Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs

A Comment on Gelman and Imbens (Journal of Business & Economic Statistics, 2019)

Melle R. Albada*

Journal of Comments and Replications in Economics, Volume 4, 2025-11, DOI: 10.18718/81781.50

JEL: C14, C15, C18
Keywords: Local polynomial regression, Treatment effect estimation, Policy analysis, False positives

Data Availability: The R code and the data to reproduce the results of this replication can be downloaded at JCRE's data archive (DOI: 10.15456/j1.2025022.0928758167).

Please Cite As: Albada, M. R. (2025). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. A Comment on Gelman and Imbens (Journal of Business & Economic Statistics, 2019). *Journal of Comments and Replications in Economics*, Vol.4 (2025-11). DOI: 10.18718/81781.50

Abstract

Gelman and Imbens (2019) argue against using global high-order polynomial models in regression discontinuity designs, recommending local linear or quadratic models instead. This comment revisits two of their arguments, showing they are contingent on specific contexts and interpretations. First, global high-order polynomial models assign extreme weights when the running variable's distribution has regions with few observations. Because so few observations receive these weights, their contribution to the treatment estimate is usually small. Second, I establish that local models can also yield excessive false positive findings, even when using best-practice modeling methods, and that this problem worsens as the sample size grows. These results improve our understanding of the limitations of global high-order polynomial models and suggest that researchers should routinely investigate false positive rates in their study.

*Vienna University of Economics and Business, Department of Socioeconomics, Welthandelsplatz 1, 1020, Vienna, melle.albada@wu.ac.at.

Received January 22, 2025; Revised July 21, 2025; Accepted September 12, 2025; Published December 30, 2025.

1 Introduction

Gelman and Imbens (2019) provide one of the most prominent critiques of estimation methods in regression discontinuity (RD) designs. Based on various applications, they present three arguments against global high-order polynomial models: the extreme weights they can give to observations far from the discontinuity, the large variance in estimates across models with a different polynomial order, and the excessive false positive findings associated with their confidence intervals. Local linear and quadratic models appear to be exempt from these issues and are therefore recommended.

These arguments have influenced subsequent methodological and applied work in the RD literature. They are frequently cited by practitioner guides to recommend local models and discourage global models (Cattaneo et al., 2019; Cattaneo and Titiunik, 2022; Melly and Lalive, 2022); by applied researchers to justify their model choice (e.g., Asher and Novosad, 2020; Colonnelli et al., 2020; Gulzar et al., 2022; Cheruvu, 2023); and form the basis for further advances in polynomial selection by Pei et al. (2022), who argue higher-order polynomials can still be useful in certain settings, as long as they are applied locally.

This comment tests the generalizability of two of the arguments in Gelman and Imbens (2019) and investigates their underlying dynamics. It does so by identifying data properties that drive these problems and extending the applications in Gelman and Imbens (2019). The findings demonstrate that, while the examples in Gelman and Imbens (2019) raise valid concerns, their conclusions are contingent on specific contexts and interpretations.

I start by showing that the weights depend on the distribution of the running variable. Global high-order polynomial models assign extreme weights in regions with few observations, typically when the distribution has long tails, contains outliers, or both. This means that if there are observations with extreme weights, they will make up a small fraction of the entire sample. Through cumulative weight plots, I demonstrate that the aggregate contribution of observations with extreme weights is marginal. Paradoxically, observations far from the cutoff can have more influence when many receive moderate weights.

Next, I establish that local models can also have higher than expected false positive rates, even with state-of-the-art bandwidth selection (Calonico et al., 2014, 2020) and confidence interval construction (Calonico et al., 2014). By analyzing subsets of data from Gerber et al. (2008), where no discontinuities are expected, I find instances where 95% confidence intervals from local models falsely reject the null hypothesis more often than 5% of the time. The severity of this issue is related to the number of observations in the sample. In cases where global or local models yield excessive false positive rates, their rates increase as the sample size grows.

The point of revisiting these arguments is not to advocate a return to global high-order polynomial models. Other relevant concerns remain, such as their instability near boundary points (Calonico et al., 2015), or Gelman and Imbens' (2019) last argument that estimates based on different polynomial orders can vary substantially. Rather, this comment makes two contributions to the RD literature.

First, extreme weights are often presented as both a chronic and major problem of global high-

order polynomial models (e.g., Calonico et al., 2015; Cattaneo and Vazquez-Bare, 2017; Cattaneo et al., 2017, 2019). The presented findings challenge those characterizations. Extreme weights occur under specific conditions, and these same conditions ensure that their influence on the treatment estimate is limited. The more fundamental criticism of global high-order polynomial models is not that a few observations far from the cutoff can receive extreme weight, but that any weight is assigned to observations that may be systematically different from those near the cutoff.

Second, the extended false positive rate analysis suggests that applied researchers should not be overly confident in results from local models. They too can falsely produce statistically significant estimates, especially with larger sample sizes. Researchers should therefore routinely investigate the false positive rate in their specific study. A good starting point is to conduct a large number of placebo cutoff tests, as proposed by Eggers et al. (2024), or to replicate the simulation procedure outlined in this comment with a pre-treatment covariate as the outcome (i.e., a variable for which there should be no discontinuities).

2 Revisiting Arguments

2.1 Weights

The weights argument poses that global high-order polynomial models can put a large weight on observations far from the cutoff when estimating the treatment effect. This is undesirable because observations far from the cutoff are likely to systematically differ from those near the cutoff. The problem becomes apparent when writing the treatment estimate as a difference in weighted averages of the outcome variable between the treatment group and the control group.

Let Y_i be the outcome variable, X_i the running variable, and c the cutoff value. Under a continuity assumption, Hahn et al. (2001) show that the average treatment effect at the cutoff is the difference between the conditional expectation functions (CEFs) of the treatment and control groups at the cutoff:

$$\tau = \lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]$$

To estimate these CEFs, polynomial regressions approximate $E[Y_i|X_i = x]$ using a polynomial expansion of degree K, with column vector $P(X_i) = (1, X_i, ..., X_i^K) \in R^{K+1}$. The estimator is evaluated at the cutoff, which the running variable is typically centered at for simplicity ($x = c = 0$):

$$\hat{f}(0) = P(0)' \left(\sum_i P(X_i) P(X_i)' \right)^{-1} \sum_i P(X_i) Y_i$$

This expression can be written as a weighted average of the outcomes:

$$\hat{f}(0) = \sum_i w_i Y_i \quad \text{where} \quad w_i = P(0)' \left(\sum_i P(X_i) P(X_i)' \right)^{-1} P(X_i)$$

The weight for each observation depends on its value for the running variable and the chosen polynomial order.

Gelman and Imbens (2019) inspect the weighting schemes for three published RD data sets (Jacob and Lefgren, 2004; Matsudaira, 2008; Lee, 2008). The weights of global models behave erratically for observations at the distributional end of the running variable when using higher-order polynomials, especially for the Jacob-Lefgren and Matsudaira data sets (see the middle row of Figure 1). In contrast, local models give zero weight to these observations by excluding them through a bandwidth (see Figure 1(b), 2(b), and 3(b) in Gelman and Imbens (2019)).

I add two points to the weights discussion. First, although Gelman and Imbens (2019) relate extreme weights to Runge's phenomenon[1], they do not connect this behavior to characteristics of RD data sets. In practice, the weights depend on the distribution of the running variable. They adjust to fit the function to the available data, amplifying observations in regions that provide less information for estimating the polynomial terms. When parts of the running variable's support contain relatively few observations, the matrix $(\sum_i P(X_i)P(X_i)')^{-1}$ can become unstable and produce large weights. This issue is most pronounced near boundaries and worsens with higher-order polynomials, both implications of Runge's phenomenon. Since the distributional ends of the running variable often have fewer observations, this is where extreme weights tend to occur.

This means that high-order polynomials mainly assign extreme weights when the distribution of the running variable has long tails, contains outliers, or both. Figure 2 illustrates this using histograms and weight plots for two variables. The variable on the left side is uniformly distributed ($X_i \sim U[0,1]$) and on the right side skewed ($X_i \sim \beta(2,8)$). In addition, the variable on the right side includes three outlier values at 0.8, 0.9, and 0.95. The weights, evaluated at 0, are calculated for polynomials up to the sixth order. Whereas the weights of the uniformly distributed variable behave smoothly throughout, the skewed variable displays erratic behavior in the segment with hardly any observations.

The point here is that not all high-order polynomials lead to extreme weights—everything depends on the distribution of the running variable. It just so happens that the Matsudaira and Jacob-Lefgren data sets have an extremely low density at the end of the distribution, including some outliers (see the top row of Figure 1).

Second, while I agree with Gelman and Imbens (2019) that assigning extreme weights to observations far from the cutoff is *prima facie* unattractive, it is not immediately clear how much those observations actually influence the treatment estimate. Standard weight plots show where large weights occur but provide no information about how many observations receive them. To address this limitation, I construct cumulative weight plots (bottom row of Figure 1). These plots sum the weights assigned to observations in order of their distance from the cutoff. By combining both the size of the weights and the number of observations receiving them, cumulative weight plots better indicate which regions contribute most to the treatment estimate.

For the Matsudaira and Jacob-Lefgren data sets, cumulative weight curves converge early, before the running variable reaches the upper end of its support. This flattening implies that the observations far from the cutoff contribute almost nothing, despite the presence of extreme weights in those regions. The explanation is straightforward: there are very few observations far from the

[1] Formally, Runge's phenomenon states that in a bounded interval with N equispaced data points, a polynomial of order $N-1$ produces oscillations in the fit near the boundaries of the interval.

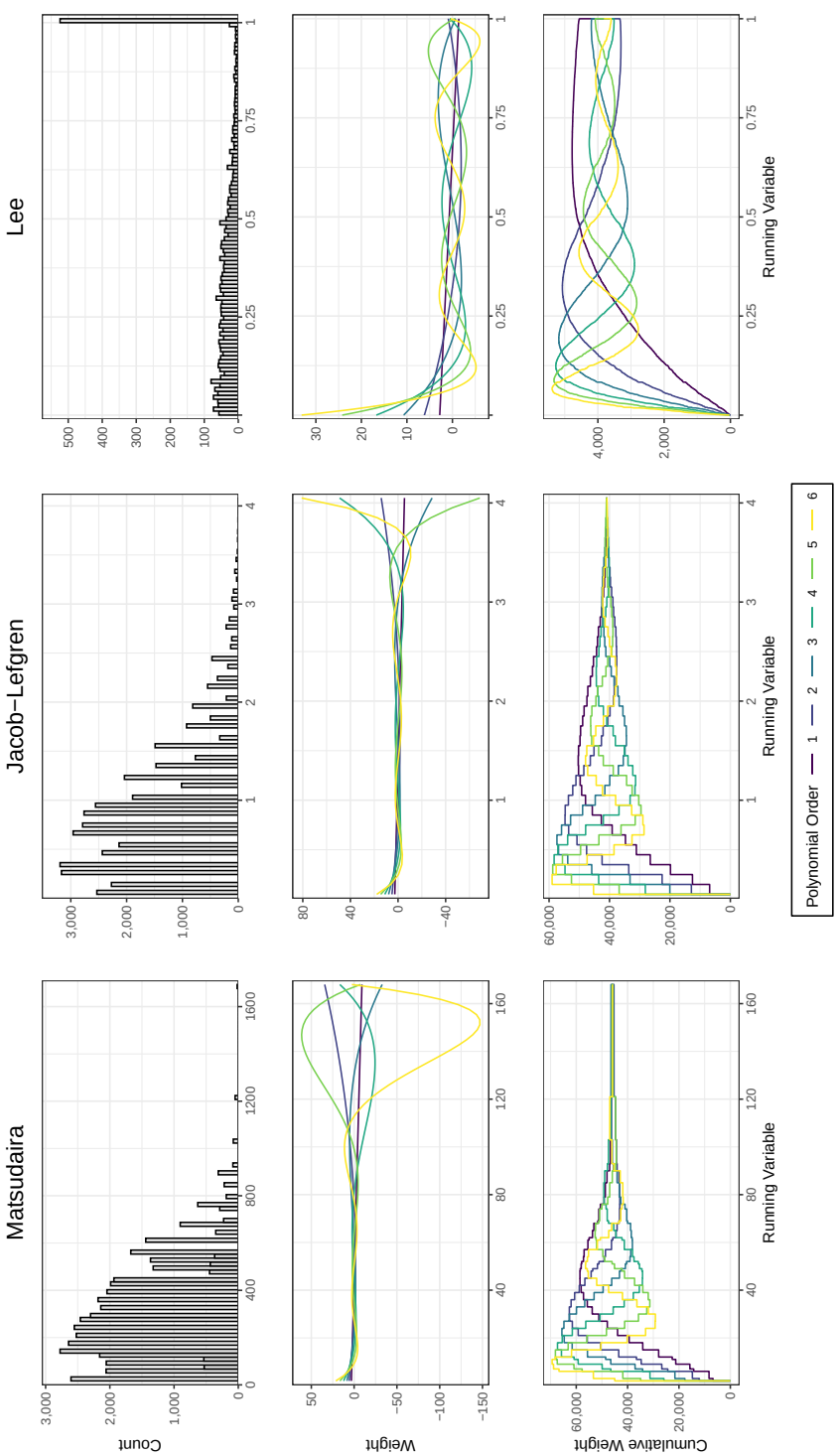

Figure 1: Plots for Matsudaira, Jacob-Lefgren, and Lee data, showing only observations with running variable values above the cutoff (cutoff is re-scaled to 0). Top row: histograms of the running variable. Middle row: standard weight plots for six global polynomial models. Following Gelman and Imbens (2019), weights are normalized to average 1. Bottom row: cumulative weight plots for the same models, accounting for the number of observations that receive each weight. Steep upward or downward slopes indicate regions that contribute strongly to the treatment estimate, while flat sections indicate regions that contribute little.

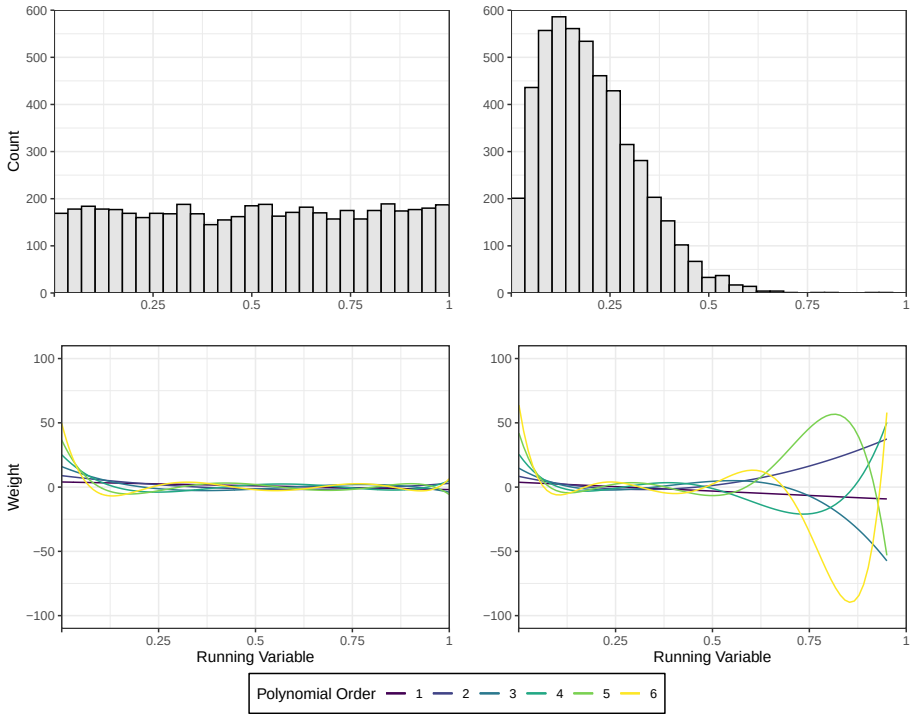

Figure 2: Plots for simulated data. Top row: histograms of a uniformly distributed and right-skewed variable. Bottom row: corresponding standard weight plots for six global polynomial models, evaluated at 0 (as if these were observations on the right side of the cutoff).

cutoff, so their aggregate contribution is limited. For example, in the Matsudaira data, the largest negative weight is –33, which appears extreme relative to an average weight of 1, but is negligible compared to a total absolute weight of 101,981.[2,3]

The Lee data set presents a different pattern. Here, weights are more stable, and cumulative weight curves only converge at the upper end of the support, where the distribution heaps. Because the region far from the cutoff is more densely populated, it continues to influence the treatment estimate. Counterintuitively, the aggregate impact of observations far from the cutoff tends to be greater when weights are well-behaved than when they are extreme. In other words, the cumulative effect of many small weights can exceed that of a few large ones.

These findings put the concern about extreme weights into perspective. Extreme weights typically occur in regions with very few observations, and precisely because so few observations receive these weights, their overall influence on the treatment estimate is limited. In fact, it may be more problematic when global models give moderate weights to many observations far from the cutoff. The more convincing critique of global high-order polynomial models is therefore not that they can assign extreme weights to observations far from the cutoff, but rather that they assign weight to such observations at all.

2.2 False Positives

Another argument that Gelman and Imbens (2019) use against global high-order polynomial models are their excessive false positive rates. In two examples where there should be no discontinuities, the authors find that 95% confidence intervals (CIs) of global models reject the null hypothesis more often than the nominal type 1 error rate (i.e., more than 5%). These models may therefore mislead researchers into thinking their estimate is statistically significant.

In their main example, Gelman and Imbens (2019) modify a data set on yearly earnings used by LaLonde (1986). The data come from the Current Population Survey and contain yearly earnings in thousands of dollars in 1974, 1975, and 1978 for 15,992 individuals. Gelman and Imbens (2019) use the average earnings from 1974 and 1975 to predict the earnings in 1978. This creates an approximately linear relation that should not contain any discontinuities. Consequently, if one takes the average of the earnings in 1974 and 1975 as a running variable, and randomly chooses a value from its distribution as a cutoff, the estimated RD treatment effect on earnings in 1978 should be close to zero.

Gelman and Imbens (2019) exploit this feature to evaluate the false positive rate of global and local models with different polynomial orders in a simulation exercise. They randomly draw 20,000 values from the empirical distribution of the artificial running variable between the 25th and 75th percentile, serving as a cutoff. Next, for each cutoff, they randomly draw 1,000 individuals from the full sample of 15,992 individuals. Given this sample and the randomly chosen cutoff, they then estimate the pseudo RD treatment effect, its standard error, and check whether the implied 95%

[2]The negative weight of -33 is for the third-order polynomial model. The sum of absolute weights for this model is 101,981.

[3]Note that the standard weight plots can be misleading. For the Matsudaira data set, there are no observations between 121 and 168, so the sixth-order polynomial model does not actually assign a negative weight of -150 to any observation.

confidence interval includes zero. Since there is no discontinuity, 95% confidence intervals should include zero 95% of the time (the coverage) and exclude zero in the remaining 5% (the false positive rate).

The analysis considers six global models (first- to sixth-order polynomial) and two local models (local linear and local quadratic). For the local models, Gelman and Imbens (2019) select bandwidths using the MSE-optimal method suggested in Imbens and Kalyanaraman (2012), IK henceforth. I reproduce their analysis with two alterations. First, I extend the analysis to larger sample sizes of 3,000 and 5,000 observations. Second, I include two additional estimation methods for local models.

At the time when Gelman and Imbens (2019) circulated their initial working paper, the IK method was the standard data-driven way to select a bandwidth. Later work by Calonico et al. (2014), CCT henceforth, introduced an improved MSE-optimal bandwidth selector with superior finite and large sample properties that comes closer to the (infeasible) theoretical MSE-optimal bandwidth. In simulations, the CCT bandwidth outperforms the IK bandwidth in terms of coverage (Calonico et al., 2014) and MSE (Arai and Ichimura, 2018).

CCT also identified that conventional confidence intervals for local models, which Gelman and Imbens (2019) use, can be incorrect. The issue lies in the inherent misspecification error of an MSE-optimal RD estimator, which conventional confidence intervals do not account for (Cattaneo and Vazquez-Bare, 2017). To maintain valid inference, CCT proposed robust bias-corrected (RBC) confidence intervals that apply a correction to the point estimate and its standard error. Using the improved MSE-optimal bandwidth with RBC confidence intervals has since become standard practice among applied researchers.

However, the CCT method is not necessarily optimal for inference. Calonico et al. (2020), CCF henceforth, show that a narrower bandwidth must be used to obtain confidence intervals with the smallest coverage error (CE). They provide theoretical and simulation evidence that combining the CE-optimal bandwidth with RBC confidence intervals dominates alternative methods in terms of coverage error, including CCT. This improvement might be particularly relevant when assessing the coverage of RD estimators.

The results for global models and local models based on the IK, CCT, and CCF methods are in Table 1. Panel A confirms Gelman and Imbens' (2019) finding that global models fail to meet the nominal false positive rate of 5%, whereas local models generally come closer. Furthermore, Panels B and C show that the false positive rates of the global models increase as the sample size grows, and that this trend is stronger for models with a lower polynomial order. False positive rates of global models can thus be even more problematic than demonstrated in Gelman and Imbens (2019). The rates of the local models, on the other hand, are stable across sample sizes and polynomial orders.

A relevant question is why global models have excessive false positive rates while local models do not. Given that the IK models achieve nominal coverage, it cannot be due to the RBC confidence intervals or the alternative bandwidth selectors. In their most basic form, local and global models use the same estimation technique, differing only in which observations they include and how they

	Panel A N = 1,000		Panel B N = 3,000		Panel C N = 5,000	
Method	FPR	Median s.e.	FPR	Median s.e.	FPR	Median s.e.
G-1	9.2%	1.07	17.5%	0.62	25.0%	0.48
G-2	8.6%	1.67	12.3%	0.96	15.9%	0.74
G-3	6.7%	2.26	8.8%	1.31	11.6%	1.01
G-4	8.5%	2.82	8.0%	1.65	9.3%	1.28
G-5	7.7%	3.37	6.3%	1.98	7.5%	1.54
G-6	9.4%	3.89	8.1%	2.31	8.3%	1.80
IK-1	7.1%	1.48	6.9%	0.89	7.1%	0.70
IK-2	6.4%	2.08	5.6%	1.22	5.7%	0.96
CCT-1	7.2%	2.99	6.4%	1.75	6.3%	1.36
CCT-2	7.5%	3.72	6.4%	2.18	6.4%	1.70
CCF-1	7.3%	3.24	5.8%	1.93	5.7%	1.52
CCF-2	7.5%	4.13	5.9%	2.48	5.9%	1.96

Table 1: Lalonde data set: False positive rates for 95% confidence intervals under a null hypothesis of no discontinuity. 'IK' uses conventional confidence intervals based on the MSE-optimal bandwidth from Imbens and Kalyanaraman (2012), 'CCT' uses robust confidence intervals based on the MSE-optimal bandwidth from Calonico et al. (2014), 'CCF' uses robust confidence intervals based on the CE-optimal bandwidth from Calonico et al. (2020), and 'G' uses HC0 confidence intervals for global models. The number indicates the polynomial order. Panel A reproduces the results from Gelman and Imbens (2019), while Panels B and C repeat the analysis with 3,000 and 5,000 observations, respectively.

weight them. This suggests that the poor performance of global models stems from the weight they assign to observations outside of the bandwidth.

To investigate this, Figure 3 presents the distribution of the running variable in the Lalonde data set alongside its corresponding weights, evaluated at a placebo cutoff of 14.65 (the median of the average of earnings in 1974 and 1975, used as an example by Gelman and Imbens (2019)). The distribution has strong heaping at the tails, rather than long tails or outliers. This rules out extreme weights as an explanation, as verified by the standard weight plots.

However, the cumulative weight plots show that observations far from the cutoff do contribute substantially to the treatment estimates. These regions may contain relevant data features that global models fail to capture. Indeed, when I repeat the analysis after removing the observations with average earnings of zero (i.e., most of the left heap in the distribution), the global models' false positive rates generally return to nominal levels (Appendix Table A1). This underlines that the disadvantage of global models lies in including observations far from the cutoff, not in assigning them extreme weights per se.

Still, there is no a priori reason to assume that problematic data structures always fall outside the bandwidth. Extreme values, heaping, or other irregularities may occur anywhere in the support of the running variable. Although Gelman and Imbens (2019) provide two examples in which local models have reliable false positive rates, this is by no means evidence of a universal result.

To assess the generalizability of these examples, I repeat the analysis with an additional data set. Utilizing a subset of the data from a randomized experiment on voter turnout by Gerber et al. (2008), previously used in RD model assessments (Green et al., 2009; Gelman and Zelizer, 2015), I examine the relation between age and voter turnout. This data set includes 30,038 target voters, who were randomly assigned to a control group (N = 24,964) and a treatment group (N = 5,074). For both groups separately, I regress age on the election day on a voting dummy. As the randomized treatment assignment ensures independence from age, and I only consider observations from either the treatment or the control group, there should not be any discontinuous treatment effect. Following the previous analysis, I draw 20,000 random cutoff values between the 25th and 75th percentiles of the age variable, draw a random sample for each cutoff, and estimate the pseudo RD treatment effect. The analysis uses sample sizes of 1,000 and 3,000 observations, omitting the sample size of 5,000 because the treatment group only contains 5,074 observations.

For the treatment group, both local and global models suffer from excessive false positive rates, and their performance worsens as the number of observations increases (Table 2, Panel A and B). In fact, some of the local models produce substantially larger false positive rates than the global models (e.g., IK-1 compared to G-6). While CCF-2 performs noticeably better than the other local

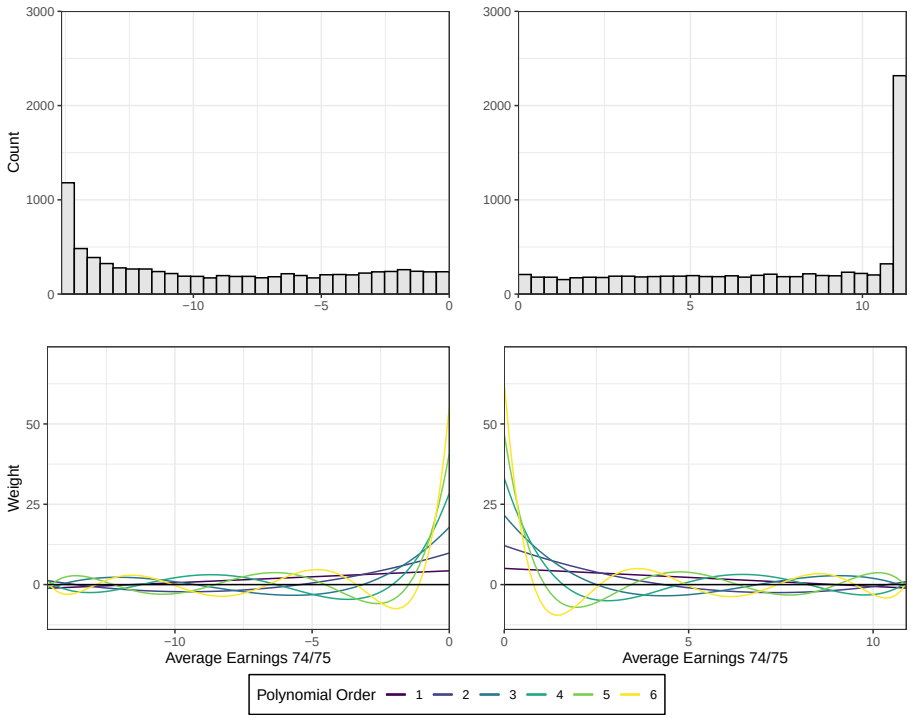

Figure 3: Plots for Lalonde data. The running variable is centered at the median of the average earnings in 1974 and 1975 (14.65, re-scaled to 0). Top row: histograms of negative and positive values of the running variable. Middle row: corresponding standard weight plots for six global polynomial models. Bottom row: cumulative weight plots for the same models, accounting for the number of observations that receive each weight.

	Panel A Treatment N = 1,000		Panel B Treatment N = 3,000		Panel C Control N = 1,000		Panel D Control N = 3,000	
Method	FPR	Median s.e.	FPR	Median s.e.	FPR	Median s.e.	FPR	Median s.e.
G-1	38.5%	0.055	77.5%	0.032	24.5%	0.053	59.0%	0.030
G-2	9.2%	0.078	16.0%	0.045	5.1%	0.074	4.5%	0.043
G-3	8.8%	0.100	15.0%	0.058	5.6%	0.096	5.8%	0.055
G-4	8.6%	0.123	15.2%	0.070	5.5%	0.117	4.6%	0.067
G-5	8.4%	0.146	13.8%	0.083	5.6%	0.139	5.1%	0.079
G-6	7.9%	0.168	11.1%	0.096	5.7%	0.161	5.0%	0.092
IK-1	10.5%	0.093	16.9%	0.056	6.1%	0.087	5.4%	0.051
IK-2	9.2%	0.125	14.3%	0.074	5.7%	0.117	5.0%	0.068
CCT-1	8.7%	0.129	13.9%	0.075	5.6%	0.122	4.9%	0.070
CCT-2	8.7%	0.158	13.8%	0.091	6.0%	0.150	5.1%	0.086
CCF-1	8.1%	0.140	13.3%	0.083	5.4%	0.132	4.6%	0.077
CCF-2	7.9%	0.176	11.5%	0.104	5.7%	0.167	4.8%	0.098

Table 2: Gerber et al. data set: False positive rates for 95% confidence intervals under a null hypothesis of no discontinuity. 'IK' uses conventional confidence intervals based on the MSE-optimal bandwidth from Imbens and Kalyanaraman (2012), 'CCT' uses robust confidence intervals based on the MSE-optimal bandwidth from Calonico et al. (2014), 'CCF' uses robust confidence intervals based on the CE-optimal bandwidth from Calonico et al. (2020), and 'G' uses HC0 confidence intervals for global models. The number indicates the polynomial order. Panels A and B only use observations from the treatment group, whereas Panels C and D only use observations from the control group.

models, it still maintains a false positive rate of 11.5% with sample sizes of 3,000 observations.[4]

Turning to the results for the control group (Table 2, Panel C and D), only G-1 has excessive false positive rates. All other models have false positive rates around 5%, for both of the sample sizes, suggesting global models do not necessarily suffer from this problem. The poor performance of the global linear model can be explained by systematic approximation bias. Since there is no reason to expect a fully linear relation between age and voting behavior across the entire support of the data, G-1 is likely an underspecified model.

[4]One possible explanation for the excessive false positive rates rates of local linear and quadratic models is that the polynomial order is held fixed across simulation rounds. In some settings, a linear model may better approximate the regression function, while in others, a nonlinear model may perform better. To assess this possibility, I apply the procedure proposed by Pei et al. (2022), which selects both the bandwidth and the polynomial order by minimizing the MSE. This approach is a generalization of the CCT method that allows the data to determine the preferred polynomial order. However, the local linear model is selected in nearly every simulation round. This suggests that the excessive false positive rates cannot be attributed to holding the polynomial order fixed, and that further model selection offers no substantial improvement in this setting.

3 Discussion

This comment adds nuance to two of the arguments in Gelman and Imbens (2019). First, global high-order polynomial models typically assign extreme weights when the distribution of the running variable contains long tails or outliers. Precisely because so few observations receive these weights, they contribute little to the estimated treatment effect. Second, excessive false positive rates are not exclusive to global models. Confidence intervals from local models, even when constructed following best practices, can also fail to include zero more often than expected. For both local and global models, this problem worsens in larger samples.

Like Gelman and Imbens (2019), the findings in this paper are based on specific examples rather than formal proofs. I therefore do not claim that local models are generally invalid or that they suffer from excessive false positive rates across all settings. On the contrary, local models remain the theoretically preferred approach in RD designs, supported by extensive methodological work in estimation and inference (see Cattaneo and Titiunik (2022) for an overview), which has been empirically validated in the context of close elections (Hyytinen et al., 2018; De Magalhães et al., 2025). However, just as Gelman and Imbens (2019) provide valuable caution against global models based on specific examples, the false positive rate analysis here suggests that local models, too, can mislead researchers into thinking their estimate is statistically significant. The broader point is that researchers should not automatically assume local models are immune to performance issues.

For applied researchers, I reiterate the latest methodological recommendations for local models. The most conservative approach separates point estimation from inference (Cattaneo and Vazquez-Bare, 2017; Calonico et al., 2020). Point estimates should be based on MSE-optimal choices, selecting the bandwidth and possibly the polynomial order that minimize the MSE, as proposed by Pei et al. (2022). Inference should be conducted with the CE-optimal bandwidth and robust bias-corrected confidence intervals. This approach reduces the risk of excessive false positive rates, though it cannot eliminate them entirely.

Consequently, researchers should investigate how often their chosen estimation method yields significant effects when none are expected. One option is to conduct many placebo cutoff tests, estimating the treatment effect at alternative cutoffs. The share of significant estimates approximates the method's false positive rate (Eggers et al., 2024). For this approach to be valid, there must be no interventions at other values of the running variable. Alternatively, researchers can conduct simulation exercises using pre-treatment variables, similar to the procedure outlined in this comment. The idea is to draw random cutoff values from the running variable's distribution, replace the outcome with a pre-treatment variable, and estimate the treatment effect at these cutoffs. Both approaches differ from conventional placebo tests, which often use just a few placebo cutoffs or test pre-treatment variables only once at the true cutoff for balance.

Bibliography

Arai, Yoichi and Hidehiko Ichimura. 2018. "Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator". *Quantitative Economics* 9 (1): 441–482. https://doi.org/10.3982/QE590.

Asher, Sam and Paul Novosad. 2020. "Rural roads and local economic development". *American Economic Review* 110 (3): 797–823. https://doi.org/10.1257/aer.20180268.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2020. "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs". *The Econometrics Journal* 23 (2): 192–210. https://doi.org/10.1093/ectj/utz022.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs". *Econometrica* 82 (3): 2295–2326. https://doi.org/10.3982/ECTA11757.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2015. "Optimal data-driven regression discontinuity plots". *Journal of the American Statistical Association* 110 (512): 1753–1769. https://doi.org/10.1080/01621459.2015.1017578.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.

Cattaneo, Matias D. and Rocío Titiunik. 2022. "10.1257/aer.20180268". *Annual Review of Economics* 14 : 821–851. https://doi.org/10.1146/annurev-economics-051520-021409.

Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare. 2017. "Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality". *Journal of Policy Analysis and Management* 36 (3): 643–681. https://doi.org/10.1002/pam.21985.

Cattaneo, Matias D. and Gonzalo Vazquez-Bare. 2017. "The choice of neighborhood in regression discontinuity designs". *Observational Studies* 3 (2): 134–146. https://doi.org/10.1353/obs.2017.0002.

Cheruvu, Sivaram. 2023. "Education, public support for institutions, and the separation of powers". *Political Science Research and Methods* 11 (3): 570–587. https://doi.org/10.1017/psrm.2022.29.

Colonnelli, Emanuele, Mounu Prem, and Edoardo Teso. 2020. "Patronage and selection in public sector organizations". *American Economic Review* 110 (10): 3071–3099. https://doi.org/10.1257/aer.20181491.

De Magalhães, Leandro, Dominik Hangartner, Salomo Hirvonen, Jaakko Meriläinen, Nelson A. Ruiz, and Janne Tukiainen. 2025. "When can we trust regression discontinuity design estimates from close elections? evidence from experimental benchmarks". *Political Analysis* 33 (3): 258–265. https://doi.org/10.1017/pan.2024.28.

Eggers, Andrew C., Guadalupe Tuñón, and Allan Dafoe. 2024. "Placebo tests for causal inference". *American Journal of Political Science* 68 : 1106–1121. https://doi.org/10.1111/ajps.12818.

Gelman, Andrew and Guido Imbens. 2019. "Why high-order polynomials should not be used in regression discontinuity designs". *Journal of Business & Economic Statistics* 37 (3): 447–456. https://doi.org/10.1080/07350015.2017.1366909.

Gelman, Andrew and Adam Zelizer. 2015. "Evidence on the deleterious impact of sustained use of polynomial regression on causal inference". *Research & Politics* 2 (1): 1–7. https://doi.org/10.1177/2053168015569830.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment". *American Political Science Review* 102 (1): 33–48. https://doi.org/10.1017/S000305540808009X.

Green, Donald P., Terence Y. Leong, Holger L. Kern, Alan S. Gerber, and Christopher W. Larimer. 2009. "Testing the accuracy of regression discontinuity analysis using experimental benchmarks". *Political Analysis* 17 (4): 400–417. https://doi.org/10.1093/pan/mpp018.

Gulzar, Saad, Miguel R. Rueda, and Nelson A. Ruiz. 2022. "Do campaign contribution limits curb the influence of money in politics?". *American Journal of Political Science* 66 (4): 932–946. https://doi.org/10.1111/ajps.12596.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and estimation of treatment effects with a regression-discontinuity design". *Econometrica* 69 (1): 201–209. https://doi.org/10.1111/1468-0262.00183.

Hyytinen, Ari, Jaakko Meriläinen, Tuukka Saarimaa, Otto Toivanen, and Janne Tukiainen. 2018. "When does regression discontinuity design work? evidence from random election outcomes". *Quantitative Economics* 9 (2): 1019–1051. https://doi.org/10.3982/QE864.

Imbens, Guido and Karthik Kalyanaraman. 2012. "Optimal bandwidth choice for the regression discontinuity estimator". *Review of Economic Studies* 79 : 933–959. https://doi.org/10.1093/restud/rdr043.

Jacob, Brian A. and Lars Lefgren. 2004. "Remedial education and student achievement: A regression-discontinuity analysis". *Review of Economics and Statistics* 86 (1): 226–244. https://www.jstor.org/stable/3211669.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data". *American Economic Review* 76 (4): 604–620. https://www.jstor.org/stable/1806062.

Lee, David S. 2008. "Randomized experiments from non-random selection in u.s. house elections". *Journal of Econometrics* 142 (2): 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004.

Matsudaira, Jordan D. 2008. "Mandatory summer school and student achievement". *Journal of Econometrics* 142 (2): 829–850. https://doi.org/10.1016/j.jeconom.2007.05.015.

Melly, Blaise and Rafael Lalive. 2022. "Regression discontinuity design". In K. F. Zimmerman (Ed.), *Handbook of Labor, Human Resources and Population Economics*, pp. 1–32. Springer.

Pei, Zhuan, David S. Lee, David Card, and Andrea Weber. 2022. "Local polynomial order in regression discontinuity designs". *Journal of Business & Economic Statistics* 40 (3): 1–9. https://doi.org/10.1080/07350015.2021.1920961.

Appendix

	Panel A N = 1,000		Panel B N = 3,000		Panel C N = 5,000	
Method	FPR	Median s.e.	FPR	Median s.e.	FPR	Median s.e.
G-1	8.3%	1.04	13.0%	0.60	18.0%	0.47
G-2	6.4%	1.61	6.7%	0.93	7.6%	0.72
G-3	5.8%	2.16	5.6%	1.25	6.4%	0.97
G-4	7.2%	2.70	6.5%	1.57	6.8%	1.22
G-5	6.9%	3.21	5.9%	1.89	5.9%	1.47
G-6	8.6%	3.71	7.2%	2.20	6.8%	1.71
IK-1	6.7%	1.43	6.8%	0.84	7.7%	0.66
IK-2	6.3%	1.99	5.2%	1.16	5.7%	0.90
CCT-1	6.7%	2.81	6.1%	1.65	6.2%	1.28
CCT-2	7.0%	3.49	6.1%	2.06	6.0%	1.61
CCF-1	6.5%	3.03	5.6%	1.82	5.3%	1.42
CCF-2	6.8%	3.88	5.9%	2.34	5.4%	1.85

Table A1: Lalonde data set after excluding observations with average earnings of zero: False positive rates for 95% confidence intervals under a null hypothesis of no discontinuity. 'IK' uses conventional confidence intervals based on the MSE-optimal bandwidth from Imbens and Kalyanaraman (2012), 'CCT' uses robust confidence intervals based on the MSE-optimal bandwidth from Calonico et al. (2014), 'CCF' uses robust confidence intervals based on the CE-optimal bandwidth from Calonico et al. (2020), and 'G' uses HC0 confidence intervals for global models. The number indicates the polynomial order. Panels A, B, and C vary the analysis with 1,000, 3,000, and 5,000 observations, respectively.