Summary of Master Thesis: "Covariate Selection via Lasso in the Regression Discontinuity Design"

Julian Streyczek
August 2020

1 Introduction

Regression discontinuity (RD) designs are becoming increasingly popular in economics and other social sciences for estimating causal effects from observational data. However, these designs can suffer from low power because treatment effects are typically estimated with few observations. While including control variables can alleviate this problem, in practice data are often high-dimensional, so that standard methods break down without a way to properly select covariates.

In this project, I address this problem by developing a data-driven approach for covariate selection in RD designs. Estimation works in the following three-step procedure: First, choose preliminary observation weights without covariates, second, select a subset of covariates via Lasso, and third, estimate the treatment effect using only the selected covariates. The approach works with any number of covariates, requires minimal tuning by the researcher, and is easily implementable using standard methods.

I contribute to a growing literature on the usage of covariates in RD designs (Frölich and Huber 2019; Calonico et al. 2019). My approach extends the latter to arbitrarily large sets of covariates. It is similar to Anastasopoulos (2020), but explores in detail the performance of different implementation choices. The results contributed to Kreiss and Rothe (2023), who study the theoretical properties of a generalized approach to Lasso-based covariate selection for RD designs.

2 Approach

The following three-step procedure allows estimating RD treatment effects with any number of control variables.

Step 1: Select a candidate bandwidth, for example the MSE-optimal bandwidth developed by Calonico et al. (2014). The bandwidth defines weights that determine the degree to which each observations contributes to the estimation (depending on their distance to the discontinuity).

Step 2: Perform selection of covariates via Lasso, using the weights obtained in Step 1, as follows:

$$arg \min_{(\alpha,\beta,\gamma',\tau)} \sum_{i=1}^{n} (Y_i - \alpha - f(\beta, \tau, W_i, X_i) - \gamma' Z_i)^2 K(X_i/h) + \lambda \sum_{j=1}^{d} |\gamma_j|,$$

The first sum represents the standard OLS minimization problem for RD designs (including kernel weighting), while the second sum adds a lasso penalty on the covariates. In particular, Y_i denotes the outcome for observations $1, ..., n$, X_i is the running variable, $f(.)$ denotes the functional form of the RD regression, and Z_i is a d-dimensional vector of covariates. The bandwidth h is estimated in Step 1. The penalization parameter λ is selected via cross-validation.

Step 3: Estimate the treatment effect τ using only the selected variables, for example using Calonico et al. (2014).

While a formal derivation of the necessary assumptions is out of the scope of this project, for best performance the standard assumptions for RD and Lasso should be satisfied. Most notably, RD requires continuity of $E[Y_i|X_i, Z_i]$ and no manipulation of the running variable X_i, while Lasso requires sparsity, meaning that only few variables are correlated with the outcome.

3 Simulations

I verify the performance of this approach in simulations similar to Imbens and Kalyanaraman (2012). I consider a running variable distributed $X_i \sim (2 * Beta(2,4) - 1)$, a vector of i.i.d. covariates $Z_i \sim N(0, 0.1295^2)$, and i.i.d. errors $\epsilon_i \sim N(0, 0.1295^2)$. The outcome Y_i is defined

$$E[Y_i|X_i = x, Z_i = z] = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 + \gamma' z, & \text{if } x < 0 \\ 0.52 + 0.84x - 3x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 + \gamma' z, & \text{if } x \geq 0 \end{cases},$$

so that the treatment effect is $\tau = 0.04$. The coefficients on the covariates are $\gamma_j > 0$ if $j \leq 5$ and 0 otherwise, so that the covariates are sufficiently sparse. I set $n = 1000$.

Table 1 reports results from 1000 estimations of τ using linear regression and a triangular kernel. The first row reports results without any covariates, while the second shows results under an ideal scenario in which only the five relevant covariates are included. In my setting, including covariates improves mean squared error by 33%.

Rows 3 and 4 report results with lasso-based selection of covariates in its preferred specification (as described above). The estimates differ only in the criterion applied to select the penalization parameter λ.[1] Both approaches achieve close to optimal performance, with improvements in MSE over the baseline of around 25%.

In rows 5 and 6, I report results for a "naive" selection step, in which the Lasso regression does not control for the functional form of the RD regression. In my thesis, I explain in detail that this performance is undesirable because it introduces negative bias due to a type of overfitting.

Table 1: Simulation results

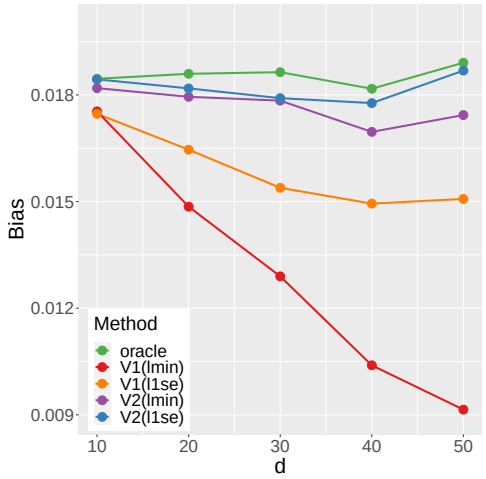
	h	N^-	N^+	λ	$covs$	$bias(\hat{\tau})$	$sd(\hat{\tau})$	$mse(\hat{\tau})$
No covariates	0.197	147.5	98.7	-	-	0.0203	0.0406	0.00206
Ideal covariates	0.184	136.4	93.4	-	5.0	0.0189	0.0321	0.00139
Lasso: Preferred, parsimonious	0.180	133.0	91.7	0.015	6.8	0.0187	0.0337	0.00148
Lasso: Preferred, generous	0.167	122.2	86.6	0.007	16.2	0.0174	0.0348	0.00152
Lasso: Naive, parsimonious	0.182	134.9	92.7	0.019	6.5	0.0151	0.0325	0.00128
Lasso: Naive, generous	0.172	126.6	88.9	0.009	15.9	0.0091	0.0324	0.00113

Notes: Results from estimating the local average treatment effect $\hat{\tau}$ 1000 times via cross-validated lasso selection. Rows 1-2 correspond to estimation without any, and with only the relevant covariates, respectively. Rows 3-4 report results for the preferred version outlined above, while rows 5-6 report results for a naive version that omits the functional form of the RD in the selection step. "Generous" and "parsimonious" selection based on criterion for selecting penalization parameter λ.

[1] The "generous" criterion chooses λ_{min} that minimizes the cross-validation error. In practice, such selection is often too generous, so that I consider a "parsimonious" alternative that chooses the largest λ for which the cross-validation error is no larger than the error under λ_{min} plus one standard deviation (Hastie et al. 2015).

Figure 1 explores stability of my approach by repeating the estimation for varying numbers of covariates. The green line refers to inclusion of ideal covariates, the blue and purple lines refer to the preferred specification, and red and orange refer to the naive specification. The preferred Lasso specification is stable for different dimensionality of covariates, while the performance of the naive specification deteriorates with dimensionality.

Figure 1: Bias by number of covariates



Notes: Bias from estimating the local average treatment effect $\hat{\tau}$ 1000 times via cross-validated lasso, for different number of covariates. *Oracle* includes only relevant covariates, *V1* and *V2* are the naive and preferred Lasso specification, and *lmin* and *l1se* denote generous and parsimonious variable selection, respectively.

In the full thesis, I explain in detail the reasons behind the poor performance of the naive Lasso by estimating the bias it introduces. Moreover, I show that my preferred approach performs well under more realistic data generating processes. I also show that iterating bandwidth and covariate selection brings only little improvements. Finally, I explore an alternative approach of choosing the penalization parameter λ by minimizing the estimated variance of the treatment effect instead of the prediction error of the explanatory variables.

4 Conclusion

In this project, I conduct simulation experiments to demonstrate that lasso-based covariate selection can improve estimation in the RD designs. I provide an intuitive approach that easily applicable to a wide range of empirical settings.

References

Anastasopoulos, L. J. (2020). "Principled estimation of regression discontinuity designs". *arXiv:1910.06381*.

Calonico, Cattaneo, and Titiunik (2019). "REGRESSION DISCONTINUITY DESIGNS USING COVARIATES".

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

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

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

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

Kreiss, A. and C. Rothe (2023). "Inference in regression discontinuity designs with high-dimensional covariates". *The Econometrics Journal* 26.2, pp. 105–123.