 k_-(t)m_q(t)m_q(t)^\top dt\right)^{-1} m_q(u)k_-(u) &(44)\\ :=& e_1^\top(\tilde{Q}_-)^{-1}m_q(u)k_-(u) &(45)\\ =& \frac{1}{f_X(0)}k_q^*(-u)\mathbb{1}_{\{u<0\}} &(46)\end{aligned}$$

where $k_q^*(u)$ is the equivalent kernel defined in Section 2.

Within-Group Local Polynomial Estimator. The base RD estimator applied within each group can be written as $\hat{\tau}_j(h) = \frac{1}{nh}\sum_{i=1}^n \tilde{k}_{j,n}(X_i/h, Z_i)Y_i$, where the subsample restriction yields:

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

The limiting weights are given by

$$\begin{aligned}\tilde{k}_{j,+}(u,z) =& e_1^\top \left(f_{X,Z}(0,z) \int_0^\infty k_+(t) m_q(t) m_q(t)^\top dt \right)^{-1} m_q(u) k_+(u) \mathbb{1}_{\{z=j\}} \\ :=& e_1^\top \left(\tilde{Q}_{j,+} \right)^{-1} m_q(u) k_+(u) \mathbb{1}_{\{z=j\}} \\ =& \frac{1}{f_{X,Z}(0,j)} k_q^*(u) \mathbb{1}_{\{u\geq 0\}} \mathbb{1}_{\{z=j\}}\end{aligned}$$

with $\tilde{k}_{j,n,-}(u,z)$ and $\tilde{k}_{j,-}(u,z)$ defined analogously.

Weight Estimator. The weight estimator can be written as $\hat{g}(h_z) = \frac{1}{nh_z}\sum_{i=1}^n \tilde{k}_n^z(X_i/h_z) Z_i$, where

$$\begin{aligned}\tilde{k}_n^z(u) =& e_1^\top \left(\underbrace{\frac{1}{nh_z}\sum_{i=1}^n k(X_i/h_z) m_q(X_i/h_z) m_q(X_i/h_z)^\top}_{:=\tilde{Q}_n^z} \right)^{-1} m_q(u) k(u) \\ =& e_1^\top (\tilde{Q}_n^z)^{-1} m_q(u) k(u).\end{aligned}$$

The limiting weights are given by

$$\begin{aligned}\tilde{k}^z(u) =& e_1^\top \left(f_X(0) \int_{\mathbb{R}} k(t) m_q(t) m_q(t)^\top dt \right)^{-1} m_q(u) k(u) \\ :=& e_1^\top \left(\tilde{Q}^z \right)^{-1} m_q(u) k(u) \\ =& \frac{1}{f_X(0)} k_q^{\text{int}}(u),\end{aligned}$$

where k_q^{int} is the interior equivalent kernel defined in Proposition 3.1.

By Lemma E.4, the weights of local polynomial estimators with a symmetric, nonnegative and bounded kernel with bounded support satisfy Assumption B.4 in Armstrong and Kolesár (2020) in probability, and by Lemma E.1 (or Lemma E.2) the same holds for their Assumption 3.1. The

worst-case bias result of their Theorem 3.1 then follows under the random design used in this paper by the argument in the proof of their Theorem B.4, with $o_p(1)$ in place of $o(1)$.

E.II Preliminary Lemmas

Lemma E.1. *Let $X_1, \dots, X_n$ be i.i.d. with density f_X that is continuous and positive at the cutoff, and let ψ be a bounded function with bounded support $\mathcal{S}$. If $h \to 0$ and $nh \to \infty$, then*

$$\frac{1}{nh} \sum_{i=1}^{n} \psi\left(\frac{X_i}{h}\right) \xrightarrow{p} f_X(0) \int_{\mathcal{S}} \psi(u)\, du.$$

This is Assumption 3.1 in Armstrong and Kolesár (2020) with $d = f_X(0)$ and convergence in probability in place of deterministic convergence.

Proof. Write

$$\begin{aligned}\frac{1}{nh} \sum_{i=1}^{n} \psi\left(\frac{X_i}{h}\right) - f_X(0) \int_{\mathcal{S}} \psi(u)\, du = &\left(\frac{1}{nh} \sum_{i=1}^{n} \psi\left(\frac{X_i}{h}\right) - \frac{1}{h}\,\mathbb{E}\left[\psi\left(\frac{X_i}{h}\right)\right]\right) \\ &+ \left(\frac{1}{h}\,\mathbb{E}\left[\psi\left(\frac{X_i}{h}\right)\right] - f_X(0) \int_{\mathcal{S}} \psi(u)\, du\right).\end{aligned}$$

By Chebyshev's inequality, the first term converges to zero in probability if the variance of the average vanishes. The second term is deterministic and converges to zero.

For the second term, a change of variables gives

$$\frac{1}{h}\,\mathbb{E}\left[\psi\left(\frac{X_i}{h}\right)\right] - f_X(0) \int_{\mathcal{S}} \psi(u)\, du = \int_{\mathcal{S}} \psi(u)\,[f_X(uh) - f_X(0)]\, du.$$

Let $\bar{s} < \infty$ satisfy $\mathcal{S} \subseteq [-\bar{s}, \bar{s}]$, which exists because $\mathcal{S}$ is bounded. Then $|uh| \le h\bar{s}$ for every $u \in \mathcal{S}$, so the absolute value of the right-hand side is at most

$$\sup_{u \in \mathcal{S}} |\psi(u)| \int_{\mathcal{S}} du \sup_{|x| \le h\bar{s}} |f_X(x) - f_X(0)|.$$

The first two factors are finite because ψ and $\mathcal{S}$ are bounded. The last factor vanishes as $h \to 0$ because f_X is continuous at 0.

For the first term, the X_i are i.i.d., so

$$\operatorname{Var}\left(\frac{1}{nh} \sum_{i=1}^{n} \psi\left(\frac{X_i}{h}\right)\right) \le \frac{1}{nh} \cdot \frac{1}{h}\,\mathbb{E}\left[\psi\left(\frac{X_i}{h}\right)^2\right] \to 0.$$

The previous step applied to ψ^2, which is also bounded with bounded support, shows that $\frac{1}{h}\,\mathbb{E}[\psi(X_i/h)^2]$ converges to $f_X(0) \int_{\mathcal{S}} \psi(u)^2\, du < \infty$. Hence the variance vanishes because $nh \to \infty$. □

Lemma E.2 (Groupwise Assumption 3.1)**.** *Let $(X_1, Z_1), \ldots, (X_n, Z_n)$ be i.i.d. and fix $z \in \{0, 1\}$. Assume that $f_{X,Z}(\cdot, z)$ is continuous and positive at the cutoff, and let ψ be a bounded function with bounded support $\mathcal{S}$. If $h \to 0$ and $nh \to \infty$, then*

$$\frac{1}{nh} \sum_{i=1}^{n} \psi\left(\frac{X_i}{h}\right) \mathbb{1}_{\{Z_i = z\}} \xrightarrow{p} f_{X,Z}(0, z) \int_{\mathcal{S}} \psi(u)\, du.$$

Proof. Analogous to the proof of Lemma E.1. □

Lemma E.3 (Limiting Gram Matrices)**.** *Let k be symmetric, nonnegative, and bounded with bounded support and let $h \to 0$ and $nh \to \infty$.*

(i) If f_X is continuous and positive at the cutoff, then

$$\tilde{Q}_{n,\pm} \xrightarrow{p} \tilde{Q}_{\pm} \quad \text{and} \quad \tilde{Q}_{n,\pm}^{-1} \xrightarrow{p} \tilde{Q}_{\pm}^{-1}.$$

(ii) If $f_{X,Z}(\cdot, j)$ is continuous and positive at the cutoff, then

$$\tilde{Q}_{j,n,\pm} \xrightarrow{p} \tilde{Q}_{j,\pm} \quad \text{and} \quad \tilde{Q}_{j,n,\pm}^{-1} \xrightarrow{p} \tilde{Q}_{j,\pm}^{-1}.$$

(iii) If f_X is continuous and positive at the cutoff, $h_z \to 0$, and $nh_z \to \infty$, then

$$\tilde{Q}_n^z \xrightarrow{p} \tilde{Q}^z \quad \text{and} \quad (\tilde{Q}_n^z)^{-1} \xrightarrow{p} (\tilde{Q}^z)^{-1}.$$

Proof. For $r, s \in \{0, \ldots, q\}$, the (r, s) entry of the empirical gram matrix $\tilde{Q}_{n,+}$ is $\frac{1}{nh} \sum_{i=1}^{n} (X_i/h)^{r+s} k_+(X_i/h)$. The function $u^{r+s} k_+(u)$ is bounded with bounded support, so applying Lemma E.1 elementwise gives $\tilde{Q}_{n,+} \xrightarrow{p} \tilde{Q}_+$. Since $\tilde{Q}_+$ is positive definite, $\tilde{Q}_{n,+}^{-1} \xrightarrow{p} \tilde{Q}_+^{-1}$ by the continuous mapping theorem. The control side follows in the same way with k_- in place of k_+.

Part (ii) follows analogously with Lemma E.2 in place of Lemma E.1, and part (iii) with h_z in place of h and k in place of k_+. □

Lemma E.4 (Equivalent Kernel Weights)**.** *Let k be symmetric, nonnegative, and bounded with bounded support and let $h \to 0$ and $nh \to \infty$.*

(i) If f_X is continuous and positive at the cutoff, then

$$\sup_u |\tilde{k}_{n,\pm}(u) - \tilde{k}_{\pm}(u)| = o_p(1).$$

(ii) If $f_{X,Z}(\cdot, j)$ is continuous and positive at the cutoff, then

$$\sup_{u,z} |\tilde{k}_{j,n,\pm}(u, z) - \tilde{k}_{j,\pm}(u, z)| = o_p(1).$$

(iii) If f_X is continuous and positive at the cutoff, $h_z \to 0$, and $nh_z \to \infty$, then

$$\sup_u |\tilde{k}_n^z(u) - \tilde{k}^z(u)| = o_p(1).$$

Proof. Lemma E.3 gives the convergence of the empirical gram matrices and their inverses. Plugging the gram matrices into the definitions of $\tilde{k}_{n,+}(u)$ and $\tilde{k}_+$ gives,

$$\begin{aligned}
\sup_u |\tilde{k}_{n,+}(u) - \tilde{k}_+(u)| &= \sup_u |e_1^\top (\tilde{Q}_{n,+}^{-1} - \tilde{Q}_+^{-1})\, m_q(u)\, k_+(u)| \\
&\leq \|e_1^\top (\tilde{Q}_{n,+}^{-1} - \tilde{Q}_+^{-1})\| \sup_u \|m_q(u)\, k_+(u)\| \\
&= o_p(1)
\end{aligned}$$

by the Cauchy–Schwarz inequality, where $\|\cdot\|$ denotes the Euclidean norm. The term $\|e_1^\top (\tilde{Q}_{n,+}^{-1} - \tilde{Q}_+^{-1})\|$ converges in probability to zero due to the elementwise convergence implied by Lemma E.3 and the last supremum is finite because k is bounded with bounded support. The control side follows by the symmetry of k.

Part (ii) follows analogously, and part (iii) with h_z in place of h and k in place of k_+. □

F Proof of Proposition 2.1

The subclass restricts heterogeneity between the two groups to a constant level shift in baseline outcomes and a constant level shift in treatment effects, ruling out slope and higher-order heterogeneity in Δm and $\Delta\tau$. Since $\mathcal{F}_p^R(M, \mathbf{C}^*, \mathbf{L}) \subseteq \mathcal{F}_p^R(M, \mathbf{C}, \mathbf{L})$, the supremum of the conditional bias over this subclass lower-bounds the supremum over the full class.

Under the vertical-shift restriction the group regression functions satisfy $m(s,1) = m(s,0) + \beta$ for some constant β with $|\beta| \le C_{\Delta m}$, and the treatment effect difference is a constant δ with $|\delta| \le C_{\Delta\tau}$.[23] The pooled regression functions in the subclass hence have the following p-th derivatives on each side of the cutoff,

$$\bar{m}^{(p)}(s) = m^{(p)}(s,0) + \beta\, g^{(p)}(s),$$
$$\bar{m}^{(p)}(s) + \bar{\tau}^{(p)}(s) = \big[m^{(p)}(s,0) + \tau^{(p)}(s,0)\big] + (\beta+\delta)\, g^{(p)}(s).$$

Following Armstrong and Kolesár (2020), the conditional bias of the base RD estimator can be written as

$$\mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}] = \int_0^\infty \bar{w}(s)\big[\bar{m}^{(p)}(s) + \bar{\tau}^{(p)}(s)\big]ds - \int_{-\infty}^0 \bar{w}(s)\, \bar{m}^{(p)}(s)\, ds,$$

where $\bar{w}(s)$ is the weight function of Lemma B.2 in the online appendix of Armstrong and Kolesár (2020).

Upper bound. Substituting the p-th derivatives with the subclass derivatives separates the bias into a within-group component and a heterogeneity component,

$$\mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}] = \underbrace{\int_0^\infty \bar{w}(s)\big[m^{(p)}(s,0) + \tau^{(p)}(s,0)\big]ds - \int_{-\infty}^0 \bar{w}(s)\, m^{(p)}(s,0)\, ds}_{\text{within-group component}}$$
$$+ \underbrace{(\beta+\delta)\int_0^\infty \bar{w}(s)\, g^{(p)}(s)\, ds - \beta\int_{-\infty}^0 \bar{w}(s)\, g^{(p)}(s)\, ds}_{\text{heterogeneity component}}.$$

By condition (i) of Assumption 3, $|m^{(p)}(s,0)| \le M$ and $|m^{(p)}(s,0) + \tau^{(p)}(s,0)| \le M$, by condition (ii), $|g^{(p)}(s)| \le L$, and in the subclass $|\beta| \le C_{\Delta m}$ and $|\delta| \le C_{\Delta\tau}$. The triangle inequality therefore

[23] For $p = 1$ condition (v) of Assumption 3 is empty, so Δm is not necessarily constant. Then $\bar{m}'(s) = \big[g(s)\, m'(s,1) + (1 - g(s))\, m'(s,0)\big] + \Delta m(s)\, g'(s)$, and analogously for $\bar{m}'(s) + \bar{\tau}'(s)$. The bracket is a convex combination bounded by M, so the argument below goes through unchanged.

gives

$$
\begin{aligned}
|\mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}]| \leq & \int_0^\infty |\bar{w}(s)||m^{(p)}(s,0) + \tau^{(p)}(s,0)|\, ds + \int_{-\infty}^0 |\bar{w}(s)||m^{(p)}(s,0)|\, ds \\
& + |\beta + \delta| \int_0^\infty |\bar{w}(s)||g^{(p)}(s)|\, ds + |\beta| \int_{-\infty}^0 |\bar{w}(s)||g^{(p)}(s)|\, ds \\
& \leq \big(M + (C_{\Delta m} + C_{\Delta \tau})L\big) \int_0^\infty |\bar{w}(s)|\, ds + \big(M + C_{\Delta m} L\big) \int_{-\infty}^0 |\bar{w}(s)|\, ds.
\end{aligned}
$$

Attainment. The bounds $M + C_{\Delta m} L$ and $M + (C_{\Delta m} + C_{\Delta \tau})L$ on the left and right pooled p-th derivatives are attained simultaneously by $\beta = C_{\Delta m}$, $\delta = C_{\Delta \tau}$, within-group p-th derivatives equal to $M \operatorname{sign}(\bar{w}(s)) \operatorname{sign}(s)$, and $g^{(p)}(s) = L \operatorname{sign}(\bar{w}(s)) \operatorname{sign}(s)$. These functions lie in the subclass, because $\mathcal{F}_p$ only bounds the p-th derivative almost everywhere, so it may change discontinuously at the cutoff and the finitely many points where $\bar{w}$ changes sign. The constraint that g lies in $[0, 1]$ because it is a conditional probability does not bind for h small enough relative to L if one sets $g(0) = 1/2$.

First-order term. By the argument in the proof of Theorem B.4 in Armstrong and Kolesár (2020), which applies under random design by Lemmas E.1 and E.4, both $\int_0^\infty |\bar{w}(s)|\, ds$ and $\int_{-\infty}^0 |\bar{w}(s)|\, ds$ equal $h^p\, B_{p,q}(k)/p!\, (1 + o_p(1))$. Hence

$$
\begin{aligned}
\sup_{(m,\tau,g) \in \mathcal{F}_p^R(M, \mathbf{C}^*, \mathbf{L})} |\mathbb{E}[\hat{\tau}(h) - \tau \,|\, \mathbf{X}]| &= \big(M + (C_{\Delta m} + C_{\Delta \tau})L\big) \int_0^\infty |\bar{w}(s)|\, ds + \big(M + C_{\Delta m} L\big) \int_{-\infty}^0 |\bar{w}(s)|\, ds \\
&= \big(2M + 2C_{\Delta m} L + C_{\Delta \tau} L\big)\, h^p\, \frac{B_{p,q}(k)}{p!}\, \big(1 + o_p(1)\big) \\
&= h^p \big(b_B^{\text{low}} + o_p(1)\big).
\end{aligned}
$$

G Proof of Proposition 3.1

G.I Central Limit Theorem for Stratified RD

To apply the Lindeberg central limit theorem to the Stratified RD estimator, I first split the estimation error in both steps of the estimator into its conditional bias and stochastic term. Then I linearize it to focus on the leading terms in the asymptotic expansion and subtract the leading bias LB, which leaves a weighted sum of the error terms. Then I replace the empirical weights in this sum with their limits, and show that this replacement is asymptotically negligible. Finally, I verify the Lindeberg conditions for the resulting sum containing the limiting local polynomial weights.

Decomposition into Conditional Bias Term and Stochastic Term. Inserting equation (36) into $\hat{\tau}_j(h) = \frac{1}{nh}\sum_{i=1}^n \tilde{k}_{j,n}(X_i/h, Z_i)\, Y_i$ gives

$$\hat{\tau}_j(h) - \tau_j = \underbrace{\frac{1}{nh}\sum_{i=1}^n \tilde{k}_{j,n}(X_i/h, Z_i)\big[m(X_i, Z_i) + \tau(X_i, Z_i)\, T_i\big] - \tau_j}_{=:B_j} + \underbrace{\frac{1}{nh}\sum_{i=1}^n \tilde{k}_{j,n}(X_i/h, Z_i)\, e_i}_{=:S_j}. \quad (47)$$

Likewise, inserting $Z_i = g(X_i) + r_i$ into $\hat{g}(h_z) = \frac{1}{nh_z}\sum_{i=1}^n \tilde{k}^z_n(X_i/h_z)\, Z_i$ gives

$$\hat{g}(h_z) - g = \underbrace{\frac{1}{nh_z}\sum_{i=1}^n \tilde{k}^z_n(X_i/h_z)\, g(X_i) - g}_{=:B_g} + \underbrace{\frac{1}{nh_z}\sum_{i=1}^n \tilde{k}^z_n(X_i/h_z)\, r_i}_{=:S_g}. \quad (48)$$

Since $\mathbb{E}[e_i \,|\, \mathbf{X}, \mathbf{Z}] = 0$ and $\mathbb{E}[r_i \,|\, \mathbf{X}] = 0$, the terms B_j and B_g are the conditional biases.

Linearization.

Lemma G.1 (Linearization of the Stratified RD Estimator)**.** *Under the assumptions of Proposition 3.1,*

$$\hat{\tau}_S(h, h_z) - \tau = g\,[\hat{\tau}_1(h) - \tau_1] + (1-g)\,[\hat{\tau}_0(h) - \tau_0] + [\tau_1 - \tau_0]\,[\hat{g}(h_z) - g] + o_p\big((nh)^{-1/2}\big).$$

Proof. Since $\tau = g\,\tau_1 + (1-g)\,\tau_0$ by equation (1), rearranging $\hat{\tau}_S(h, h_z) = \hat{g}(h_z)\,\hat{\tau}_1(h) + (1-\hat{g}(h_z))\,\hat{\tau}_0(h)$ gives the exact identity

$$\hat{\tau}_S(h, h_z) - \tau = g\,[\hat{\tau}_1(h) - \tau_1] + (1-g)\,[\hat{\tau}_0(h) - \tau_0] + [\tau_1 - \tau_0]\,[\hat{g}(h_z) - g] + [\hat{g}(h_z) - g]\,\big([\hat{\tau}_1(h) - \tau_1] - [\hat{\tau}_0(h) - \tau_0]\big).$$

By equations (47) and (48), each bracketed estimation error on the right-hand side is a sum of the conditional bias and the stochastic terms. The conditional biases are $O_p(h^p)$ uniformly over the primitives that satisfy Assumption 3 by the argument in the proof of Theorem B.4 in Armstrong and Kolesár (2020), which applies under random design by Lemmas E.2 and E.4. The stochastic parts are $O_p((nh)^{-1/2})$ because the weights are bounded in probability by Lemma E.4 and vanish

outside the bounded support of the kernel, and the conditional variances of e_i and r_i are bounded by Assumption 5 and $|r_i| \leq 1$. Under Assumption 1, $h^p \asymp (nh)^{-1/2}$, so the last term involving products is $O_p((nh)^{-1}) = o_p((nh)^{-1/2})$. □

Subtract Leading Bias. By the definition of LB in Proposition 3.1, $LB = g\,B_1 + (1-g)\,B_0 + [\tau_1 - \tau_0]\,B_g$, so combining Lemma G.1 with equations (47) and (48) gives

$$\hat{\tau}_S(h, h_z) - \tau - LB = g\,S_1 + (1-g)\,S_0 + [\tau_1 - \tau_0]\,S_g + o_p((nh)^{-1/2}). \tag{49}$$

Replace stochastic local polynomial weights with limiting weights. Adding and subtracting the limiting weights $\tilde{k}_j$ and $\tilde{k}^z$ in S_1, S_0 and S_g, and using $h_z = \kappa h$, equation (49) becomes

$$\sqrt{nh}(\hat{\tau}_S(h, h_z) - \tau - LB) = \sum_{i=1}^{n} Z_{n,i} + \sqrt{nh}\,R + o_p(1),$$

where $R := g\,R_1 + (1-g)\,R_0 + [\tau_1 - \tau_0]\,R_g$ with $R_j := \frac{1}{nh}\sum_{i=1}^{n}[\tilde{k}_{j,n}(X_i/h, Z_i) - \tilde{k}_j(X_i/h, Z_i)]e_i$ and $R_g := \frac{1}{nh_z}\sum_{i=1}^{n}[\tilde{k}^z_n(X_i/h_z) - \tilde{k}^z(X_i/h_z)]r_i$ and

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

The goal is to show $\sqrt{nh}\,R = o_p(1)$. I show the result for the treatment side of R_j. The control side and R_g follow in the same manner, using Lemma E.3(iii) and $|r_i| \leq 1$ for R_g. Since $\tilde{Q}_{j,n,+}$ and $\tilde{Q}_{j,+}$ do not depend on i, they can be moved in front of the sum over i,

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

The first factor converges to zero in probability by Lemma E.3(ii). The second factor has mean zero and a bounded second moment, because the kernel is bounded with bounded support and σ_L^2 is bounded by Assumption 5, so it is $O_p(1)$ by Chebyshev's inequality. It therefore suffices to determine the limiting distribution of $\sum_{i=1}^{n} Z_{n,i}$.

Lindeberg conditions. The $Z_{n,i}$ form a triangular array, because the bandwidth changes with the sample size n, and for each n they are i.i.d. To apply the Lindeberg central limit theorem to $\sum_{i=1}^{n} Z_{n,i}$, I verify the three Lindeberg conditions,

(i) $\mathbb{E}[Z_{n,i}] = 0$,

(ii) $\sum_{i=1}^{n}\mathbb{E}[Z_{n,i}^2] \to \mathcal{V}_S$,

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

(i) By the law of iterated expectations,

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

(ii) Expanding the square gives

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

The last three terms in the second line are zero due to the following reasons. Note that by definition of the limiting weights on the treated side

$$\tilde{k}_{0,+}(u,z)\,\tilde{k}_{1,+}(u,z) = e_1^\top(\tilde{Q}_{0,+})^{-1}m_q(u)\,k_+(u)\;e_1^\top(\tilde{Q}_{1,+})^{-1}m_q(u)\,k_+(u)\;\underbrace{\mathbb{1}_{\{z=0\}}\,\mathbb{1}_{\{z=1\}}}_{=0} = 0$$

due to the mutually exclusive dummies in the regression weights. The remaining terms, which involve the control side weights, are zero for the same reason. Further, an application of the law of iterated expectations shows that $\mathbb{E}[s(X_i, Z_i)\,e_i\,r_i] = \mathbb{E}\big[s(X_i, Z_i)\,r_i\;\mathbb{E}[e_i \,|\, X_i, Z_i]\big] = 0$ for any function $s(X_i, Z_i)$. Finally, the last line inserts $h = \frac{1}{\kappa}h_z$.

Summing over i and using the law of iterated expectations gives

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{nh}\sum_{i=1}^{n}[\tilde{k}_j(X_i/h, Z_i)\, e_i]^2\Big] &= \mathbb{E}\Big[\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_j(X_i/h, Z_i)^2\, \sigma_L^2(X_i, Z_i)\Big] \\ &= \mathbb{E}\Big[\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_{j,+}(X_i/h, Z_i)^2\, \sigma_L^2(X_i, Z_i)\Big] \\ &+ \mathbb{E}\Big[\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_{j,-}(X_i/h, Z_i)^2\, \sigma_L^2(X_i, Z_i)\Big].\end{aligned}$$

The second line follows by plugging in $\tilde{k}_j(X_i/h, Z_i) = \tilde{k}_{j,+}(X_i/h, Z_i) - \tilde{k}_{j,-}(X_i/h, Z_i)$ and using the fact that the weights of the estimators on the left and right side of the cutoff have disjoint supports and hence $\tilde{k}_{j,+}(X_i/h, Z_i)\, \tilde{k}_{j,-}(X_i/h, Z_i) = 0$.

Since $\tilde{k}_{j,+}(u, j) = k_q^*(u)/f_{X,Z}(0, j)$ for $u \geq 0$, a change of variables and dominated convergence give

$$\mathbb{E}\Big[\frac{1}{nh}\sum_{i=1}^{n}\tilde{k}_{j,+}(X_i/h, Z_i)^2\, \sigma_L^2(X_i, Z_i)\Big] \to \frac{\sigma_{L,+}^2(0, j)\, \overline{k_q^2}(k)}{f_{X,Z}(0, j)}.$$

Dominated convergence applies because the integrand is bounded with bounded support, and the limit uses the one-sided limits of σ_L^2 in Assumption 5. The term involving $\tilde{k}_{j,-}$ follows likewise with $\sigma_{L,-}^2(0, j)$. Since $f_{X,Z}(0, 1) = g\, f_X(0)$ and $f_{X,Z}(0, 0) = (1 - g)\, f_X(0)$, the two group terms sum to $\mathcal{V}_O$ by the definition of $\sigma_L^2(x)$. The weight term similarly converges to $\mathcal{V}_g$, using the continuity of σ_Z^2 from Lemma D.1. Hence $\sum_{i=1}^{n}\mathbb{E}[Z_{n,i}^2] \to \mathcal{V}_O + [\tau_1 - \tau_0]^2\, \kappa^{-1}\, \mathcal{V}_g = \mathcal{V}_S$.

(iii) Let $\eta > 0$ be the constant from Assumption 5. Note that $\mathbb{1}_{\{|Z_{n,i}|>\epsilon\}} = \mathbb{1}_{\{|Z_{n,i}|^\eta>\epsilon^\eta\}} \leq \frac{|Z_{n,i}|^\eta}{\epsilon^\eta}$ gives

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

Bound $\sum_{i=1}^{n}\mathbb{E}[|Z_{n,i}|^{2+\eta}]$ using the c_r inequality in Lemma J.1, with $r = 2 + \eta$. For any single $Z_{n,i}$ only two summands are nonzero, the term of its own group and the weight term, because the group indicators in the equivalent kernels are mutually exclusive, so the bound holds with the constant $2^{(1+\eta)}$:

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

The above line follows after applying the fact that $|a \cdot b|^p = |a|^p \cdot |b|^p$ and Assumption 5 together with the law of iterated expectations. Note that r_i is automatically bounded since Z_i is binary. Without loss of generality I assume $\eta \leq 1$. If Assumption 5 holds for some $\eta > 1$, then for any $\eta' \in (0, 1]$ conditional Jensen's inequality gives $\mathbb{E}[|e_i|^{2+\eta'}|X_i, Z_i] \leq (\mathbb{E}[|e_i|^{2+\eta}|X_i, Z_i])^{\frac{2+\eta'}{2+\eta}} \leq (1/\eta)^{\frac{2+\eta'}{2+\eta}} < 1 \leq 1/\eta'$, so the assumption also holds with η' in place of η. Since $\eta \leq 1$, the third term satisfies $|r_i|^{2+\eta} \leq 1 \leq 1/\eta$, so the factor $1/\eta$ applies to all three terms:

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

And likewise for the other terms.

Since k is bounded with bounded support, a change of variables and the continuity of the densities at the cutoff give

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

The third line follows because $|\tilde{k}_j(u,j)| = \frac{1}{f_{X,Z}(0,j)}|k_q^*(|u|)|$ and hence $|g\,\tilde{k}_1(u,1)| = |(1-g)\,\tilde{k}_0(u,0)| = \frac{|k_q^*(|u|)|}{f_X(0)}$. The fifth line follows because $g + (1-g) = 1$.

Putting everything together one can conclude that,

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

It follows that the Lindeberg conditions hold, $L_n(\epsilon) \to 0$ for any $\epsilon > 0$, since $nh \to \infty$ implies $\frac{1}{(nh)^{\eta/2}} \to 0$ for any $\eta > 0$, in particular for the fixed η from Assumption 5, which allows the application of the CLT.

By the Lindeberg central limit theorem, $\sum_{i=1}^{n} Z_{n,i} \xrightarrow{d} N(0, \mathcal{V}_S)$, and the claim follows after replacing the empirical weights with the limiting weights and applying Slutsky's lemma.

G.II Worst-Case Leading Bias of Stratified RD

Applying the same integral representation as in the proof of Proposition 2.1 to the conditional biases in $LB = g\,B_1 + (1-g)\,B_0 + [\tau_1 - \tau_0]\,B_g$ gives

$$
\begin{aligned}
B_j &= \int_0^\infty \bar{w}_j(s)\,[m^{(p)}(s,j) + \tau^{(p)}(s,j)]\,ds - \int_{-\infty}^0 \bar{w}_j(s)\,m^{(p)}(s,j)\,ds, \\
B_g &= \int_{-\infty}^\infty \bar{w}^z(s)\,g^{(p)}(s)\,ds,
\end{aligned}
$$

where $\bar{w}_j$ is defined analogously to $\bar{w}$ in the proof of Proposition 2.1, but using the within-group weights of group j and $\bar{w}^z$ using the weights of the weight estimator. By conditions (i)–(iii) of

Assumption 3 and the triangle inequality,

$$|LB| \leq g\,M \int_{-\infty}^{\infty} |\bar{w}_1(s)|\,ds + (1-g)\,M \int_{-\infty}^{\infty} |\bar{w}_0(s)|\,ds + C_{\Delta\tau} L \int_{-\infty}^{\infty} |\bar{w}^z(s)|\,ds.$$

The construction in the proof of Proposition 2.1, with $\Delta m(x) = 0$, $\Delta\tau(x) = C_{\Delta\tau}$ for all x, and $g^{(p)}(s) = L\,\mathrm{sign}(\bar{w}^z(s))$, attains this bound to first order, because the integrated weights of both groups converge to the same function. By the argument in the proof of Theorem B.4 in Armstrong and Kolesár (2020), which applies by Lemmas E.2 and E.4, $\int |\bar{w}_j(s)|\,ds = 2h^p B_{p,q}(k)/p!\,(1+o_p(1))$ for both groups and $\int |\bar{w}^z(s)|\,ds = h_z^p B^{\mathrm{int}}_{p,q}(k)/p!\,(1+o_p(1))$. Hence, with $h_z = \kappa h$,

$$\begin{aligned}
\sup_{(m,\tau,g)\in\mathcal{F}_p^R(M,\mathbf{C},\mathbf{L})} |LB| &= g\,M \int |\bar{w}_1(s)|\,ds + (1-g)\,M \int |\bar{w}_0(s)|\,ds + C_{\Delta\tau} L \int |\bar{w}^z(s)|\,ds + o_p(h^p) \\
&= (g + (1-g))\,\frac{2M}{p!}\,h^p B_{p,q}(k) + \kappa^p\,C_{\Delta\tau}\,\frac{L}{p!}\,h^p B^{\mathrm{int}}_{p,q}(k) + o_p(h^p) \\
&= h^p(b_S + o_p(1)).
\end{aligned}$$

H Proof of Theorem 4.1

Part (i). Take the leading terms of the worst-case bias for the Stratified RD estimator and the lower bound on the worst-case bias of the base RD estimator from Proposition 3.1 and Proposition 2.1. To guarantee $b_S \leq b_B^{\text{low}}$, it suffices to choose $\kappa \leq \left(\frac{B_{p,q}(k)}{B_{p,q}^{\text{int}}(k)}\right)^{\frac{1}{p}} = U_{p,q}(k)$, because with such a choice the following inequality holds:

$$\begin{aligned}
b_S &\leq b_B^{\text{low}} \\
\frac{2M}{p!}B_{p,q}(k) + \kappa^p C_{\Delta\tau}\frac{L}{p!}B_{p,q}^{\text{int}}(k) &\leq [2M + (2C_{\Delta m} + C_{\Delta\tau})L]\frac{1}{p!}B_{p,q}(k) \\
\kappa^p C_{\Delta\tau}\frac{L}{p!}B_{p,q}^{\text{int}}(k) &\leq [(2C_{\Delta m} + C_{\Delta\tau})L]\frac{1}{p!}B_{p,q}(k) \\
\kappa^p C_{\Delta\tau} L\, B_{p,q}^{\text{int}}(k) &\leq C_{\Delta\tau} L\, B_{p,q}(k) + 2C_{\Delta m} L\, B_{p,q}(k) \\
C_{\Delta\tau} L[\kappa^p B_{p,q}^{\text{int}}(k) - B_{p,q}(k)] &\leq 2C_{\Delta m} L\, B_{p,q}(k)
\end{aligned}$$

where the factor κ^p follows from plugging in Assumption 4. Since all constants are greater than or equal to zero, the right-hand side, $2C_{\Delta m}LB_{p,q}(k)$, is weakly larger than zero. To guarantee the inequality I pick κ such that $[\kappa^p B_{p,q}^{\text{int}}(k) - B_{p,q}(k)] \leq 0$, which is guaranteed for any choice $\kappa \leq U_{p,q}(k)$. If $C_{\Delta m}L > 0$, the right-hand side is strictly positive and the inequality is strict.

Part (ii). From Armstrong and Kolesár (2020) one knows that the base RD estimator has asymptotic variance

$$\mathcal{V}_B = \frac{\overline{k_q^2}(k)\left(\sigma_{P,+}^2(0) + \sigma_{P,-}^2(0)\right)}{f_X(0)}.$$

Now by plugging in Lemma C.1 and noting that $\sigma_{P,-}^2(0)$ does not include a term on the difference in treatment effects, I get:

$$\begin{aligned}
\mathcal{V}_B &= \frac{\overline{k_q^2}(k)\left((\Delta m(0)^2 + [\Delta m(0) + \Delta\tau(0)]^2)\sigma_Z^2(0) + \sigma_{L,+}^2(0) + \sigma_{L,-}^2(0)\right)}{f_X(0)} \\
&= \frac{\overline{k_q^2}(k)\left([2\Delta m(0)^2 + 2\Delta m(0)\Delta\tau(0) + \Delta\tau(0)^2]\,\sigma_Z^2(0)\right)}{f_X(0)} + \frac{\overline{k_q^2}(k)\left(\sigma_{L,+}^2(0) + \sigma_{L,-}^2(0)\right)}{f_X(0)} \\
&= \mathcal{V}_O + \frac{\overline{k_q^2}(k)\left(\left[2\left(\Delta m(0) + \tfrac{1}{2}\Delta\tau(0)\right)^2 + \tfrac{1}{2}\Delta\tau(0)^2\right]\sigma_Z^2(0)\right)}{f_X(0)} \\
&= \mathcal{V}_O + \frac{2\left(\Delta m(0) + \tfrac{1}{2}\Delta\tau(0)\right)^2\sigma_Z^2(0)\,\overline{k_q^2}(k) + \tfrac{1}{2}\Delta\tau(0)^2\sigma_Z^2(0)\,\overline{k_q^2}(k)}{f_X(0)}
\end{aligned}$$

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

$$\Delta m(0)^2 + [\Delta m(0) + \Delta\tau(0)]^2 = 2\Delta m(0)^2 + 2\Delta m(0)\Delta\tau(0) + \Delta\tau(0)^2 = 2\left(\Delta m(0) + \tfrac{1}{2}\Delta\tau(0)\right)^2 + \tfrac{1}{2}\Delta\tau(0)^2.$$

On the other hand from Proposition 3.1 and $\sigma_Z^2(0) = g(1-g)$ (Lemma D.1) I get that:

$$\begin{aligned}\mathcal{V}_S &= \mathcal{V}_O + \frac{1}{\kappa}\Delta\tau(0)^2\,\mathcal{V}_g \\ &= \frac{\overline{k_q^2}(k)\left(\sigma_{L,+}^2(0)+\sigma_{L,-}^2(0)\right)}{f_X(0)} + \frac{1}{\kappa}\frac{\widetilde{k_q^2}(k)\,\Delta\tau(0)^2\sigma_Z^2(0)}{f_X(0)} \\ &= \mathcal{V}_O + \frac{1}{\kappa}\frac{\widetilde{k_q^2}(k)\,\Delta\tau(0)^2\sigma_Z^2(0)}{f_X(0)}.\end{aligned}$$

Hence

$$\begin{aligned}\mathcal{V}_B - \mathcal{V}_S &= \underbrace{\frac{2\left(\Delta m(0)+\frac{1}{2}\Delta\tau(0)\right)^2\sigma_Z^2(0)\,\overline{k_q^2}(k)}{f_X(0)}}_{\geq 0} + \frac{\frac{1}{2}\Delta\tau(0)^2\sigma_Z^2(0)\,\overline{k_q^2}(k)}{f_X(0)} - \frac{1}{\kappa}\frac{\widetilde{k_q^2}(k)\,\Delta\tau(0)^2\sigma_Z^2(0)}{f_X(0)} \\ &\geq \Delta\tau(0)^2\sigma_Z^2(0)\frac{1}{f_X(0)}\left(\frac{1}{2}\overline{k_q^2}(k) - \frac{1}{\kappa}\widetilde{k_q^2}(k)\right).\end{aligned}$$

I pick κ such that $\left(\frac{1}{2}\overline{k_q^2}(k) - \frac{1}{\kappa}\widetilde{k_q^2}(k)\right) \geq 0$ to guarantee that $\mathcal{V}_B - \mathcal{V}_S \geq 0$, which holds for $\kappa \geq 2\,\widetilde{k_q^2}(k)/\overline{k_q^2}(k) = L_{p,q}(k)$. If $\Delta m(0) + \Delta\tau(0)/2 \neq 0$, the first term is strictly positive because $\sigma_Z^2(0) = g(1-g) \geq \underline{g}(1-\underline{g}) > 0$, and the inequality is strict.

Part (iii). Taken together the results in parts (i) and (ii) imply part (iii), where the respective bounds form the interval $\kappa \in I_{p,q}(k) = [L_{p,q}(k), U_{p,q}(k)]$. Both terms in equation (28) are increasing in the first-order worst-case bias and in the asymptotic variance, so equation (29) holds, strictly if either inequality is strict.

I Extension to Continuous Covariates

Proofs for the results in Section 6.2 are available from the author upon request.

J Auxiliary Lemmas

Lemma J.1 (c_r Inequality)**.** *Let $a, b \in \mathbb{R}$ and $r \geq 1$, then*

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