Principled estimation of regression discontinuity designs

Jason Anastasopoulos
`ljanastas@uga.edu` *

First Version: August 30th 2018†
Current draft: January 24, 2020

Abstract

The regression discontinuity design (RDD) has become a popular method for making causal inferences with observational data under minimal assumptions. Local average treatment effects (LATE) for RDDs are typically estimated using local linear regressions with pre–treatment covariates added to increase the efficiency of treatment effect estimates. In political science applications where there are typically few observations around the cutpoint, covariate selection can have a large impact on treatment effect and standard error estimates. In this paper, I propose a principled, efficiency-maximizing approach for selecting covariates to include in RDDs. This approach allows researchers to combine substantive insights with regularization via a novel adaptive LASSO algorithm. When combined with currently existing robust estimation methods, this approach improves the efficiency of LATE RDD with pre–treatment covariates[1].

Keywords: regression discontinuity design, RDD, causal inference, treatment effect, adaptive lasso, machine learning, regularization, covariates, model selection, shrinkage.

Word count: 4,099.

*I thank Mariliz Kastberg-Leonard, Gary King, Max Gopelrud, Molly Offer-Westort, Erin Hartman, Marc Ratkovic, Jonathan Katz, Kevin Esterling, Luke Miratrix , Richard Nielsen and Rocio Titunik for their helpful comments and assistance. This is a draft, please do not cite without permission.

†Prepared for the annual *American Political Science Association Conference* in Boston, MA

[1]The approach will be implemented in a forthcoming *R* package, *AdaptiveRDD* which can be used to estimate and compare treatment effects generated by this approach with extant approaches.

The regression discontinuity design (RDD) has become a popular method for making causal inferences with observational data under minimal assumptions (see eg. Caughey and Sekhon (2011); Erikson and Rader (2017); Green et al. (2009); Imai (2011); Skovron and Titiunik (2015)). The premise of the RDD is conceptually simple and intuitive. Around a narrow interval of a threshold for a variable that assigns a treatment (running variable), treatments can be plausibly considered to be "as-if" randomly assigned. While bandwidth selection, kernel choice and estimation strategy for RDDs are well understood, work on theoretical considerations regarding the common practice of including covariates to adjust local average treatment effect (LATE) estimates is relatively new (Frölich 2007; Calonico et al. 2019).

Calonico et al. (2019) in particular provide strong theoretical and empirical grounds for continuing the practice of estimating RDDs with pre–treatment covariates. Despite the substantial achievements in this area, however, applied researchers are left will little guidance regarding which covariates they should include or exclude when estimating RDDs. In this paper, I develop a framework and a method for selecting pre-treatment covariates to incorporate into RDDs to address this gap for applied researchers.

This approach is flexible and allows researchers to combine substantive knowledge with regularization via a novel Adaptive LASSO that is employed here on the basis of its demonstrated model selection (oracle) properties (Zou 2006). This method allows applied researchers to initially choose covariates to include based on context-specific substantive knowledge of the estimation problem and then subjects this initial choice to further optimization through covariate trimming via regularization using the Adaptive LASSO, a machine learning algorithm that is used primarily for dimensionality reduction when consistent and correct model selection is important[2].

[2]As I describe in more detail below, this contrasts with the more "traditional" version of the

The remainder of this paper is as follows. Section 1 provides a brief introduction to LATE estimation for sharp RDDs with local linear regression (LLR), the focus of this paper; Section 2 introduces the adaptive lasso and accompanying implementation algorithm along with relevant theoretical derivations; Section 3 provides an applied example of enhanced LATE estimation using a close election RDD study of the effect of holding political office on profit margins in Russian firms published in the *American Political Science Review* by (Szakonyi 2018). Section 4 provides empirical evidence of the bias reduction and efficiency gains of this method using a series of simulated close election RDDs with covariates and finally, Section 5 concludes with a discussion of future research in this area.

1 Covariate adjusted LATE in regression discontinuity designs

Regression discontinuity designs are a framework for the causal estimation of local average treatment effects with observational data. This is accomplished using a running variable $F_i; i = 1, \cdots, n$ which assigns treatment T_i on the basis of some threshold value f such that if $F_i > f$, a unit (individual, geographic unit etc) is assigned to treatment $T_i = 1$ and is not assigned to treatment otherwise $T_i = 0$. Assuming continuity of the forcing variable, the sharp RDD leverages this mechanism by allowing for the causal estimation of LATE around a narrow window of the threshold $f - \epsilon < f < f + \epsilon$ by making the assumption that, in the limit of this window, units are as "as if" randomly assigned to a treatment (Hahn, Todd, and Van der Klaauw 2001). In line with other work on the RDD, this paper is concerned primarily with

LASSO developed by Tibshirani (1996) which is concerned primarily for MSE reduction at the expense of consistent model selection and specification.

the sharp RDD, the most commonly used design in the literature (Calonico et al. 2016).

Under the potential outcomes framework (Rubin 2005), define Y_i as the observed outcome for i, $Y_i(1)$ as the outcome, had unit i received treatment, and $Y_i(0)$ as the outcome had unit i not received treatment, RDDs allow us to estimate the local average treatment effects (LATE) at the threshold $F_i = f$. For purposes of illustration, we assume that $f = 0$:

$$LATE = \tau = \lim_{F_i \downarrow 0} E[Y(1)_i | F_i = f + \epsilon] - \lim_{F_i \uparrow 0} E[Y(0)_i | F_i = f - \epsilon] \tag{1}$$

Estimation of τ is typically accomplished through a local linear regression (LLR) in a neighborhood of the cutpoint $F_i \in [c-h, c+h]$ which is determined through optimal bandwidth selection procedures designed to minimize cross-validated MSE Imbens and Kalyanaraman (2012).

$$\hat{Y}_i = \beta_0 + \hat{\tau} T_i + f(T_i, F_i) \tag{2}$$

In Equation 2, $\hat{\tau}$ is the estimated local average treatment effect, T_i is a binary treatment indicator function which equals 1 when $F_i > 0$ and $f(T_i, F_i)$ is a function of the forcing variable which often takes the form of a non-parametric kernel or p^{th} order polynomial. A common LLR model estimated in the literature is the model shown in Equation 3 (Calonico et al. 2019):

$$\hat{Y}_i = \beta_0 + \hat{\tau} T_i + \delta(F_i \cdot T_i) + X\beta \tag{3}$$

In Equation 3, a set of covariates X are added to increase the precision of LATE. Calonico et al. (2019) derive the covariate adjusted estimator of $\hat{\tau}$ and demonstrate that pre–treatment covariate adjustment typically leads to more efficient estimates of $\hat{\tau}$ but, as mentioned above, there is little guidance regarding *which* pre–treatment covariates should be included to maximize the efficiency of LATE. Table 1 which lists the types of covariates chosen for similar close-election RDD designs highlights this problem. This is particularly problematic in small N estimation contexts and when covariates are correlated with the running variable, cases in which covariate selection can have a much greater impact on LATE efficiency and point estimates.

As a solution to a similar problem in the context of randomized experiments (Bloniarz et al. 2016) propose selecting covariates using a shrinkage and variable selection method known as the LASSO, a practice which I modify and extend to LATE estimation in the regression discontinuity design here by employing the adaptive lasso, the only version of the lasso which has oracle (correct model selection) properties (Zou 2006).

Covariate selection using the adaptive lasso has a number of benefits. First, given any initial set of covariates chosen by the researcher, subsequent covariate selection using this method can improve optimal bandwidth choice via model MSE minimization independent of the bandwidth estimation algorithm; second, this method can maximize LATE efficiency and; third, the method constrains the extent which a treatment effect estimate can be "p-hacked" through the practice of adding covariates. Each of these properties are demonstrated below.

Journal (Year),Author(s)	Title	DV	Forcing	Covariate Types	Lowest N
APSR (2009) Eggers and Hainmueller	"MPs for Sale? Returns to Office in Postwar British Politics"	Logged wealth at death	Vote share margin	Candidate/official level traits	165
APSR (2014), Ferwerda and Miller	"Political devolution and resistance to foreign rule: A natural experiment"	attacks	commune distance from demarcation line	mean elevation, train station distance, communications available, farmed area, ruggedness of the landscape, population	15
APSR (2015), Hall	"What happens when extremists win primaries?"	party victory	Vote share margin	Congress fixed effects	35
APSR (2018), Szakonyi	"Businesspeople in elected office: Identifying private benefits from firm-level returns"	Revenue and profit margins	vote share margin	sector, region, year fixed effects, candidate level covariates	136
AJPS (2011), Boas and Hidalgo	"Controlling the airwaves: Incumbency advantage and community radio in Brazil"	radio station coverage	vote share margin	municipal population	33
JOP (2014), Boas, Hidalgo, and Richardson	"The spoils of victory: campaign donations and government contracts in Brazil"	total contracts	vote share margin	firm level fixed effects	45

Table 1 – Covariate types chosen for RDD estimation in top political science journals.

2 Regularization, machine learning and variable selection

Regularization methods are tools used primarily for prediction problems and machine learning applications as a means of reducing the dimensionality of a feature space to avoid over fitting of a prediction model. In the context of linear models, ridge regression and lasso regression are the primary regularization methods used for linear prediction problems Tibshirani (1996). Each method applies a term which penalizes each additional variable added to an OLS model in a different way. For instance, in all OLS problems our goal is to find coefficient estimates β which minimize the squared error loss:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X\beta)^2$$

OLS under mild assumptions is guaranteed by Gauss-Markov to be the best linear unbiased estimator (BLUE) of the coefficient values. However, if our ultimate goal is *prediction* using a linear model, as is typically the case in the machine learning context, the bias–variance trade-off allows us to exchange unbiasedness of coefficient estimates for a model that makes better out-of-sample predictions (lower MSE) Tibshirani, Wainwright, and Hastie (2015). This was first demonstrated by statistician and mathematician Charles Stein in 1956 and improved upon by statistician Williard James and Stein in 1961 and came to be known as James-Stein *shrinkage* estimation of linear models (Stein 1956; James and Stein 1961).

2.1 Shrinkage and ridge regularization

As its name suggests, shrinkage estimation is a means of optimizing the predictive abilities of linear models through shrinking coefficient estimates toward zero. One of the first shrinkage methods developed for linear models was ridge regression which added a $\updownarrow_2$ penalty to the OLS minimization problem (Tihonov 1963):

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \underbrace{\sum_{i=1}^{N}(Y_i - X\beta)^2}_{OLS\ Loss} - \underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{Ridge\ Penalty}$$

In ridge regression equation above, the original OLS loss function is estimated with a penalty which penalizes the addition of more variables and is determined by the tuning parameter λ which is typically estimated using cross-validation (Tibshirani 1996).

2.2 Shrinkage and selection with lasso regularization

This ridge regression estimator ends up introducing biased (shrunken) coefficient estimates, but through the introduction of this bias, minimizes MSE and improves ability of the model to make better predictions in out of sample data. Unfortunately, ridge regression cannot be used as a variable selection tool because it will never shrink coefficients to zero (Tibshirani, Wainwright, and Hastie 2015), however, the LASSO, an acronym for "least absolute shrinkage and selection operator," which slightly modifies the penalty term above to an $\updownarrow_1$ norm allows the model to serve as both a shrinkage and selection method:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \underbrace{\sum_{i=1}^{N}(Y_i - X\beta)^2}_{OLS\ Loss} - \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{Lasso\ Penalty}$$

Due to the nature of the constrained optimization problem presented by the objective function above, some coefficients will be shrunk toward zero, thus allowing for the lasso to be model selection and shrinkage tool. Additional versions of the lasso which involved tweaks of the penalty for specific high dimensional problems include the elastic net, which can be thought of as a middle ground between ridge regression and the lasso, and the "group lasso", which is used to select out large groups of covariates.

2.3 Variable selection and oracle properties of the adaptive lasso

Most variations of the LASSO applicable to high dimensional ($p > n$) data often do a good job of minimizing MSE, but fare poorly in simulations in which the ultimate goal is to retrieve the correct subset of covariates from a relatively large pool Zou (2006). As such, the usefulness of the ordinary lasso for LATE adjustment in RDDs, which do not typically involve high dimensional problems with covariates, is somewhat questionable. Fortunately, the *adaptive lasso* first introduced by Zou (2006) was developed with the goal of maximizing "correct" variable selection for both low and high-dimensional estimation problems, making it an ideal candidate for covariate adjustment of LATE in RDDs and other causal inference contexts which call for covariate adjustments. As with other flavors of the LASSO the adaptive LASSO involves some alterations of the regularization term:

$$\hat{\beta}_{adaptive} = \arg\min_{\beta} \underbrace{\sum_{i=1}^{N}(Y_i - X\beta)^2}_{OLS\ Loss} - \underbrace{\lambda \sum_{j=1}^{p} \omega_j |\beta_j|}_{Adaptive\ Penalty}$$

The inclusion of a set of weights ω, differentiates the adaptive LASSO from other

LASSO varieties. In the adaptive LASSO as developed by Zou (2006), weights are chosen from the OLS estimates of the coefficients such that:

$$\omega_j = \frac{1}{|\beta_j|^\gamma}$$

where the β_j are the coefficients estimated from an OLS model:

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j$$

and γ is a tuning parameter that can take on positive values. In his original adaptive lasso study,Zou (2006) tuned the adaptive lasso with γ values of 0.5,1 and 2 or cross-validation. In his simulations, the best results were achieved with $\gamma = 2$ followed by γ selected by cross validation. The tuning parameter λ is estimated in the ordinary way via k-fold cross-validation[3].

What makes the adaptive lasso appealing for causal inference, in general, is that with the appropriate value of λ estimated from the data, the adaptive lasso exhibits oracle properties: it tends to consistently select a correct subset of variables out of a larger set and has asymptotic guarantees of unbiasedness and normality (Zou 2006). This is especially useful when the lasso is used as a variable selection, rather than shrinkage tool, which will be true more often in the context of covariate adjustments of LATE in RDDs and other causal inference contexts more generally.

Indeed, as with ridge regression and other varieties of lasso, however, raw parameter estimates ($\hat{\beta}_{adaptive}$) can be biased in finite samples, which seemingly limits the utility of this method for causal inference more generally. Fortunately, however, as Bloniarz et al. (2016) and others point out, estimation through a two-step procedure

[3] In most software packages k is set to 10 but this should be adjusted depending upon sample size.

in which the lasso is used as a model selection tool and final parameter values are estimated using OLS allows us to obtain BLUE coefficient estimates with appropriate standard errors in an easily interpretable model.

Accordingly, this is the approach that we employ here discussed in more detail below. Furthermore, here, as in (Bloniarz et al. 2016), we argue that adaptive lasso covariate adjustment of LATE can improve the precision of estimates and also function as a means of “principled” model selection that can avoid some of the pitfalls of model manipulation to recover statistically significant treatment effects (ie “p-hacking”) for RDDs. Based on a series of simulations and on the basis of the theoretical results discussed here and previously in (Bloniarz et al. 2016), we recommend a four–step process for RDD treatment effect estimation when covariates are included. This process is outlined in Table 2 and each step is described in more detail below.

Table 2 – Overview of an RDD estimation algorithm with the adaptive lasso.

Step 1	Researcher pre-treatment covariate selection	Covariates are selected by the researcher on the basis of substantive concerns. and data issues.
Step 2	Adaptive lasso regularization	The model from Step 1 is estimated using an adaptive lasso as described below.
Step 3	Covariate adjustment	Covariates and higher-order terms whose coefficients are shrunk to 0 are excluded from the final model. The adaptive lasso is tailored in this case such that **the treatment effect, forcing variable and variables included in the kernel chosen are NOT penalized.**
Step 4	CCT robust estimation of final model	The modified model from Step 3 is estimated via the CCT robust procedure (Calonico, Cattaneo, and Titiunik 2014).

Briefly, the four steps involve researcher model selection based on substantive or theoretically motivated concerns, the application of a adaptive lasso regularization with tuning parameter cross validation; variable selection based on the results of adaptive lasso estimation in the previous step and finally CCT robust estimation of the model selected from Step 3. Each of these steps along with treatment effect estimates produced by this method in the context of RDDs with local linear regression and covariates are derived below.

3 Adaptive lasso estimation of LATE for RDDs

3.1 Step 1: Researcher Pre-Treatment Covariate Selection

The purpose of including pre-treatment covariates in RDD estimation, as in randomized experiments, is to increase the precision of treatment effect estimates (Bloniarz et al. 2016; Calonico et al. 2018). This increase in precision can be the result of improved bandwidth selection, reduced model variance or a combination of the two. Some questions that researchers may struggle with, however, are: (1) *which* pre-treatment covariates to include in the model and; (2) whether pre-treatment covariates should be included before or after optimal bandwidth selection.

This is a thorny issue because all of these decisions can have significant downstream consequences for LATE estimation and efficiency, particularly when covariates included are highly correlated with the forcing variable and in small N local linear regression contexts which tend to be common in RDD estimation within the political science literature. As a result, temptations to manipulate covariate selection to maximize the statistical significance of LATE estimates is high, particularly in cases where LATE estimated without covariates are marginally significant (i.e. $0.05 < p < 0.10$).

While the automated model selection algorithm proposed here in Table 2 cannot eliminate "p-hacking", it is a procedure that can at the very least attenuate the ability of researchers to engage in this practice while simultaneously providing LATE estimates when covariates are introduced than researcher model selection alone. That being said, initial decisions regarding which pre-treatment covariates to include should **always** be made on the basis of expert judgment/the researcher's expectation of which are the most relevant to the problem at hand. Since RDDs in the political science literature are typically conducted with close election vote share as the forcing variable F_i and the treatment of interest is an election win T_i where $T_i = 1$ if $F_i > c$ and $T_i = 0$

otherwise we focus on this type of RDD to illustrate the method.

$$Y_i = \alpha + \tau T_i + \gamma F_i + \delta(F_i \cdot T_i) + X\beta + \epsilon_i \tag{4}$$

Equation 4 is a typical local linear regression model estimated to obtain the treatment effect estimate $\hat{\tau}$ where the observations i are in some neighborhood, a of the forcing variable F_i around the cutpoint c, i.e. $i \in F_a \pm c$ and X is a matrix of covariates. In these circumstances, the covariates included are often characteristics of the candidate (age, sex, etc) and characteristics of an electoral unit that they represent (pre-treatment demographics etc). Szakonyi (2018), for instance includes candidate controls such as age, gender, incumbency, ruling party membership, state ownership, foreign ownership, and logged total assets in the pre-election year in his estimates. As I mentioned above, selection of this initial set of covariates should *always be dictated by a substantive understanding of the problem at hand.*

3.2 Step 2: Adaptive lasso regularization

Once the model in Equation 4 has been selected, a question that remains is whether this is the **best** possible model that can be fit which invariably raises the question of what "best" means in this context. Here we define "best" as a model in which a set of covariates X^* are chosen out of the original set of covariates X which minimizes the variance of LATE, $Var(\hat{\tau})$. All things equal, it can be shown that minimizing $Var(\hat{\tau})$ can be accomplished by minimizing the mean squared error (MSE) of the local linear regression.

Formally, if X^s is a subset of covariates from X, $X^s \subseteq X$, we seek to choose an X^{s*} that minimizes the mean squared error (MSE) of Equation 4 w.r.t to the LLR parameters which we describe as the vector $\mathbf{\Theta} = (\tau, \gamma, \delta, \beta)$ for convenience. Thus:

$$\arg\min_{\boldsymbol{\Theta}} \sum_{i=1}^{N} (Y_i - [\alpha + \tau T_i + \gamma F_i + \delta(F_i \cdot T_i) + X^{s*}\beta])^2 \tag{5}$$

While many methods exist for choosing X^{s*}, LASSO regularization is suited directly to the estimation of linear models and has been found to outperform other automated variable selection methods (Tibshirani, Wainwright, and Hastie 2015). Also, since we are primarily concerned with an optimal *variable selection* and minimizing the bias of the LATE in a low-dimensional context, the adaptive lasso discovered by Zou (2006) is a natural choice since it is the only lasso variety which possesses the oracle property, as mentioned above. This is important because it guarantees that it will be consistent in both estimation of τ and in variable selection. Formally this implies asymptotic unbiasedness of $\hat{\tau}$ in the ordinary sense:

$$\sqrt{n}(\hat{\tau} - \tau) \to N(0, \mathbb{I}^{-1}(\tau))$$

while simultaneously identifying the correct set of non-zero coefficients. These properties ensure that adaptive lasso estimates of τ are asymptotically *at least as good, in terms of efficiency and bias, as LLR without adaptive lasso variable selection.*

Learning about which covariates to exclude in RDDs, however, requires modifying the adaptive lasso to the RDD context. In particular, we do not want to penalize the treatment effect, forcing variable or kernel, but do want to penalize any additional covariates. This can be accomplished by simply estimating a modified version of the adaptive lasso in which the weights for these coefficients are set to 0 while the weights of the added covariates are identical to those of the adaptive lasso. The full initial model to be estimated is thus:

$$\arg\min_{\mathbf{\Theta}} \sum_{i=1}^{N} \left[Y_i - (\alpha + \tau T_i + \gamma F_i + \delta(F_i \cdot T_i) + X\beta)\right]^2 + \lambda \left[\sum_{j=3}^{p} \omega_j |\beta_j|\right] \tag{6}$$

Where $\omega_j = 1/|\beta_j|^\gamma$ are obtained through OLS estimation of β_j and γ is determined through cross-validation as described above. Again, the tuning parameter λ is estimated with k-fold cross validation.

3.3 Step 3: Automated Model Selection

Once parameters from the adaptive lasso model in Equation 7 are estimated using the optimal penalty value λ^{RDD} and optimal weights, those covariates which are shrunk to zero are excluded from the model prior to calculating the optimal bandwidth. The resulting model used to estimate optimal bandwidth and subsequently, robust treatment effects, is will thus be:

$$\mathbb{E}(Y_i|T_i, F_i, X^o) = \alpha + \tau T_i + \gamma F_i + \delta(F_i \cdot T_i) + X^o\beta \tag{7}$$

where $X^o \subseteq X$ is the truncated set of covariates selected out by the adaptive lasso described above. Since optimal bandwidth selection algorithms such as Imbens–Kalyanaraman use cross-validated MSE as criteria for selecting the "best" possible bandwidth, MSE for bandwidth values estimated using covariates pre-processed by the adaptive lasso method described should be equal to or less than model MSE for bandwidth values estimated using the full model from Step 1.

As I demonstrate below, this method can be incorporated into RDD estimation with covariates *before* bandwidth selection, which will alter the optimal bandwidth chosen, or *after* bandwidth selection if the bandwidth is set to a predetermined value (eg. 1%, 5% etc for close election RDDs).

3.4 Step 4: Regularized CCT Robust Estimation

Steps 1-3 involve selecting an optimal LLR conditional expectation function (CEF), $\mathbb{E}(Y_i|T_i, F_i, X^o)$, and estimating an optimal bandwidth h_o^* based on the CEF. Once this has been accomplished, final treatment effect estimates are produced using CCT robust estimation (Calonico et al. 2018). The final LATE estimated using this procedure, $\hat{\tau}_{adaptive}$ will be at least as efficient as the un-adjusted treatment effect estimated from the original model, $\hat{\tau}$:

$$Var(\hat{\tau}_{adaptive}) \leq Var(\hat{\tau})$$

A proof of this is provided in the Appendix.

4 Empirical Illustration: Do Firms Profit from Having Elected Board Members?

Knowledge of whether politicians benefit financially from holding office-holding is essential is essential for ensuring the legitimacy of democratic institutions. Earlier work using RDDs to estimate the returns to office found large lifetime earnings effects by barely (initially) elected members of the British Parilament (Eggers and Hainmueller 2009). Subsequent work in different national contexts has found similar results as well (see eg Fisman, Schulz, and Vig (2014) (India), Truex (2014) (China), etc). Recent innovative work published in the *American Political Science Review* by Szakonyi (2018) adds an interesting dimension to this literature by using a close election RDD to explore whether office-holding affects the profits of firms whose board members held political office in Russia. Using a close election RDD, Szakonyi (2018) finds that office holding positively affects both firm profitability and firm revenue. In the empirical illustration below, I apply both the adaptive lasso algorithim described above to optimize covariate adjusted treatment effects.

In the following analysis, I replicate the local linear regression in Table 2 of Szakonyi (2018). In this table, the author uses a close election RDD to estimate the causal effect of holding political office on firm profitability with and without covariates using a 5% bandwith as recommended by CCT???? and the I-K optimal bandwidth estimated without covariates. The general form of the local linear regression estimated is:

$$\text{Firm Profits} = \alpha + \hat{\tau}(\text{District Win}) + \gamma\text{Margin} + \delta(\text{District Win} \times \text{Margin}) + X\beta + \mathbf{Y}_j + \mathbf{S}_j + \mathbf{R}_j \quad (8)$$

In Equation 8, the outcome variable is firm profit margins and the treatment indicator is whether the businessperson won election in their district and the running variable is the vote margin. In these analyses are also included a set of covariates X and year, sector and region fixed effects (Y, S, R). This regression is estimated around a threshold of the cutpoint $c \pm h^{full}$ where $c \pm h^{full}$ is determined through cross-validation. Define the original treatment effect of the full model (i.e. the model entered in Step 1 above), $\hat{\tau}(h^{full})$.

After selecting covariates via either Bayesian SS or LASSO through Steps 2-4, we are left with the model:

$$\text{Firm Profits} = \alpha + \check{\tau}(\text{District Win}) + \gamma\text{Margin} + \delta(\text{District Win} \times \text{Margin}) + X^o\beta^o + \mathbf{Y}_j + \mathbf{S}_j + \mathbf{R}_j \tag{9}$$

Note that the primary difference between the two equations above is the new set of covariates $X^o\beta^o$ which satisfies the condition $rank(X^o) \leq rank(X)$ through the removal of covariates and a new optimal bandwidth h^{optim} as a result of the addition of new covariates. We thus define the new adjusted treatment effect as $\check{\tau}(h^{optim})$. It can be shown that $Var[\check{\tau}(h^{optim})] \leq Var[\hat{\tau}(h^{full})]$, a result that is expanded upon in the Appendix. While coverage properties of this new estimator is less clear theoretically, results from CCT and others suggest that the regularization adjusted estimator will have superior coverage properties as well under a variety of circumstances. This is confirmed in a series of simulations below.

One important thing to note is that all fixed effects in the model were not regularized. This was done deliberately through setting the penalty parameters to 0 when estimating the LASSO model. While efficiency gains can theoretically be made through the exclusion of some fixed effects, this estimation strategy does not make substantive sense in any context.

Table 3 contains original and regularization adjusted treatment effects and standard errors. One thing of note is that the standard errors of all adaptive lasso treatment effects are smaller than those of the original covariate adjusted treatment effects published. As simulations below demonstrate, this is due to the oracle property enjoyed by the adaptive Lasso, which has been demonstrated produce "correct" model specification under a wide variety of conditions.

	Original (APSR)	Adaptive	Original 5% (APSR)	Adaptive 5%	Adaptive CCT Robust
District Win	0.146***	0.102*	0.198**	0.097**	0.140***
	(0.065)	(0.060)	(0.090)	(0.038)	(0.052)
Bandwidth	0.113	0.120	0.050	0.050	0.120
Covariates Dropped	*	4	*	2	4
Firm and Cand Covariates	Full	Select	Full	Select	Select
Region,Sector Year FE	Full	Full	No	No	No
Observations	481	520	201	201	520

Table 3 – Replication of Political Connections and Firm Profitability Analysis in Szakonyi (2018) with Adaptive LASSO Adjusted Treatment Effects.

5 Simulations

In order to explore the bias and coverage properties of the adaptive lasso method in as realistic a scenario as possible, I perform a series of simulations using the same election and profit data from Szakonyi (2018) described above but with a true simulated treatment effect. To accomplish this, I use the correlation matrix of the covariates and vote margin used by Szakonyi (2018) to construct 2,000 simulated data sets which have the same covariance structure and mean of the original dataset and set the true treatment effect τ_{RDD} to 0.30.

Define Ξ as a matrix which contains the set of covariates plus the vote margin used in Szakonyi (2018) discussed above. Furthermore, assume that the data generating process of Ξ is that of a multivariate normal distribution defined by some mean parameters $\mu = (\mu_1, \mu_2, \cdots, \mu_p)$ and a covariance matrix Σ. Thus:

$$\Xi \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

Using this data generating process along with empirically defined parameters μ and covariance structure Σ, I generate $s = 1, \cdots, 2000$ simulated data sets $\boldsymbol{\Xi}^s$ such that the d.g.p of each simulated dataset adheres to:

$$\boldsymbol{\Xi}^s \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$$

Through generating the data in this manner, we're insuring that each simulated dataset conforms to a realistic d.g.p in the context of a close-election RDD. For each simulation the outcome Y, and thus the true model, is thus defined by

$$Y_s = 0.3(District\ Win_s) + \gamma(Margin_s) + \delta(District\ Win_s \times Margin_s) + \eta_s$$

where the error term $\eta_s \sim N(0,1)$, the simulated vote share, $Margin_s$ is simulated as one of the variables within $\boldsymbol{\Xi}^s$ and $District\ Win_s = \mathbb{I}(Margin_s > 0)$ is a simulated forcing variable based on $Margin_s$. The true treatment effect that we estimate using the conventional RDD approach and adaptive lasso approach with is $\tau_{RDD} = 0.3$. Reported coefficient values, bandwidths and standard errors are CCT robust estimates using the standard and adaptive approaches.

The model estimated for each simulation is the full model including covariates:

$$\text{Y}_s = \alpha^s + \hat{\tau}^s_{RDD}(\text{District Win}_s) + \gamma^s \text{Margin}_s + \delta^s(\text{District Win}_s \times \text{Margin}_s) + X^s\beta^s + \epsilon^s \tag{10}$$

For the simulations, the average bias of $\hat{\tau}^s_{RDD}$, $SE(\hat{\tau}^s_{RDD})$ and % coverage of the confidence intervals were recorded for models in which the bandwidth was allowed to vary according to the adaptive lasso procedure outlined above or was fixed at a certain value with the adaptive lasso applied afterwards.

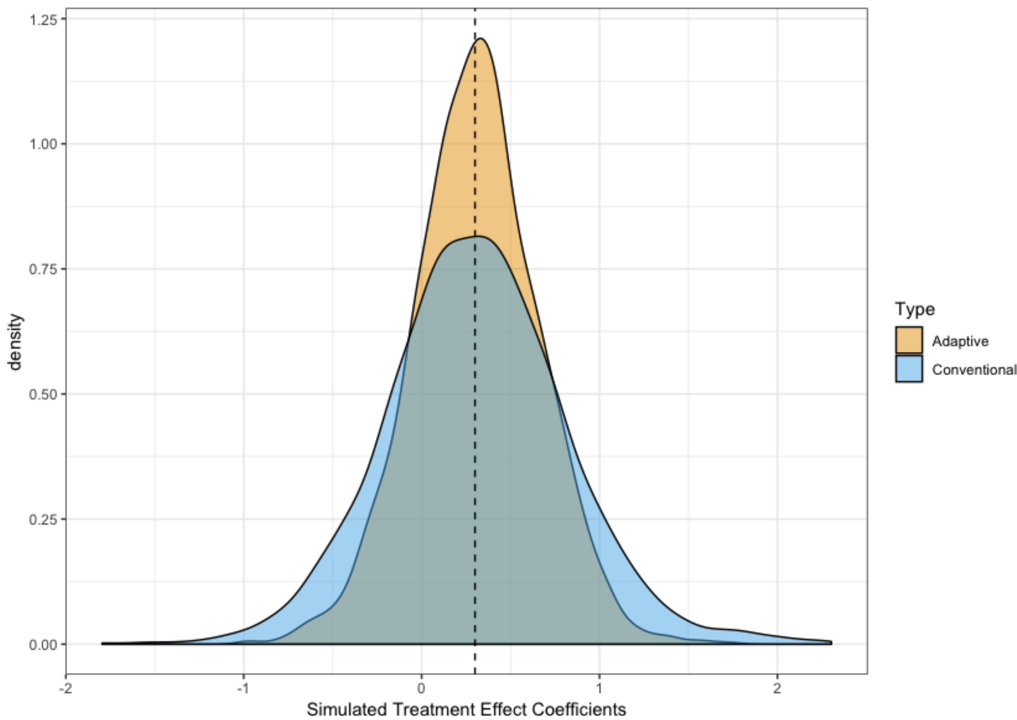

Figure 1 – Distribution of simulated treatment effects $\hat{\tau}^s_{RDD}$, for adaptive lasso adjusted treatment effects and conventional treatment effects across 2,000 simulated data sets with variable bandwidth select. The true $\tau_{RDD} = 0.30$ is denoted by the black dotted line.

Figure 1 contains the distribution of simulated treatment effects estimated using conventional and adaptive lasso methods. Here we see that the adaptive lasso restricts the treatment effects estimated to a much narrower band around the true treatment effect.

Variable Bandwidth*

	Adaptive	*Conventional*	*Difference (Adaptive - Conventional)*
τ_{RDD} Bias	0.274	0.397	- 0.123***
% Coverage	0.944	0.699	+ 0.245***
τ_{RDD} Estimate	0.308	0.308	-
Bandwidth	0.38	0.292	+ 0.088***

Fixed Bandwidth$^{\mp}$

	Adaptive	*Conventional*	*Difference (Adaptive - Conventional)*
τ_{RDD} Bias	0.375	0.375	- 0.001
% Coverage	0.931	0.796	+ 0.135***
τ_{RDD} Estimate	0.300	0.300	- 0.001
Bandwidth	0.200	0.200	-

Table 4 – Performance of Adaptive Lasso v. Conventional Treatment Effect Estimates in Simulations$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$ for t-test of mean difference with $H_0 : \mu_{Adaptive} = \mu_{Conventional}$. Average simulation results across 2,000 simulations comparing "Adaptive" vs. "Conventional" treatment effect bias and coverage results. Final bias and coverage results are both estimated using CCT robust point estimates and confidence intervals. * "Variable bandwidth" results are produced through Imbens-Kalyanaraman optimal bandwidth selection based on models selected by the adaptive algorithm described above or the full model mentioned in this section. ∓ For fixed bandwidth simulations, bandwidth was set to 0.20.

Table 4 contains estimates of the bias, % coverage and other statistics from the simulation. The adaptive lasso here provides some very striking efficiency improvements which are reflected in the % coverage estimates in both variable and fixed bandwidth selection procedures. In the variable bandwidth scenario, the adaptive lasso combined with CCT robust estimation produces confidence intervals on treatment effects that achieves an average of 94% coverage versus 70% coverage under conventional estimation while under the fixed bandwidth scenario, adaptive lasso estimation achieved 93% coverage compared to about 80% coverage under conventional estimation. Each of these differences was statistically significant at the $p < 0.01$ level.

6 Discussion

In this paper we have demonstrated that our algorithm which employs the adaptive LASSO can improve the efficiency of treatment effects for RDDs estimated with covariates and provide a principled framework of treatment effect adjustment for RDDs. As we emphasize above, however, this does not imply that substantive considerations in the estimation process should be abandoned and replaced by automated machine learning methods. To the contrary, substantive considerations, as reflected in the algorithm that we developed above, are and should always be at the forefront of model estimation whether in the context of RDDs or estimation strategies.

References

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu. 2016. "Lasso adjustments of treatment effect estimates in randomized experiments." *Proceedings of the National Academy of Sciences* 113 (27): 7383–7390.

Calonico, Sebastian, Matias D Cattaneo, Max H Farrell, and Rocıo Titiunik. 2016. "Regression discontinuity designs using covariates." *URL http://www-personal. umich. edu/˜ cattaneo/papers/Calonico-Cattaneo-/Farrell-Titiunik_2016_wp. pdf.*

Calonico, Sebastian, Matias D Cattaneo, Max H Farrell, and Rocio Titiunik. 2018. "Regression discontinuity designs using covariates." *Review of Economics and Statistics* (0).

Calonico, Sebastian, Matias D Cattaneo, Max H Farrell, and Rocio Titiunik. 2019. "Regression discontinuity designs using covariates." *Review of Economics and Statistics* 101 (3): 442–451.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust non-parametric confidence intervals for regression-discontinuity designs." *Econometrica* 82 (6): 2295–2326.

Caughey, Devin, and Jasjeet S Sekhon. 2011. "Elections and the regression discontinuity design: Lessons from close US house races, 1942–2008." *Political Analysis* 19 (4): 385–408.

Eggers, Andrew C, and Jens Hainmueller. 2009. "MPs for sale? Returns to office in postwar British politics." *American Political Science Review* 103 (4): 513–533.

Erikson, Robert S, and Kelly Rader. 2017. "Much ado about nothing: rdd and the incumbency advantage." *Political Analysis* 25 (2): 269–275.

Fisman, Raymond, Florian Schulz, and Vikrant Vig. 2014. "The private returns to public office." *Journal of Political Economy* 122 (4): 806–862.

Frölich, Markus. 2007. "Regression discontinuity design with covariates." *University of St. Gallen, Department of Economics, Discussion Paper* (2007-32).

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.

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.

Imai, Kosuke. 2011. "Introduction to the Virtual Issue: Past and Future Research Agenda on Causal Inference." *Political Analysis* 19 (V2): 1–4.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal bandwidth choice for the regression discontinuity estimator." *The Review of economic studies* 79 (3): 933–959.

James, W, and C Stein. 1961. "Proc. Fourth Berkeley Symp. Math. Statist. Probab." In *Estimation with quadratic loss.* Vol. 1. Univ. California Press.

Rubin, Donald B. 2005. "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association* 100 (469): 322–331.

Skovron, Christopher, and Rocıo Titiunik. 2015. "A practical guide to regression discontinuity designs in political science." *American Journal of Political Science*: 1–47.

Stein, Charles. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, . Berkeley, Calif. pp. 197–206. https://projecteuclid.org/euclid.bsmsp/1200501656.

Szakonyi, David. 2018. "Businesspeople in Elected Office: Identifying Private Benefits from Firm-Level Returns." *American Political Science Review* 112 (2): 322–338.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*: 267–288.

Tibshirani, Robert, Martin Wainwright, and Trevor Hastie. 2015. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Tihonov, Andrei Nikolajevits. 1963. "Solution of incorrectly formulated problems and the regularization method." *Soviet Math.* 4: 1035–1038.

Truex, Rory. 2014. "The returns to office in a rubber stamp parliament." *American Political Science Review* 108 (2): 235–251.

Zou, Hui. 2006. "The adaptive lasso and its oracle properties." *Journal of the American statistical association* 101 (476): 1418–1429.