Optimal Regression Discontinuity Asymmetric Bandwidth Selector with Local Polynomial Progression: with an Application to U.S. Congressional Election Data

Jiawei Yang[*]
University of California, San Diego

December 2022

Abstract

I investigate the choice for the optimal local bandwidth in sharp regression discontinuity designs. I propose the adoption of local polynomial progression instead of local linear regression for its parsimonious capture of both linear and non-linear data-generating processes. The computation of the bandwidth results from comparing the combination of the order of regression polynomials and bandwidth. Specifically, I propose the objective function that penalizes the length with bandwidth, fit of the polynomial, and the order. This strategy differs from the traditional ones to allow asymmetric optimization. Applying this bandwidth selector to Lee (2008) data to study the incumbency advantage in the U.S. congressional election, I show local polynomial regression identifies narrower asymmetric bandwidths to yield regression discontinuity causal estimates of similar effect size but smaller estimated variance.

1 Introduction

Regression discontinuity design (RDD) was first introduced for causal inference by Thistlewaite and Campbell (1960) to evaluate the effect on academic achievement of a merit-based scholarship whose eligibility was determined by a threshold in exam scores. Their context of RDD was later categorized as a sharp regression discontinuity (SRD). When Hahn et al. (2001) formally imported the concept to economics by formulating its estimation of treatment effects with the potential outcome framework, it was increasingly used in empirical studies. It has long been recognized as one of the family of quasi-experimental tools; however, Lee and Lemieux (2010) proved it more reliable than the others to be "as good as" the golden standard, randomized controlled trials (RCT).

In an SRD setting, the automatic failing of the overlapping assumption pins down the key distinction between the SRD and the RCT causal estimates: SRD's

[*] Deep thanks to Professor Emanuel Vespa and Simone Galperti for organizing the thesis seminar and guiding me through the process and to Professor Yixiao Sun and Xinwei Ma for carefully discussing my econometric research proposals.

ability to house almost perfect randomization just around the threshold, on the other hand, imposes the limitation to only estimate a local treatment effect specific to hypothetical individuals right at the threshold. Similar conditions exist for other types of RDD's such as fuzzy regression discontinuity designs.

In practice, researchers set a bandwidth near the treatment threshold and only model the expected outcomes for individuals within the bandwidth to ensure locality and therefore goodness of randomization. Moreover, local linear regression is most commonly used to capture the expected outcomes within the local window. There had been little consensus on the choice of bandwidth that led to *ad-hoc* decisions, though later literature attempted to address the problem. The caveat should be clear that totally discretionary choice of the bandwidth leaves room for repetitive trials in search for a bandwidth where an outcome of favored effect size or significance could be obtained.

In this paper, I propose a data-driven and side-specific selector of neighborhood window sizes, a key set of hyper-parameter that governs the goodness in any local methods and therefore influences the robustness of the estimates they produce. I specifically focus on the SRD setup and construct the selector with the parametric approach of local polynomial regressions; the selector indicates the optimal neighborhood window sizes separately for the treated and untreated subjects, with respect to my proposed objective function that jointly measures locality, the goodness of fit of polynomial regressions, and the order of the polynomial; optimization is achieved via iterative computation algorithms. For a given order, the naive objective function indicates the smallest possible neighborhood window size that allows the polynomial regression to reasonably capture the pattern in conditional expectation of the outcome. The objective function is further modified to enable comparing polynomial regressions of different orders. The modified version of the strategy provides guidance on both the order of appropriate polynomial approximations and the neighborhood window size optimal in the error criterion.

I start by revisiting inference with SRD and current approaches of bandwidth selection in section 2. In section 3, I reveal the specifics of polynomial progression in terms of its goal, intuition, mathematics regarding its constrained optimization, and computational solutions. Section 4 illustrates the working of my proposed selector with a Monte-Carlo simulated example. In section 5, I apply the selector and follow up with SRD analysis of the congressional election data initially investigated in Lee (2008). In section 6, I talk about potential issues of polynomial progression and suggest possible remedies to those.

2 Related Literature on SRD and Selectors

2.1 Setup: SRD and Properties

In this paper, I focus on the specific scenario of sharp regression discontinuity designs with one forcing variable. This section reviews its basics of causal inference and specifies the notation used thereafter in line with Rubin's potential outcome language.

Consider a sample indexed by unit i, for $i = 1, ..., N$, for each individual we observe a scalar baseline characteristic X_i and the continuous outcome of interest Y_i. Let the treatment status be denoted by the binary D_i for $D_i = 1$

as treated and $D_i = 0$ as untreated, we then have the potential outcomes:

$$Y_i = Y(D_i)$$

In the SRD setting, the treatment assignment is completely determined by a threshold in one of the scalar baseline covariates. In this case, let X_i be that variable. Such variables are often referred to as the forcing or the running variable. Without loss of generality, with a threshold of $X = c$, let treatment be given to individuals whose $X_i \geq c$ and not assigned to those whose $X_i \leq c$. This can be denoted as:

$$D_i = \mathbb{1}_{X_i \geq c}.$$

To estimate the causal effect of the treatment D in a standard RCT, the average treatment effect is computed as:

$$\begin{aligned}\tau &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] \\ &= \mathbb{E}\left[Y_i(1) \middle| D_i = 1\right] - \mathbb{E}\left[Y_i(0) \middle| D_i = 0\right]\end{aligned}$$

given statistical independence between the treatment assignment and outcome that $\mathbb{E}\left[Y_i(d)\right] = \mathbb{E}\left[Y_i(d) \middle| D_i = d\right]$ for $d \in \{0, 1\}$.

Inference with SRD's, or RDD's in general, faces more restrictive conditions and so differentiates from that with RCT's. For one, the statistical independence only holds at or within an arbitrarily small window around the threshold c of the forcing variable X where one believes the subjects have no control over on which side of the threshold she falls. Therefore, for subjects infinitely close to the threshold, their receiving the treatment is considered completely randomized to be "as good as" an RCT. Therefore, the SRD local average treatment effect could be computed as:

$$\begin{aligned}\tau_{\text{SRD}} &= \mathbb{E}\left[(Y_i(1)|D_i = 1) - (Y_i(0)|D_i = 0)|X_i = c\right] \\ &= \mathbb{E}\left[Y_i(1) - Y_i(0)|X_i = c\right]\end{aligned}$$

For two, the continuity condition must be met for the above SRD local average treatment effect to be computed. Specifically, it requires the conditional expectation of the outcome Y on the forcing X to exist and be continuous at the cutoff $X = c$ separately for the treatment and control group; Indeed, the above τ_{SRD} is a simplification of:

$$\tau_{\text{SRD}} = \lim_{x \to c^+} f_1(x) - \lim_{x \to c^-} f_2(x)$$

where $f(x) = \mathbb{E}[Y|X = x]$ in general, with f_1 and f_2 respectively denoting the conditional expected outcome for the treatment and control group.

The continuity condition is necessary for the discontinuous jump in the conditional expectation to be at all observable at the threshold. This is, however, generally perceived as an intangible criteria[1] that I assume to hold for the remaining discussion here.

Given the practical infeasibility of capturing and estimating data right at or within an arbitrarily small interval between the cutoff, researchers often use the local methods: for example, a non-parametric local linear regression over only

[1] Lee and Lemieux (2010)

data points "close" to the threshold[2]. The notion of "closenes" can be regarded as a relaxation from the limit to a fixed neighborhood with length, say k. The focus of this paper is the optimal choice for size k of the local window.

2.2 Literature on local bandwidths and their selectors

Restricting RDD analysis to observations within local bandwidths has been widely recognized and adopted as the standard procedure in many econometric texts such as Hill et al. (2018). Moreover, Cattaneo and Vazquez-Bare (2017) offers a comprehensive review of different kinds of established bandwidth selectors and summarizes their development over time.

The main standards based on which the selectors have been invented are *ad-hoc* selections, local polynomials, local randomization, and falsification and validation. The branch based on local polynomials consists of multiple selectors referencing distinctive error criteria including coverage error, robust bias correction, and mean-squared-error minimization. They mostly incorporate the examination of the global fit of the estimated regression models and are designed with assumptions suitable for economic analysis. The branches based on local randomization and falsification/validation impose sets of different assumptions with the goal of providing purely data-driven tests against the existence of nearly perfect randomization near the threshold. Their approaches employ more general techniques such as cross-validation.

Local polynomial progression proposed in this paper features a mix of techniques used in local-polynomial-based and falsification-based selectors. It resembles the polynomial-based selectors in its application of polynomial regressions but differs in two key ideas. Firstly, it only considers parametric estimation of the conditional expected outcome while some polynomial-based selectors treat local linear approximation as a non-parametric strategy. Secondly, it does not concern the global performance of the regression models. Additionally, it allows room for cross-validation.

3 Local Polynomial Progression

3.1 Formulating SRD analysis

By centering the distribution of the forcing variable X at the cutoff c, one could compute the SRD causal estimate directly as the difference between the local linear regressions for the conditional first moments above and below the threshold evaluated at 0. With a fixed local window size of k enforced on both sides of the cutoff, a simplification of the current textbook approach[3] can be seen as

$$\hat{Y}(c) = \begin{cases} \hat{\alpha}_{+}(c) & \text{if} \quad x \geq c \\ \hat{\alpha}_{-}(c) & \text{if} \quad x \leq c \end{cases}$$

[2] Fan and Gijbels (1992) have shown the attractive bias properties of local linear regressions at the boundary with finite samples

[3] Hill et al. (2018) pp. 350

where for the below-threshold sub-sample

$$(\hat{\alpha}_{-}(x), \hat{\beta}_{-}(x)) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbb{1}_{X_i \in [c-k, c)} \cdot (Y_i - \alpha - \beta(X_i - c))^2,$$

and similarly the above-threshold sub-sample

$$(\hat{\alpha}_{+}(x), \hat{\beta}_{+}(x)) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbb{1}_{X_i \in [c, c+k]} \cdot (Y_i - \alpha - \beta(X_i - c))^2.$$

Only the intercept terms are relevant as

$$\hat{\tau}_{\text{SRD}} = \hat{\alpha}_{+}(c) - \hat{\alpha}_{-}(c).$$

3.2 A Current Symmetric Selector

More recent studies provide guidance to identify the optimal k mentioned above. The most influential one is Imbens and Kalyanaraman (2012). They developed a selector that adopts local polynomial strategies and specifically focused on local linear approximations. The approach seeks to maximize the local randomization by minimizing the mean-squared error of the SRD causal estimation while smoothly approximating the marginal distribution of the running variable. Using chosen kernel functions $K(\cdot)$, they formulated the local linear regressions as

$$(\hat{\alpha}_{-}(c), \hat{\beta}_{-}(c)) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbb{1}_{X_i < c} \cdot (Y_i - \alpha - \beta(X_i - c))^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

and

$$(\hat{\alpha}_{+}(c), \hat{\beta}_{+}(c)) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbb{1}_{X_i \geq c} \cdot (Y_i - \alpha - \beta(X_i - c))^2 \cdot K\left(\frac{X_i - x}{h}\right).$$

The optimal bandwidth is then chosen with respect to the error criterion based on the expected asymptotic expansion of the squared error of τ_{SRD} as $h \to 0$. Their selection process can be summarized by taking

$$\mathbf{MSE}(h) = \mathbb{E}[(\hat{\tau}_{\text{SRD}} - \tau_{SRD})^2] = \mathbb{E}[((\hat{\mu}_{+} - \mu_{+}) - (\hat{\mu}_{-} - \mu_{-}))^2],$$

and solve for the optimal bandwidth h^*

$$h^* = \arg\min_{h} \mathbf{MSE}(h)$$

3.3 Goal: allowing data-driven asymmetric selection

Before the inventions of data-driven approaches to determine the k to choose, analysts commonly adopt either *ad-hoc* window sizes or experiment and report a set of sizes with the estimates thereby produced. For the *ad-hoc* selections, it is clear that no generalized criterion governs the selection process and the robustness of the chosen bandwidths can not be effectively compared across different studies once the distribution of the running variable around the boundary is not shared. Experimenting with varying windows to some extent enhances the robustness by providing a range of tested window sizes. When the estimated SRD effects are similar in either effect size or significance over different k, one may conclude some magnitude or sign properties of the causal estimates given such similarity. However, a different issue of not being able to pin down one best estimate arises due to the lack of a neutral way to compare the estimates. Moreover, it is not appropriate to collectively present and infer from these estimates as they do not form a sampling distribution of the SRD causal estimates. In fact, each of the estimates that corresponds to one specific k_j is an instance of its own sampling distribution of the τ_{SRD,k_j}.

The *ad-hoc* selection of neighborhood windows is obviously not ideal for fully relying on discretionary consideration. One may come up with case-by-case explanations to justify their selection; for example, when the assignment of the treatment is determined by intervals in the running variable instead of a single threshold. Imbens and Kalyanaraman (2012) specifically discussed why a uniform bandwidth on both sides of the boundary is optimized as this criterion creates unfavorable conditions [4] not to allow effective choices of side-specific bandwidths.

A uniform bandwidth may be harmless when the distribution of the running variable is roughly symmetrical around the cut-off. It is automatic that a bandwidth that is optimized on one side also fits the other.

I propose further regulation on the optimal choice of the local window described in the textbook approach of estimating SRD with regression models. In essence, my strategy will provide guidance on the order of the polynomial to use and the size of the local windows, which are specific to either side[5] of the treatment threshold in the forcing variable.

The strategy proposed in this paper is designed to address two issues potentially undermining the current bandwidth-optimizing strategy: the lack of methods to allow asymmetric optimization and the inclusion of non-local observations that are far from the threshold c.

3.4 Intuition

Local polynomial progression seeks to provide a data-driven solution to the optimal window size in SRD settings with optimal order suggested as a by-product. It fits occasions where the conditional expected outcome is specified as parametric, polynomial in particular, functions of the running variable. The name "progression" can be understood by the algorithm running progressively in two dimensions: from fewer observations to more observations, and from lower-order polynomials to higher-order ones. In each dimension, progression faces constraints: more observations mean a lower level of randomness, and higher-order polynomials expose a higher risk of misspecification error. So the

[4] implementing asymmetric bandwidth optimization with respect to the MSE and AMSE objective leads to the concern of bias undesirably cancels out each other when a specific function exists between h_+ and h_-. Even if the function does not exist, large heterogeneous bandwidths may still feature close to 0 bias due to mutual canceling. See example pg. 937 Imbens and Kalyanaraman (2012)

[5] A uniform window size could also be computed with minor changes in the optimization algorithm. This will be more extensively discussed in later sections.

central question then becomes if the progression is worthwhile. In later sections, I uncover the constrained optimization algorithms that answer this question by signaling the stop of the progression when it is no longer worthy and thereby identify the optimal window size.

I view polynomial as a good parametric choice for it being parsimonious but informative in the SRD context: given a subset of the data, even low-order (no more than the third) could effectively capture the underlying data-generating process. Recall the sources of randomization in SRD settings is individuals close to the threshold are comparable to each other except for their treatment status determined by being above or below the threshold. Therefore, one would theoretically always prefer a narrower window to a wider one since closer proximity to the threshold approximates better randomization.

The conflict, however, arises in that a narrower window means a smaller share of observed data which leads to worse performance of regression models given the noisiness. This limitation of data visibility exists even if the correct parametric form of the conditional expectation of the outcome is specified.

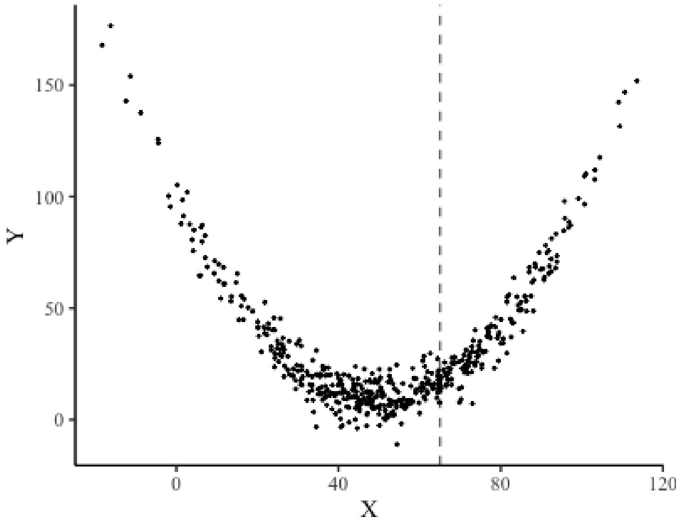


Figure 1: Quadratic functional form example

Figure 1 shows a simulated data set where the underlying functional form of the conditional expectation of the outcome is quadratic in the running variable. The dotted line marks an arbitrary treatment threshold at $X = 65$, and I let the individuals below that threshold be untreated. Additionally, there is no treatment effect. Consider only the untreated group, we run polynomial regressions up to the third to observe the following pattern in the goodness of fit:

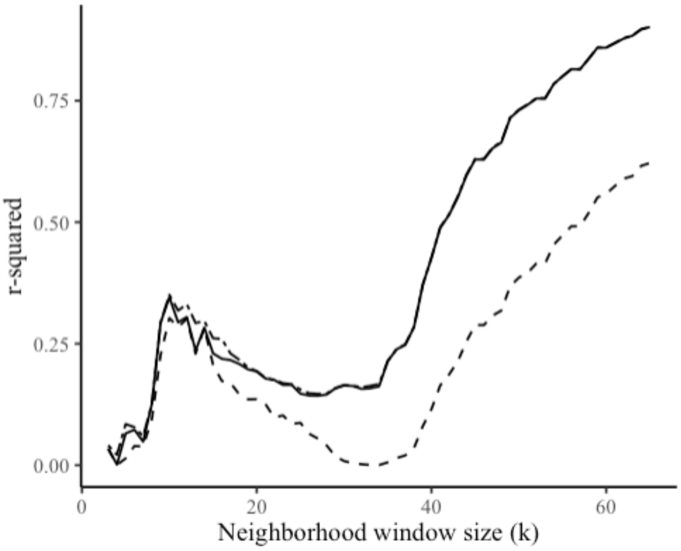


Figure 2: fit and neighborhood size for the untreated $X < 65$

In figure 2, the solid line plots the trend in the fit of the correct quadratic specification and the dotted lines are those of incorrect linear and cubic specification. One can clearly observe that the fit increases for all three specifications in the neighborhood window size up to peaking out at roughly $k = 9$. This effectively demonstrates an overall increasing pattern of the fit in the size of the neighborhood window: in fact, the additional inclusion of data along with wider windows contributes to a more precise picture of the pattern of correlation. In this simple example, the quadratic specification does not outperform the linear one until around $k = 15$. The cubic specification yields almost the same fit outcome over the entire support of the forcing variable, which is self-evident in that it contains the quadratic and linear terms and leaves the cubic term unnecessarily included.

Note that the above discussion on data presented in Figure 1 only concerns the subset of the sample to the left of the threshold on X. It is set up deliberately so one may see how the fit and pattern visibility trade-off could be viewed for one side at a time at no cost of theoretical insufficiency. The same procedure can be interchangeably applied to the right half of the data, blind to the left half as well.

3.5 Naive Objective Function Construction

In this section, I introduce the error criterion, optimization algorithm, and computation procedure of local polynomial progression to determine the optimal neighborhood window size k^*.

The example above provides numerical evidence of why a wider window is required for a better depiction of the data pattern in contrast to a narrower window favored for randomization in treatment assignment in SRD. A desirable, or optimal, neighborhood window size should achieve the balance between the two conflicting criteria: that is for a given polynomial to pin down a size as close to the threshold as possible while depicting well the conditional expectation. The idea of constrained optimization is then relevant for providing a solution to solve for the k^*.

The error criteria proposed result from combining different component objective functions, and is side-indifferent that the treated and untreated group in SRD are just two cohorts subjective to two instances of the same objective functions run independently.

The first component, the locality proxy $L(\cdot)$, governs how close the local samples captured in a window size h are to the threshold, it follows

$$L^-(k) = \frac{1}{|N^-|} \sum_{X \in N^-} \mathbb{1}_{X_i \in [c-k, c)} \text{ where } |N^-| = \sum \mathbb{1}_{X_i < c}$$

for the below-threshold-group,

$$L^+(k) = \frac{1}{|N^+|} \sum_{X \in N^+} \mathbb{1}_{X_i \in [c, c+k]} \text{ where } |N^+| = \sum \mathbb{1}_{X_i \geq c}$$

for the above-threshold group. $|N^-|$ and $|N^+|$ are respectively the total number of samples below and above the threshold. The locality proxy measures the share of data points visible within a window size.

To define the second component, the fit proxy $G(\cdot)$, adjustments on the regression functions are needed to allow the inclusion of higher-order terms. They are now

$$(\hat{\alpha}_-(c, P), \hat{\beta}_-(c, P)) = \arg\min_{\alpha, \beta} \sum (y_i - \alpha - \sum_{p=1}^{P} \beta_p (x_i - c)^p)^2$$

$$(\hat{\alpha}_+(c, P), \hat{\beta}_+(c, P)) = \arg\min_{\alpha, \beta} \sum (y_i - \alpha - \sum_{p=1}^{P} \beta_p (x_i - c)^p)^2$$

for $P \in \mathbb{N}$.[6] Then the fit proxy can be defined as

$$G^-(p, k^+) = R^2(\hat{\alpha}_-(p, k^+))$$

$$G^+(p, k^-) = R^2(\hat{\alpha}_+(p, k^-))$$

where it is a function of the order of the polynomial and the window size to measure the R-squared from fitting a p-th order regression on data enclosed by a window of size h.

The naive objective function can be constructed by taking the product of the two component proxies to be

$$O^-(\bar{p}, k^-) = (1 - L(k^-)) \cdot G(p, k^-)$$

$$O^+(\bar{p}, k^+) = (1 - L(k^+)) \cdot G(p, k^+)$$

then the optimal window size could be computed as

$$k^{-*} = \arg\max_{k} O^-(k^-, \bar{p})$$

[6] in theory there is no upper limit of p, but practically going to and above $p = 4$ only brings misspecification bias.

$$k^{+*} = \arg\max_{k} O^{+}(k^{+}, \bar{p})$$

to allow the computation of the asymmetric optimal window size for a fixed order $\bar{p}$.

Note since no assumption is made about the distribution of the forcing variable, there is no analytical solution to this constrained optimization, but it could be solved computationally by iterating over all observed x_i and picking the one that maximizes the objective function. A complete procedure guide is provided in the supplementary section 8 to uncover the specifics of the numerical computational algorithm.

3.6 Properties of the Objective Function and Score

The objective function meets the goal of penalization; the product construction of the objective itself and the evaluated value (hereafter the "score") lead to nice properties in terms of robustness against extreme scoring in components and an interpretable range. I summarize these as the three properties of the objective and score below.

Property 1. *Monotone Penalization Component with Equal Weights*

The objective function is monotone in respectively in the components. The first-order partial derivatives can be checked easily that

$$\frac{\partial O}{\partial L} = \frac{\partial}{\partial L}(1 - L) \cdot G = G$$

and

$$\frac{\partial O}{\partial G} = \frac{\partial}{\partial G}(1 - L) \cdot G = 1 - L$$

to see the goal of penalizing wider, and therefore non-local, neighborhoods as well as poorer fits.

In particular, the locality proxy is monotone in the size of the neighborhood.

Property 2. *Robustness against Extreme Component Values*

The objective function is an aggregation by product of the component functions, a strategy commonly observed in other optimization problems. Product aggregation, in particular, ensures robustness against extreme values in the components. One way to view this is to see the marginal contribution, the partial derivative, of one component is not constant: it is precisely only a function of the other.

One may also take the order-preserving log transformation of the objective function. The maximizer, optimal neighborhood size k^*, is invariant to such transformation. Namely,

$$k^* = \arg\max_{k}(O) = \arg\max_{k}(ln(O))$$

where

$$ln(O) = ln((1 - L) \cdot G) = ln(G) + ln(1 - L)$$

the logged components provide a more direct intuition that extreme values of the component scores are log-reduced.

Property 3. *Interpretable Range*

The resulting score always lies in the $(0,1]$ range. The proof is straightforward given the components are respectively a percentage share and an R-squared: $L(k) \in (0,1], G(p,k) \in (0,1]$ therefore $O(p,k) \in (0,1]$). Explicitly the score range is fixed, instead of a function of the order, threshold value, and any other feature of the observations at hand. This allows one to compare the score, or the performance of this optimization procedure, across different orders and more importantly across different studies. This property fundamentally allows more modifications to enable systematically comparing cross-order performance, which is discussed in the subsequent section.

3.7 Comparing Different Orders: Cross Validation

With simple RD settings displaying a favorable condition to visualize the data in two-dimensional graphs with outcome against the running variable, determining the other hyper-parameter, the order of polynomials, appear to require less objective justification for data sets at hand. One may naturally base on the scatter plot to examine whether a non-linear functional form is necessary and then specify the regression formula accordingly. The selection of p, however, can still be quantitatively regulated where a modified objective function enables comparing different orders of polynomial progression with cross-validation.

I acknowledge the fact that for a fixed data set, any higher-order polynomial regression is guaranteed to yield a raw r-squared at least as high as that from any lower-order ones. The naive objective score for a lower-order polynomial therefore will be bounded at least weakly above by the higher-order specifications, which effectively prevents comparison across orders. That is true even if the underlying pattern is polynomial and the lower order specification already suffices when additional inclusion of higher order terms undesirably memorizes the noise. Two schemes based on cross-validation described below offer a solution to the issue.

3.7.1 Modified Objective with Cross-Validation

To enable cross-order comparison of all candidate polynomial functional forms, I propose a modified version of the naive objective (thereafter, the "modified" objective), which incorporates cross-validation to account for misspecification of the orders. The modified objective differs from the naive only in terms of the fit proxy.

Recall the fit proxy $G(\cdot)$ in the naive objective is defined separately for the control and treatment group as

$$G^{-}(p,k^{+}) = R^{2}(\hat{\alpha}_{-}(p,k^{+}))$$

$$G^{+}(p,k^{-}) = R^{2}(\hat{\alpha}_{+}(p,k^{-}))$$

where $R^2(\cdot)$ represents the in-sample R-squared[7] value of fitting a p-th order polynomial regression on the subsample enclosed by (c,k). In the modified

[7] The R^2_{oos} here is a generalization representation of out-of-sample fit metrics. The specific computational strategy taken in the later sections is a standard five-fold cross-validation and the mean cross-validated R-squared value is used for demonstration purposes and by default in polynomial progression with the modified objective.

objective, the definition of the fit proxy G_m is replaced by

$$G_m^-(p, k^+) = R^2_{\text{oos}}(\hat{\alpha}_-(p, k^+))$$

$$G_m^+(p, k^-) = R^2_{\text{oos}}(\hat{\alpha}_+(p, k^-))$$

where R^2_{oos} represents the out-of-sample R-squared value computed from cross-validating the fit of the polynomial. The modified objective then is the product of the locality proxy as usual and the modified fit proxy as

$$O_m^-(p, k^-) = (1 - L(k^-)) \cdot G_m^-(p, k^-)$$

$$O_m^+(p, k^+) = (1 - L(k^+)) \cdot G_m^+(p, k^+)$$

and it is trivial the score resulting from the modified objective function preserves the same set of properties discussed in 3.6 too.

The introduction of cross-validation, in terms of fit performance tested out-of-sample, sets grounds for cross-order performance comparison. It resolves the upper bound issue of a higher-order specification deterministically beating lower-order ones: for an improper but higher-order polynomial approximation of the conditional expected outcome, the marginal increase of the in-sample R-squared is a consequence of memorizing the noise in the process which is penalized upon validation against an unseen test set.

3.7.2 Optimization with the Modified Objective

Notice the modified objective function takes in varying orders p as an argument instead of a fixed $\bar{p}$ order. The computational algorithm to solve for the optimal k^* remains the same in its iterative search across possible k along the axis of the running variable, but differs only in computing the score for multiple orders to be compared with each other simultaneously. With the modified objective, one solves for

$$(k_p^*, p^*) = \arg\max_{(k,p)} O_m(p, k)$$

to obtain the optimal pair of order and the corresponding window size. The procedure can be viewed in terms of a two-step breakdown. The first step is to find the optimal order p^* from

$$p^* = \arg\max_{p} O_m(p) \text{ for any } k$$

which looks for the order that achieves the highest possible scoring in the modified objective for any particular objective-maximizing window size k only specific to this order. The objective-maximizing k's need not be the same across different orders. Per my discussion in the intuition section about the trade-off between data visibility and goodness of polynomial fit, they in fact are not the same for most of the practical cases. For example, consider an overall non-linear process with an approximately linear shape around the threshold, and the only candidate orders considered are linear and quadratic. Suppose the highest scoring for the linear approximation $\max O_m(1)$ appears at $k = k_1$ and that for the quadratic approximation $\max O_m(2)$ appears at $k = k_2$. Without further prior assumptions of the process, it can be expected that $k_1 \leq k_2$ given a more local subset of the data can be well described by a linear specification while a quadratic one works better when presented with a more complete picture of the global process. However, the comparison of k_1 and k_2 is not of one's concern here for being only a partial picture; Observing $k_1 \leq k_2$ does not lead to choosing k_1 and the linear functional form as the solution. Indeed, the guidance comes from comparing the $\max O_m(1)$ and $\max O_m(2)$, the evaluated modified objective scores.

Once a p^* is fixed by the first step, the remaining procedure is the same as with the naive objective

$$k_p^* = \arg\max_h O(k, \bar{p}) \text{ by letting } \bar{p} = p^*.$$

Note in the actual computation, the optimal (k, p) is found simultaneously, and the two-step breakdown provides a conceptual framework to intuitively trace the process with.

4 Simulations

In this section, I go through the same Monte-Carlo simulated example given in Figure 1 to illustrate the workings of polynomial regression with both the naive and the modified objective. The example is inspired by a simulated data set presented on page 350 from Hill et al. (2018). The original data set is not disclosed, and the following strategy is taken for simulation based on summary statistics and the shape:

$$y_i = 100 - 3.6x_i + 0.036x_i^2 + \varepsilon$$

where the running variable and the noise are respectively Gaussian that $X \overset{\text{iid}}{\sim} N(50, 25)$ and $\varepsilon \overset{\text{iid}}{\sim} N(0, 6)$.

4.1 Naive Objective

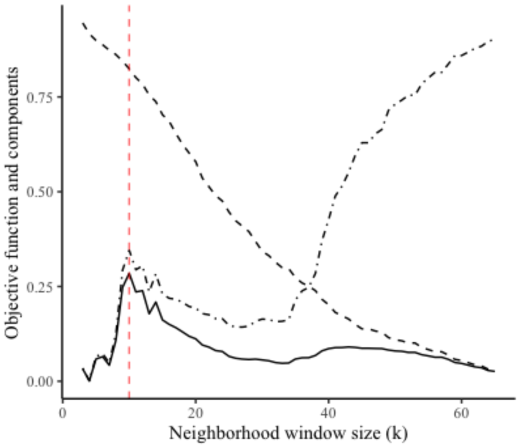


Figure 3: Quadratic example with naive objective function

Recall the quadratic example with only the untreated in Figure 1, Figure 3 visualizes the dynamic optimization process when fixing the current second-order polynomial. The objective scores, in terms of the naive objective function and two components, are plotted at all iterated window sizes. The overall objective score is plotted in solid black. The locality proxy is plotted in dotted black with a downward-sloping trend. The fit proxy is plotted in dotted black.

With the naive objective, local quadratic progression pins down $k^* = 10$, indicated by the red line, as the optimal window size for the untreated. The monotonic decreasing trend in the locality score can be easily observed in line with the goal to penalize the inclusion of more non-local observations. The fit scores vary with the window size in a slightly more complicated fashion: it gradually increases approximately before hitting the optimal window size where the marginal observable data provides a better picture for quadratic models to capture. It falls afterward for window sizes smaller than 38 to suggest that the inclusion of more data points within this range does not disclose an informative quadratic pattern. It then increases more rapidly after the kink at $k = 38$ as the complete control group is reached moving towards the tail. The identification of the optimal window size at $k^* = 10$ near the threshold reinforces the goal of polynomial regression to select the optimal window that gives a reasonably good picture of the correlation pattern with the fewest data points possible. Even if there are window sizes larger than 40 where the quadratic fit is reassured with considerably more observations, the cost of non-locality in that range prevails in the objective score.

4.2 Modified Objective

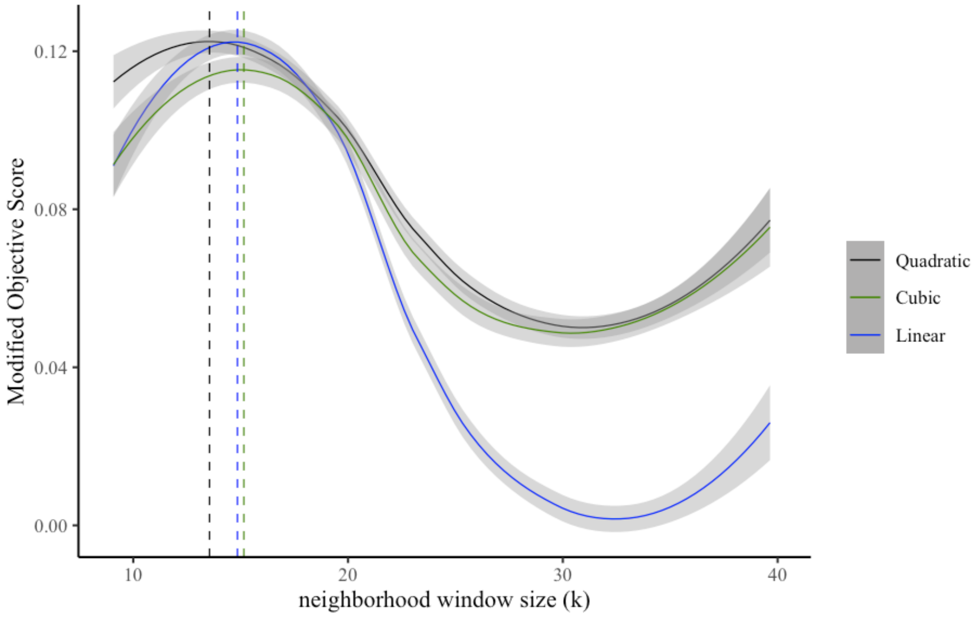


Figure 4: Quadratic example with modified objective

I apply the modified objective functions on the same set of data, the treatment group in the quadratic sample, and the results are given in Figure 4. Since cross-validation is adopted to enable cross-order comparison, the optimal pair of (k, p) differ across each execution for randomness during the K-fold sampling process. Figure 4 plots the average modified objective score against experimented window sizes with smoothing by LOWESS (Locally Weighted Scatterplot Smoothing), for visualization purposes. The objective-maximizing average window sizes are identified by vertical dashed lines. Note this plot is not based on any single particular outcome but aims to provide an average sense of how the dynamic optimization work with the modified objective. The actual precise plot of a single run looks nosier, one instance can be found in the appendix.

Following the two-step breakdown of the modified optimization procedure, one begins with pinning down the appropriate order by locating the order that achieves the highest modified score. As shown in Figure 4, the quadratic polynomial, which is the true underlying order, peaks highest in black with nearly a tie with the linear one. There is no ambiguity caused by the tie since the quadratic model is optimized at the narrowest window as well. With cross-validation, the figure also shows the upper-bound issue is addressed: the quadratic score dominates at least weakly the cubic score throughout the domain of the running variable.

5 Empirical Illustration

In this section, I apply the local polynomial progression selector on the data set originally investigated by Lee (2008) where SRD is used to detect the effect of vote share margin on incumbency advantage in the U.S. congressional elections. I further compare the optimization results from those produced by Imbens and Kalyanaraman (2012) via their kernel estimation approach on the same data set.

5.1 Data on Lee (2008)

Lee (2008) used SRD to estimate the incumbency advantage of the democratic nominees in the U.S. House congressional elections. The data used in the analysis is the record of U.S. congressional returns ranging from 1946 to 1998. Expositionally, the causal inference tested is the effect of the share of vote from the previous(t) period influence that from the subsequent$(t + 1)$ next period for the nominees from the democratic party.[8]

The forcing variable X is therefore the share of votes received by the democrat nominees. A special centering is applied here, where the share margin is computed as the democrat share minus the largest share received by a single non-democrat party, which in most of the cases was the Republican share. This centering differs from the classical $X - c$ where c be a constant shift equal to the threshold value, but is however reasonable and necessary to replace a $c = 0.5$ centering; as the appointment is exclusively determined by the majority vote,

[8] Only the democrat party discloses their specific received vote share data on a district level, which provides grounds for this research. Lee (2008) motioned that in a strict two-party system, the vote share of the other party is perfectly symmetrical and inference results stay unchanged. The assumption is largely comparable to the U.S. election where the major parties present are the democratic and republic.

the share margin at 0 always identifies an SRD where the democrat nominee wins the election at any positive margin and loses at any negative margin. The treatment D_i follows directly to be whether the democrat is elected in period t.

Figure 5 is the histogram plot of the distribution of the running variable.

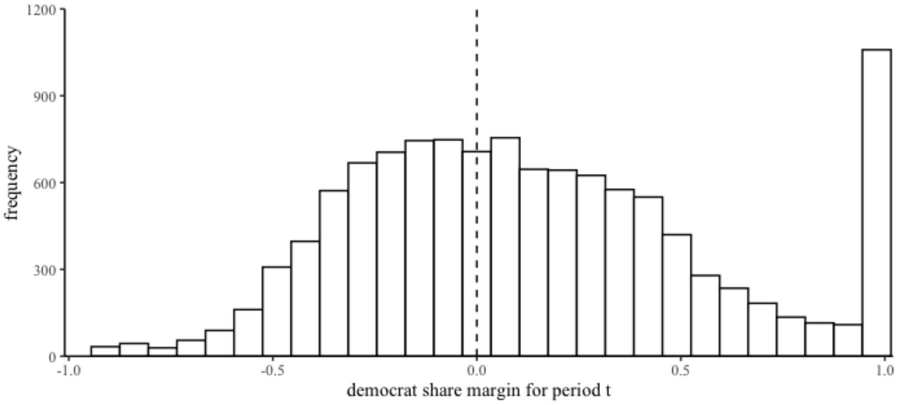


Figure 5: Distribution of Democrats' share margin

Notice that the distribution of the running variable is largely symmetrical about the threshold of the margin at 0, despite a spike at margins close to 1 where the democrats receive almost all votes in elections during certain periods. Such a spike[9], is however, not of much concern with respect to the local estimation at hand since polynomial progression theoretically should not reach as far as the spike due to severe penalization from the locality proxy. Additionally, the distribution of X is actually clustered around the center on top of its symmetry. This is a nice, but not required, property in favor of linear search algorithms like the iterative progression from the threshold implemented in polynomial regression for marginal expansion around the threshold uncover considerable additional visible data points.

The outcome Y is the raw vote share received by the democrat nominees (may be a different individual from the previous period) in period $t+1$. The scatter plot of the outcome against the running variable is given in Figure 6.

Figure 6 presents a positive correlation between the outcome and the running variable over the entire domain of the running variable. A discontinuous jump of the conditional average at the threshold is also relatively clear: the outcome of the observations to the right of the threshold are systematically higher, on average than to the left.

5.2 SRD analysis with Local Linear Progression

To make estimation comparable to that carried out by Imbens and Kalyanaraman (2012), the raw data is transformed accordingly by averaging the outcome

[9] The problem of the spike arises in global estimation procedures as significant leverage points that may need to be taken care of by weighting the observations

[10] There are observations that lie exactly on the upper and lower bounds of both X and Y axis for reason not specified in the original data release. Such observations were omitted from visualization in Figure 6 but included in the following computation. Full scatter available in the appendix.

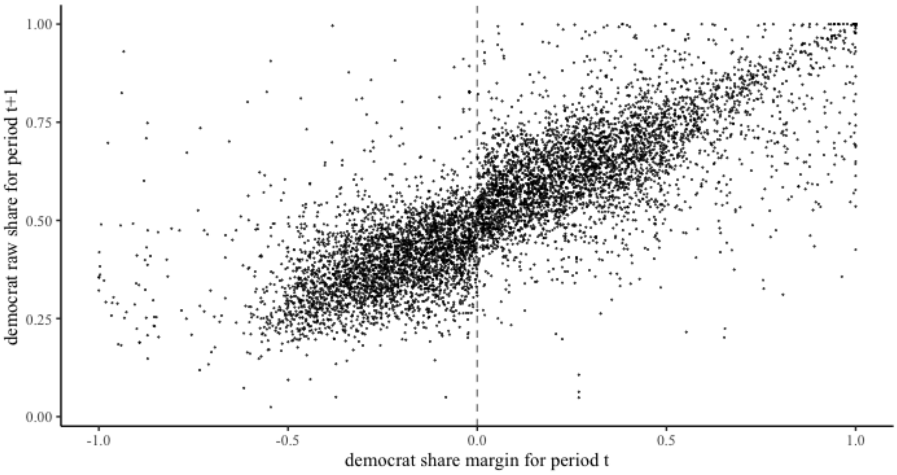


Figure 6: Scatter plot of Democrats' raw vote against current margin[10]

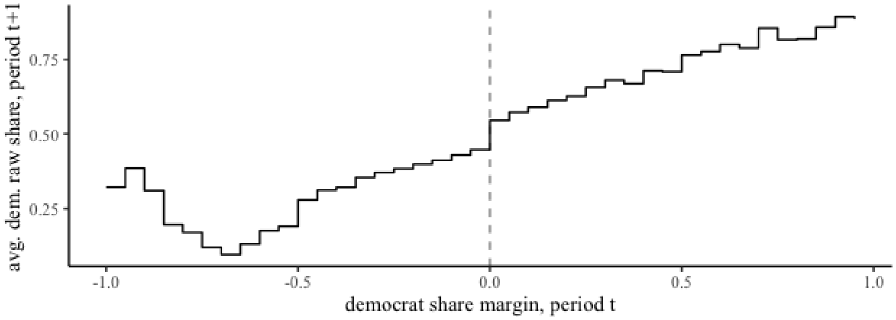


Figure 7: Bin-wise average of dem. raw share of vote against current margin

Y in bins of the running variable X. Figure 7 is the bin-wise average outcome plotted as step functions against the running variable, with bin size of 0.05 used in their paper. The discontinuous jump in the averaged outcome appears more obvious.

Applying the same local linear regression strategy, I carry out first-order polynomial regression to estimate the τ_{SRD} of incumbency advantage. For now, let me keep the outcome and the running variable as Y and X as usual for simplification while in this particular setting, they result from bin-wise averaging. To estimate the variance of the effect for hypothesis testing, the following pooled specification is adapted allowing different slopes on both sides of the threshold:

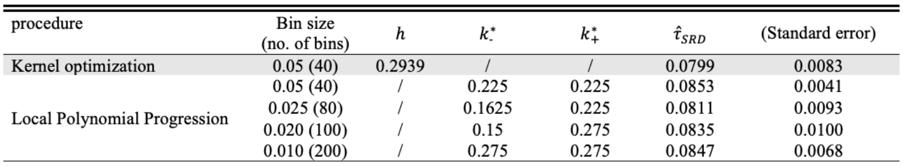
$$Y_i = \beta_0 + \tau_{\text{SRD}} D_i + \beta_1 X_i + \beta_2 (D_i \cdot X_i) + \epsilon_i \text{ for } X_i \in [k_-^*, k_+^*]$$

where the range $[k_-^*, k_+^*]$ is optimized by applying local polynomial progression.

5.3 Results

Results from the standard kernel optimization as given in Imbens and Kalyanaraman (2012) and polynomial progression are listed in Table 1.

Table 1: SRD estimates and bandwidths for Lee (2008) data



procedure	Bin size (no. of bins)	h	k^*_-	k^*_+	$\hat{\tau}_{SRD}$	(Standard error)
Kernel optimization	0.05 (40)	0.2939	/	/	0.0799	0.0083
Local Polynomial Progression	0.05 (40)	/	0.225	0.225	0.0853	0.0041
	0.025 (80)	/	0.1625	0.225	0.0811	0.0093
	0.020 (100)	/	0.15	0.275	0.0835	0.0100
	0.010 (200)	/	0.275	0.275	0.0847	0.0068

Additional bin sizes are introduced and the data sets derived accordingly with bin-average transformation are tested with local polynomial (linear in this case) progression. Note that the bin size is another hyper-parameter that measures the level of randomization displayed in estimation: larger bins average more observations in one bin to lead to worse loss of information. Therefore, only bin sizes smaller than that tested in Imbens and Kalyanaraman (2012) are included here to investigate how local polynomial progression compares to the kernel optimization given stricter conditions. One can observe from all levels of bin sizes, the local polynomial pins down strictly narrower neighborhood window sizes while the SRD analysis followed gives roughly the same effect estimate of around eight percent. Most importantly, when holding the same bin size and applying the polynomial regression, local polynomial (in this case linear) progression yields a $\tau_{\mathrm{S\hat{R}D}}$ with very similar effect size with almost half the estimated standardized error compared to that based on the kernel estimation SRD approach. Note the window sizes are the same for 40 bins, 0.225 for both the treatment and the control group. This is a coincidence resulting from applying polynomial progression independently for observations on two sides. Looking at additional bin sizes reveals that different optimized bin sizes are identified respectively for cases with 80, 100, and 200 bins. The $\tau_{\mathrm{S\hat{R}D}}$ estimates are significant in all cases as well.

6 Discussion

I discuss additional concerns about the current version of polynomial progression and suggest possible improvements in this section.

6.1 Uniform Window Size and bandwidth

Currently, the procedure specifies an asymmetric optimization procedure of the window size and order of polynomials. This is done by separating the locality (L^+, L^-) and fit (G^+, G^-) proxies. It follows the natural assumption that the data-generating process does not have to be best described with the same order polynomial, and the distribution of forcing variables could vary across the border. This assumption can be easily loosened upon additional discovery of similarities between the treated and the untreated. One would just need to enforce a collective objective function across two sides of the threshold, as the optimization is still feasible with iterative computation.

6.2 Sampling Randomness from the Modified Objective

Given the nature of cross-validation, the k-fold sample split gives rise to randomness when the modified objective is applied. Though with no specific control over the sample split process, each run of the modified polynomial progression gives a faithful estimation of the optimal window size and the order. To address the undesired volatility, I suggest combining the modified and the naive objective. By repeating the modified objective for sufficiently many rounds, one may identify the optimal order p^* if it is suggested to maximize the modified objective in a considerably larger share of the experiments. With the p^* fixed, the remaining search for k^* can be then achieved with the naive objective, which gives a deterministic answer.

6.3 Customized tuning

At this stage, I offer no tuning parameters in the objective function; this is intended to eliminate subjective intervention in the optimization process. However, in special cases (for example when locality is extremely important due to a sparse distribution of the forcing variable), tuning can be achieved by editing the objective function, one way is to redistribute the powers of the components such that

$$O(p,h) = (1 - L^{\gamma}(k)) \cdot G^{\theta}(p,k) \text{ where } \gamma + \theta = 1$$

, the indices represent the relative adjusted marginal contribution to the objective by either proxy.

7 Conclusion

This paper proposes local polynomial progression as an option for bandwidth selectors in SRD settings. Though noting it as “optimal” in many places throughout the above discussion, I would like to stress that the optimality of the selector lies fundamentally in the objective functions. The ultimate goal of the selector is to provide data-driven evidence for asymmetric bandwidths. As mentioned in section 6, there is room for additional tuning, and it is intentionally simplified to offer transparency and avoid being adopted just as one other way to hack the data at hand. Given the existence of other selectors whose properties in their corresponding estimators have been extensively explored and rigorously proved, one might consider experimenting with different selectors and comparing the performance of local polynomial progression to the rest. The special cases where a non-linear data-generating process is readily visible with a limited amount of observations or significantly different distribution of the running variable across the threshold are good places to justify the usage of local polynomial progression.

8 Optimization Supplement

In this supplementary section, I go through the specific step-by-step computational procedure of polynomial progression with the naive objective.

8.1 General Procedure with the Naive Objective

As discussed in section 3, the optimization procedure is shared across the threshold but carried out independently; therefore, the same algorithm applies to both sides and only differs in the direction of progression. Specifically, the iterative search begins at the threshold and progressively includes more non-local observations, so each additional iteration examines additional subjects with higher running variable values for the above-threshold group (and vice versa, lower running variable values for the below-threshold group). For demonstration purposes, I will go through the steps for the above-threshold group as an example.

Consider subjects indexed by unit i, for $i \in \{1, 2, ..., M-1, M, M+1, ..., N\}$, for each individual we observe the scalar running variable X_i and the continuous outcome of interest Y_i. The assignment of treatment is determined by a threshold c such that $D_i = \mathbb{1}_{X_i \geq c}$.

Without loss of generality, denote the treatment group by first M (including the M-th) subjects in an observed sample to result in

$$D_i = 1 \text{ for } i \in \{1, 2, ..., M-1, M\}$$

and order the index by running variable values so that

$$\{x_i : x_1 \leq x_2 \leq ... \leq x_{M-1} \leq x_M\}.$$

Centering at c defines a complete ordered set of candidate bandwidths,

$$K = \{k_i \equiv x_i - c : k_1 \leq k_2 \leq ... \leq k_{M-1} \leq k_M\}$$

given

$$x_1 - c \leq x_2 - c \leq ... \leq x_{M-1} - c \leq x_M - c \implies k_1 \leq k_2 \leq ... \leq k_{M-1} \leq k_M.$$

The optimization of k^* with naive objectives at a given order $\bar{p}$ is then simply

$$k^* = \sup_{k_j \in K'} O(k_j, \bar{p}) \text{ for some } K' \subset K$$

where K' refers to an ordered subset of candidate bandwidth that starts at some observations away from the threshold, which namely is $K' = \{k_S, k_{S+1}, ..., k_M\}$ for $1 < S < M$.

$\sup O(\cdot)$ can be easily computed as the maximum obtained from looping through the bandwidths indexed by $\{S, S+1, ..., M-1, M\}$. Plotting the overall objective scores and the component scores against the window sizes yields the visualization of the dynamic optimization process of polynomial progression (check examples of such plots in section 4).

8.2 Choice of the Starting Point

K' is introduced to manually prevent locality traps in general and local misspecification errors against raw data that is not bin-wise average.

Locality traps are a common issue if the optimization procedure starts with the first observation right next to the threshold without out-of-sample testing. Consider the extreme case that only $i = 1, 2$ are included and they deterministically max out the naive objective given both component objectives achieve their respective upper bounds. In particular, with only two data points specifying a linear regression

$$G(\cdot) = \max G(\cdot) = 1 \text{ and } 1 - L(\cdot) = 1 - \min L(\cdot) = \frac{n-2}{n}$$

where n is the count of the observations in the treatment group.

Bin-wise averaged data is a common approach in RDD analysis (see an empirical example in section 5) to address skedastic issues in the outcome variable. With raw data in the absence of bin-wise averaging, local misspecification arises from this progressive iterative search algorithm. An intuitive example is that, when imposing a sufficiently small window size with few observations captured, the regression functions will in most cases output some near-vertical patterns. Such patterns, however, are a natural result of the variance of the outcome itself.

Therefore the construction of K' depends on the issue to be addressed. For the simpler case of preventing locality traps, a rule-of-thumb of starting with S as the nearest integer around $0.1M$, $0.15M$, and $0.2M$ suffices. To address the local misspecification error, the same rule could still apply with proper visual evidence from the scatter plot of Y against X. However, there are more complicated scenarios: for example, when additional control co-variates are present, there is no effective ways to visualize high-dimensional correlations. In these cases, comparing the average skedasticity of the outcome and the variance of the running variable backs up the construction of K' and therefore the choice of S. I recommend estimating the expected conditional variance of the outcome $\mathbb{E}\left[\sigma^2(Y)|X\right]$ by the approach suggested in Fan and Yao (1998).

9 Appendix

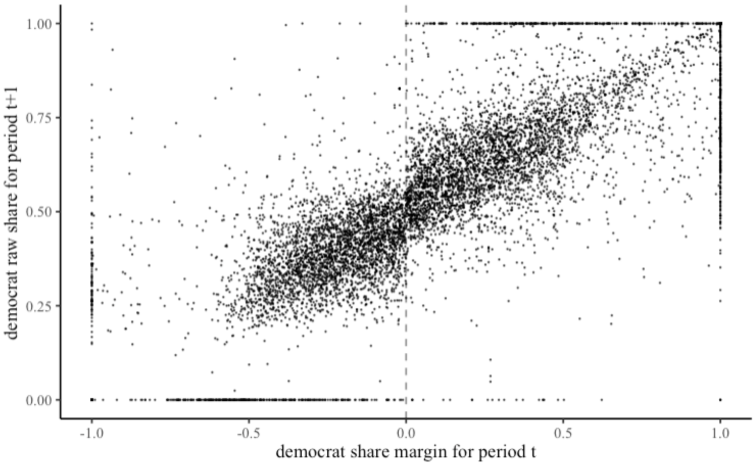


Figure 8: Raw Scatter of Lee (2008) data

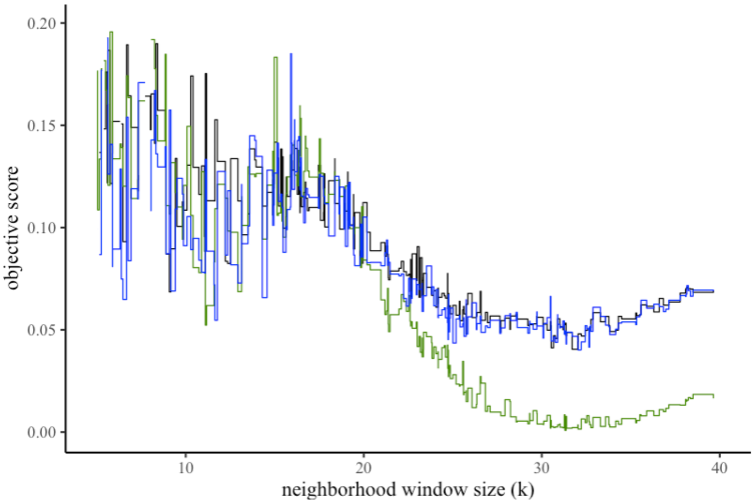


Figure 9: An instance of cross-order comparison

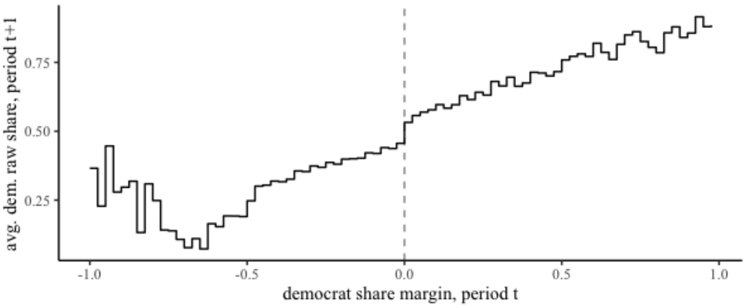


Figure 10: Bin-wise average: bin size = 0.025

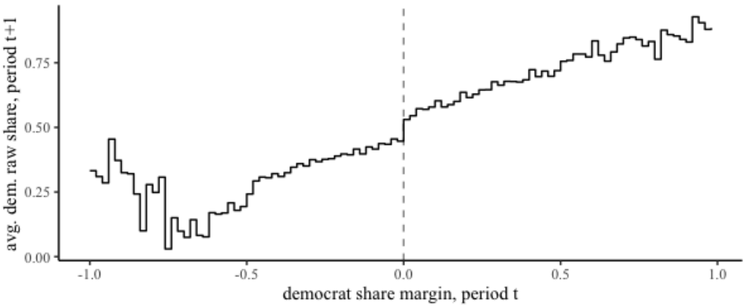


Figure 11: Bin-wise average: bin size = 0.02

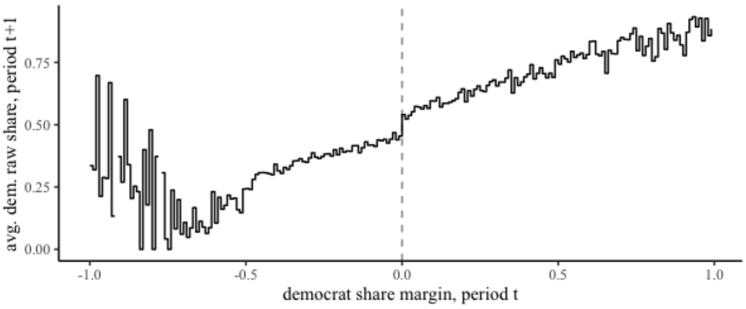


Figure 12: Bin-wise average: bin size = 0.01

References

Cattaneo, M. and Vazquez-Bare, G. (2017). The choice of neighborhood in regression discontinuity designs. *Observational Studies*, 3:134–146.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4).

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3).

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1).

Hill, R. C., Griffiths, W. E., and Lim, G. C. (2018). *Principles of Econometrics, 5th Edition*. WILEY.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economics Studies*, 79(3).

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142(2).

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2).

Thistlewaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(1).