The Minimax Concave Penalty (MCP) Variable selection regularization method for Regression Discontinuity Designs

Bahr kadhim Mohammed , Ashwaq Abdul Sada Kadhim

Department of Statistics, University of Al-Qadisiyah, Iraq.

E-mails: bahr.mahemmed@qu.edu.iq ashwaq.abdul.sadah.kazem@gmail.com

ABSTRACT

The classical method faced a big problem with estimating and selecting important variables when the dataset has a cut-off point. Therefore, we propose a new method to solve these problems. In this paper we suggested a new approach by combining the Regression Discontinuity Designs (RDD) with the Minimax Concave Penalty (MCP) method. Local linear regression (LLR) method was used to estimate the effect of processing on the cut-off region of the observations within the optimum bandwidth selection for the RDD design to obtain the best model. Three models were used to determine the IK (Iembens and kalyanman) bandwidth, cross-validation (CV) method, and The CCT (Calonico, Cattaneo & Titiunik) bandwidth. A simulation study and real data are conducted to investigate the performance of the proposed method. The mean squared errors (MSE) is used to choose the best model.

Keywords: Regression Discontinuity Designs (RDD), Minimax Concave Penalty (MCP), variable selection, Local linear regression, bandwidth selection, IK, CV, CCT.

1- Introduction

RDD is a quasi-experimental pretest and posttest, design that extract the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is unfeasible. Variable selection methods are one of the well-sophisticated field in the modern statistics. In this study, we deal with one of the most commonly used models in this area, which is the Regression-Discontinuity Design model. First applied by (Donald Thistlethwaite and Donald Campbell (1960)) to the evaluation of scholarship programs, the RDD has become increasingly popular in recent years.

In regularization methods the Variable Selection (V.S) is implemented with the process of the parameter estimation. Examples of regularization approaches are the Lasso (Tibshirani, 1996) explained that ridge regression and lasso regression, each method in which the penalty is

applied to each additional variable added to the OLS, , Elastic Net (Zou and Hastie, 2005) which combined Ridg's penalty and Lasso's penalty with a "group lasso " used for the purpose of selecting a large set of covariates. , adaptive Lasso (Zou, 2006) developed adaptive Lasso method for the purpose of maximizing the selection of the correct variable to solve problems of estimating low and high dimensions , group Lasso (Yuan and Lin, 2006), MCP (Zhang, 2010) that estimates and selects linear regression variables simultaneously using the MCP penalty function, overcomes the Lasso method in terms of its inconsistency in the selection of variables. (Anastasopoulos, L. J. (2019)) employ method adaptive lasso with RDD model. .

In this paper, we will employ one of the variable selection methods, which is MCP method with RDD, where use local linear regression (LLR) in the cut-off region of the observations within the optimal bandwidth range chosen for the RDD either side of the cut-off point $F_i \in$ (c - h , c + h) where h bandwidth and (c) the cut-off point to obtain the lowest MSE. Three models were used to select the bandwidth, The IK method proposed by (Imbens and Kalyanaraman, (2009)), The Cross-validation (CV) approach proposed by (miller and Ludwig, (2007)), the CCT method was proposed by (Calonico et al. (2014) where the MSE criterion was adopted to compare the proposed method and some previous methods, where we used this criterion to determine the performance of those methods.

This paper is organized as follows: We have been shown basics about discontinuity regression designs and Local Linear Regression (LLR) in Section 2 ; In section 3, basics about Bandwidth Selection and some methods that were used by the researcher has been presented ; In section 4 ,we have displayed a method for selecting a variable using MCP method ; In section 5 we have explained the selection of the variable by combining each of the MCP penalty function and model (RDD). In Section 6 we have summarized the results of the simulation study and present the data for the sample analysis. A brief conclusion has been included in Section 7.

---

2 . Regression-Discontinuity Design (RDD)

RDD model is divided into two groups on the basis of a specific threshold limit or the so-called breakpoint (Thistlethwaite and Campbell, 1960). This point is determined in advance according to the study conditions and requirements. The importance of calling it a discontinuity design (RDD) comes from the fact that the treatment effect will lead to a 'jump' or discontinuity 'in the regression function point of the relationship between Classification variable (F_i) (an explanatory variable) in a discontinuity design with response variable Y_i (Lee, D. S., & Lemieux, T., 2010) .

When estimating RDD, covariates should be included before treatment, for the purpose of obtaining the most accurate treatment effect estimates (Bloniarz et al. 2016; Calonico et al. 2018). The most important part of the accuracy depends mainly on the bandwidth, or on the low variance in the model, and it may be due to both. Making preliminary decisions regarding the covariates variables that must be included before performing a treatment should always be based on expert judgment and the researcher's expectations that are closely related to the problem at hand (Frölich and Huber (2019).

The simplest method to estimate the treatment effect is by using local linear regression (LLR) in the cut-off region of the observations within the optimal bandwidth range chosen for the RDD method on either side of the cut-off point $F_i \in$ (c - h, c + h) where (c) denotes the cutoff point, (h) denotes the bandwidth, to obtain the lowest MSE. Three models were used to select the bandwidth, The IK method proposed by (Imbens and Kalyanaraman, (2009)), The Cross-validation (CV) approach proposed by (miller and Ludwig, (2007)) and the CCT method was proposed by (Calonico et al. (2014)) .

$$Y_i = \alpha + \tau T_i + \gamma F_i + \delta(F_i . T_i) + XB + \epsilon_i \quad \ldots . (1)$$

where:

α = the average value of the outcome for those in the treatment group after controlling for the rating variable. $\hat{\tau}$: Estimated local average treatment effect. F_i : The forcing variable or rating variable for observation i, centered at the cut-point. T_i: An indication whether or not to receive treatment:

$$T_i = \begin{cases} 1 & \text{receiving treatment} \\ 0 & \text{Not receiving treatment} \end{cases}$$

γ: Coefficient of the forcing variable. $f(T_i, F_i)$: It is a function of the force variable which is in the form of a nonparametric kernel or a polynomial of order p^{th}. δ: Coefficient $(F_i . T_i)$, β: Coefficient vector $\beta_{p\times 1}$, X : Matrix of covariates $(n \times P)$. (Anastasopoulos, J. (2019)).

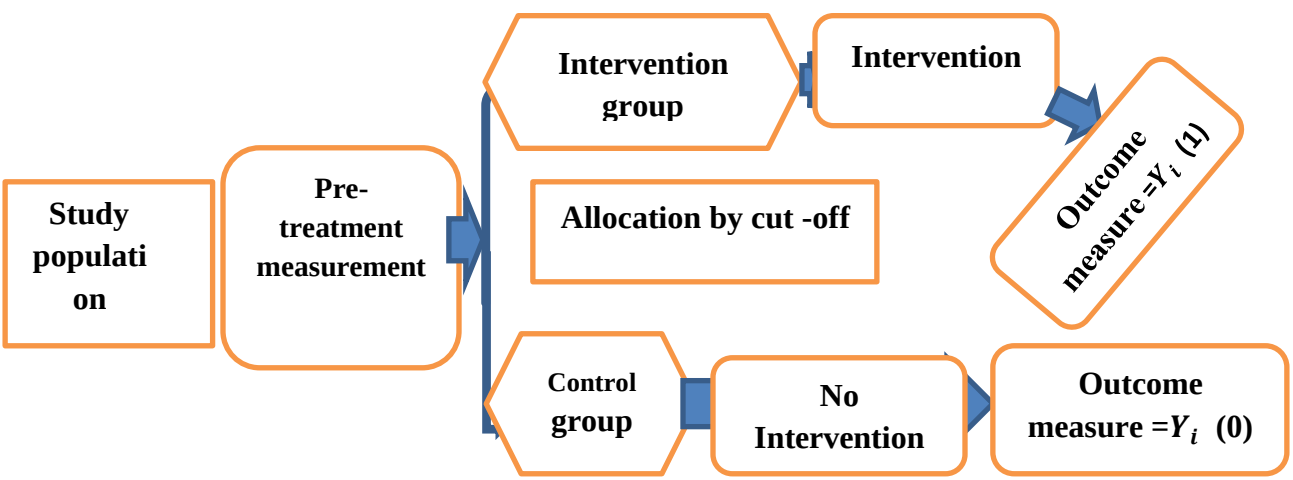

Figure (1): Regression discontinuity design diagram (RD)

2-1 Local Linear Regression (LLR)

Local linear regression is a non-parametric method that is used to continuously estimate the treatment effect in RDD model (Hahn et al.,2001; Porter, 2003; Imbens and Lemieux,2008). Neighborhood idea ***h*** is the basis of the LLR. Where ***h*** bandwidth is chosen. In this method, points within the radius ***h*** of x_0 are determined. Points near x_0 are given greater weights than those further away from x_0 . The average weight to weigh the adjacent observed data by the kernel which is a statistical technique for estimating the reality of the function. The kernel function K (u): R → R, , has the following properties (Mutair, Hafez Muhammad, 2011).

1. $0 \leq k(u) \leq \infty$, $K(u)$ is a continuous function with non-negative real values
2. $\int_{-\infty}^{\infty} K(u)du = 1$

3. K(u) is a symmetric function around zero, $\int u\, K(u)\, du = 0$

$$and\ \sigma^2 = \int u^2\, K(u)\, du > 0$$

3. Bandwidth Selection

Bandwidth is an unrestricted parameter (Free parameter) that has a clear role in the estimation process as it greatly affects bias and variance, as the more bandwidth increases, the bias increases and the variance decreases and vice versa, and as a result it will have a clear effect on smoothing the curve and the rate of its approach to the original curve. (Imbens, G. W., & Lemieux, T., 2008). The basic idea of choosing a bandwidth in the SRD is a trade-off between bias and variance for ($\hat{\tau}_c^{SRD}$) where the greater the bandwidth, the greater the bias and the less variance. There are several methods for choosing the optimal value of the bandwidth that have been used by many researchers such as cross validation and plug-in method, and many other methods (Hill, R. Carter and Kang-sun Lee., 2001). The methods that were used by the researcher will be covered in this research:

3.1 IK method

The IK method was suggested by (Imbens and Kalyanaraman, 2009). The researchers explained that the optimal choice of the Bandwidth Optimal is by substituting the six unknown quantities shown in the equation below, which will ultimately lead to the consistent estimators.

$$h_{opt}^{IK} = \arg\min_h AMSE\ (h) = CK\ .\left(\frac{\sigma_+^2(c) + \sigma_-^2(c)}{f(c).\ (m_+^{(2)}\ (c) - m_-^{(2)}\ (c))}\right)^{1/5} .\ N^{-1/5} \ldots (2)$$

When obtaining the six unknown estimators in equation (2), the optimum bandwidth estimate is according to the following formula:

$$\hat{h}_{opt}^{IK} = \arg\min_h AMSE(h) = CK\left(\frac{\hat{\sigma}_+^2(c) + \hat{\sigma}_-^2(c)}{\hat{f}(c).(\hat{m}_+^2\ (c) - \hat{m}_-^2\ (c))}\right)^{1/5} N^{-1/5} \ldots (3)$$

3.2 The Cross-validation method

The Cross-validation (CV) approach proposed by (miller and Ludwig, 2007). This method is considered one of the best and most used methods of selecting the bandwidth, and it is called the method (leave – one - out) in which one observation is excluded from the values of the observations, as it is the main part of the process of balance between both the variance and the bias, as the more the variance value decreases, the value of the bandwidth increases and the bias value begins to increase. The package width that has the lowest value for the Cross-validation criterion (CV) is chosen according to the following formula:

$$h_{CV} = \min_{h>0} CV_Y(h) \quad \ldots\ldots \quad (4)$$

3.3 The CCT method.

The CCT method was proposed by (Calonico et al.(2014)). We estimate the bounds of (asymptotic variance) by finding the initial bandwidth (Vn, Cn) denoted by (V) where:(Calonico et al.(2014)).

$$\hat{V} = 2.58\, \omega\, n^{-\frac{1}{5}} \quad \ldots\ldots (5)$$

where:

$$\omega = min\left\{ S_x , \frac{IQR_x}{1.349} \right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

(S_x) Denotes the sample variance, (IQR_x) indicates the interquartile range, and the bandwidth ($\hat{C}_n$) where :

$$\hat{C}_n = \begin{pmatrix} 1/(2q+5 \\ v \;,\; p \;,\; q \end{pmatrix} . \; n^{-1/(2q+5)} \quad \ldots (7)$$

We find the bandwidth ($\hat{b}_{CCT}$) and it is calculated according to the following formula(Ali, O. A. et al. (2020)):

$$\hat{b}_{CCT} = \begin{pmatrix} 1/(2q+5 \\ 0 \;,\; 1 \;,\; 2 \end{pmatrix} \quad \ldots\ldots\ldots\ldots (8)$$

$$\hat{C}_{0,1,2} = \frac{5n\, \hat{v}_n^5 \;, \hat{V}_{2,2}\; (\hat{v}_n)}{2\, \beta_{2,2}^2 \{ \left(é_3\; \hat{\beta}_{+,3}\; (Cn) + é_3\; \hat{\beta}_{-,3}\; (\hat{C}_n) \right)^2 + 3\hat{V}_{3,3}\, (\hat{C}_n) \}} \ldots\ldots (9)$$

We find the basic bandwidth ($\hat{h}_{CCT}$) according to the following formula:

$$\hat{h}_{CCT} = \begin{pmatrix} 1/5 \\ 0 \;, 1 \;, 0 \end{pmatrix} . \; n^{-1/5} \quad \ldots\ldots (10)$$

$$\hat{C}_{0,1,0} = \frac{n\, \hat{v}_n^5 \;, \hat{V}_{0,1}\; (\hat{v}_n)}{4\, \beta_{0,1}^2 \{ \left(é_2\; \hat{\beta}_{+,2}\; (\hat{b}_{CCT}) + é_2\; \hat{\beta}_{-,2}\; (\hat{b}_{CCT}) \right)^2 + 3\hat{V}_{3,3}\, (\hat{C}_n) \}} \ldots (11)$$

3. Variable selection in the Minimax Concave Penalty (MCP) method

The Minimax Concave Penalty (MCP) is another alternative to get less biased regression coefficients in sparse models. Zhang (2010) proposed the MCP method, that estimates and selects linear regression variables simultaneously using the MCP penalty function, overcomes the Lasso method in terms of its inconsistency in the selection of variables. The MCP estimator is obtained by the following formula: (Choon, C. L. (2012)

$$\hat{\beta}_j^{MCP} = arg \min_p [\|Y - X\beta\|^2 + \sum_{j=1}^{p} P_{\lambda,\gamma}^{MCP}] \quad .... \ (12)$$

Where: $\sum_{j=1}^{p} P_{\lambda,\gamma}^{MCP}$ the MCP penalty function. .
The MCP function takes the following form:

$$p_{\lambda}(|\beta|) = \begin{cases} \lambda \left(|\beta| - \frac{|\beta|^2}{2\lambda\gamma} \right), & |\beta| < \lambda\gamma \\ \frac{\lambda^2\gamma}{2}, & |\beta| \geq \lambda\gamma \end{cases} \quad .... (13)$$

Where : $\gamma > 1$ (Breheny, P. (2016)).

Many concave penalties depend on λ , as well as include a tuning parameter (γ) that controls the concavity of the penalty (i.e. how quickly the penalty decreases) . It should be noted that the MCP function has an interval of values on which all estimates are flat - across this region, estimates are the same as those for least squares regression . The adaptive lasso and (MCP) methods differ from lasso in that they allow the estimated coefficients to reach large values more quickly than lasso, since all of these methods shrink most of the coefficients towards zero, but these two methods (adaptive lasso and MCP) operate on apply less shrinkage to non-zero coefficients; this indicates less bias.

4 . MCP with RDD Model

In this paper, we will employ the method (MCP) with model (RDD) for the purpose of estimating and selecting the variable by integrating the MCP penalty function with the model (RDD) according to the following formula:

$$arg\ min_{\Theta} \sum_{i=1}^{N} [Y_i - (\alpha + \tau T_i + \gamma F_i + \delta (F_i, T_i) + X\beta)]^2 + \sum_{j=3}^{p} P_{\lambda,\gamma}^{MCP} \quad .... (14)$$

Θ = (τ, γ, δ, β) represents the vector of the estimated coefficients in RDD.

5. Application

5.1 SIMULATION STUDY :

Step1: A sample was generated in the following sizes (50, 100, 150, 250) and p = 50 variables include (s) nonzero variables. That is mean, null variables are p-s.

Step2: Correlations have been formed between the variables from 1 to k; (1 ,…,k). Where k represents the number of variables related to the amount of correlation (r = 0.75).

Step3: Two types of variables were created, where the first type is a treatment variable. It has been generated according to a uniform distribution (Anastasopoulos, J. (2019)). Sample size n with terms a = -1 represents the minimum and b = 1 represents the upper bound. With a parameter value of (10 , 2) (treatment variables and the two treatment parameters).The variables of the second type $\acute{x}\beta$ (Szakonyi, David. (2018)) (the rest of the variables) were generated according to the normal distribution with a parameter vector μ of (zero) and with degree $1 \times p$ and a common variance matrix of sigma (σ_{ij}) of degree ($p \times p$) where the main diameter elements of this matrix are (1) As for the rest of the elements, it is equal to (Rou) when i ≠ j and that i, j is less than k where $i, j < k$)) and zero when ($i, j > k$).

Step4: The random error term (e) was generated according to the standard normal distribution N (0,1), and the data were generated based on the following model (TrBeta). and repeat each experiment (IT=1000) for all of the simulation experiments.

Step 5: Calculate the MSE.

Example 1: Samples size (n=50,100,250) , number of variables (p=15) ,(s=5),(p-s=10) and $\rho = 0.75$. where

$$\beta = (\underbrace{0.5,1,1.5,2,3}_{S},\underbrace{0,0,\dots,0}_{P-S})$$

Table 1: MSE values for methods of study for n=100, p= 15, s=5 and ρ =0.75 .

Table (1)

Methods	Cut off point	IK	CCT	CV
ad lasso	0.0	0.0528	0.0486	0.0498
MCP		0.0468	0.0450	0.0466
ad lasso	0.5	0.0620	0.0520	0.0533
MCP		0.0523	0.0422	0.0470
ad lasso	2	0.0734	0.0621	0.0640
MCP		0.0559	0.0442	0.0472

In this example **1**, From Table 1 with n = 100, p= 15, s=5 and ρ =0.75 , we notice the superiority of our suggested method (MCP) over the adaptive lasso through MSE values. In addition, it's clear to see that the best method of the bandwidth is CCT at all of cut- off point.

Although both methods have the advantage that they apply less shrinkage to non-zero operands, this indicates less bias. Theoretical results, simulations show that the MCP function is a penalty function to be reckoned with.

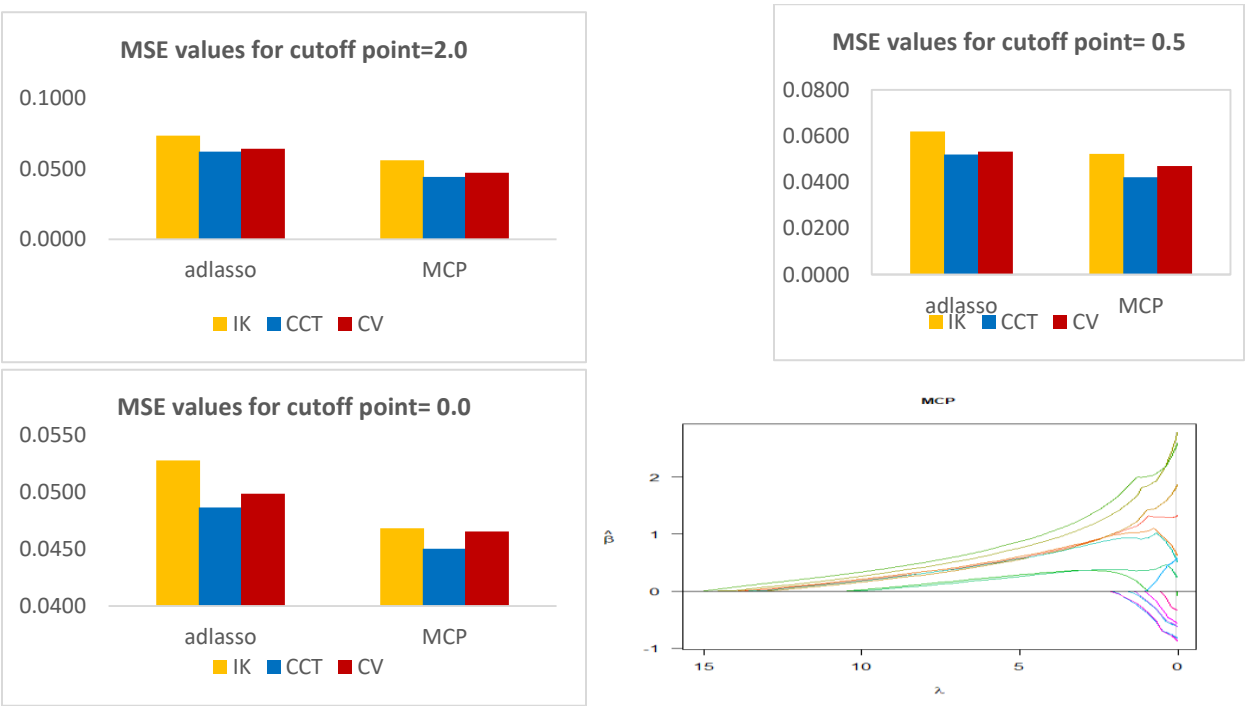

Figure 2: MSE values for methods of study for n=100, p= 15, s=5 and ρ =0.75 .

Figure (2) shows three cases of cut-off point (0.0, 0.5 and 2) for the preference of our suggested method over the adaptive lasso method in the RDD model. according to the (MSE) criterion Although both methods have the advantage that they apply less shrinkage to non-zero operands, this indicates less bias, we note that the number of important variables appeared far from zero, as is true coefficients that assumed by the simulation and the figure that shows the features in the method MCP.

Example 2: Samples size (n=100) , number of variables (p=25) ,(s=10),(p-s=15) and $\rho = 0.75$. where

$$\beta = (\underbrace{0.5,1,1.5\,,2,3}_{S}\,,\underbrace{0,0,\ldots,0}_{P-S})$$

Table 2: MSE values for methods of study for n=100, p= 25, s=10 and ρ=0.75

methods	Cut off point	IK	CCT	CV
ad lasso	0.0	0.1053	0.0996	0.1015
MCP		0.1049	0.0896	0.0951
ad lasso	0.5	0.1300	0.1142	0.1248
MCP		0.1305	0.1045	0.1152
ad lasso	2	0.1699	0.1519	0.1506
MCP		0.1276	0.1011	0.1116

In this example **2**, From Table 2 with n= 100, p= 25, s=10 and ρ =0.75, we notice the superiority of our suggested method (MCP) over the adaptive lasso through MSE values. We also note that the best method of the bandwidth is CCT.

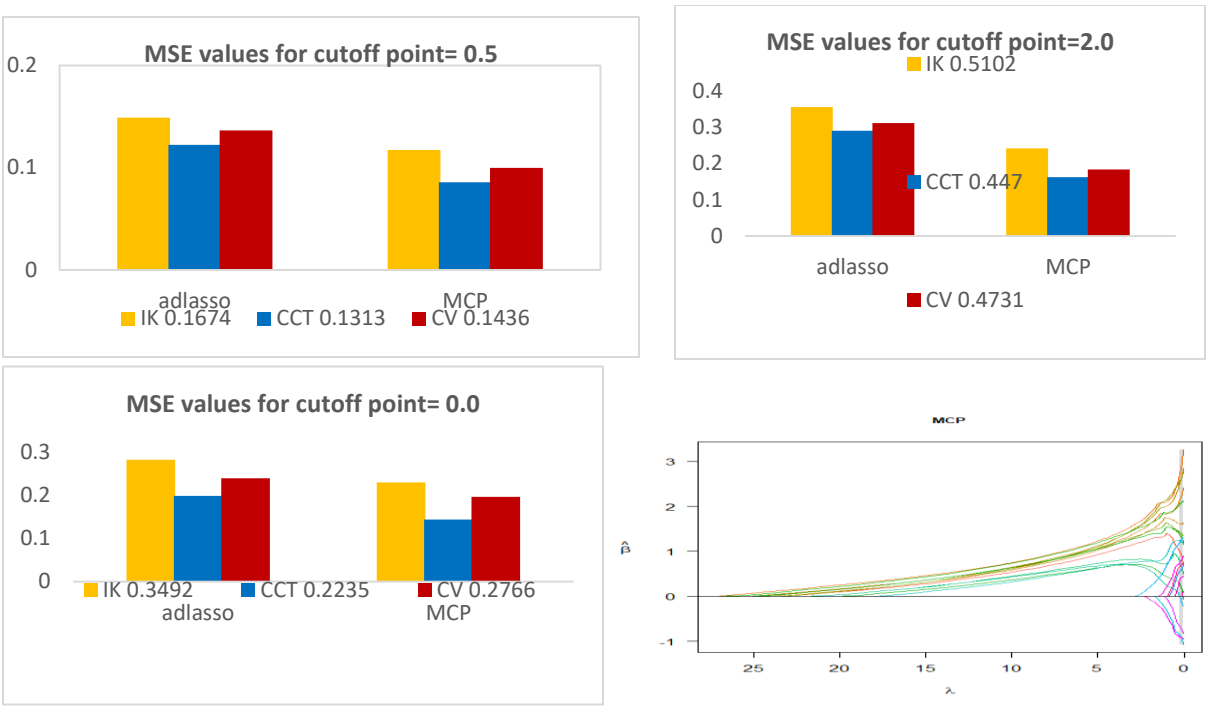

Figure 3: MSE values for methods of study for n=100, p= 15, s=5 and ρ =0.75 .

Figure (3) shows the preference of our suggested method over the adaptive lasso method for RDD model. According to the (MSE) values . Although both methods have the advantage that they apply less shrinkage to non-zero operands, this indicates less bias, we note that the number of important variables appeared far from zero, as is assumed by the simulation through and the figure that shows the features in the method MCP.

Example 3: Samples size (n=100) , number of variables (p=50) ,(s=20),(p-s=30) and $\rho = 0.75$. where

$$\beta = (\underbrace{0.5,1,1.5\,,2,3}_{S}\,,\underbrace{0,0,\ldots,0}_{P-S})$$

Table 3: MSE values for methods of study for n=100, p= 50, s=20 and ρ =0.75

methods	Cut off point	IK	CCT	CV
ad lasso	0.0	0.2834	0.1991	0.2398
MCP		0.2307	0.1443	0.1974
ad lasso	0.5	0.1488	0.1223	0.1364
MCP		0.1173	0.0858	0.0999
ad lasso	2	0.3554	0.2902	0.3112
MCP		0.2421	0.162	0.1837

In the example 3, From Table 3, with n= 100, p= 50, s=20 and ρ =0.75 , We notice the superiority of our suggested method (MCP) over the adaptive lasso. and that is through MSE, we also note that the best method of the bandwidth is CCT for all cut- off points.

Figure 4: MSE values for methods of study for n=100, p= 15, s=5 and ρ =0.75

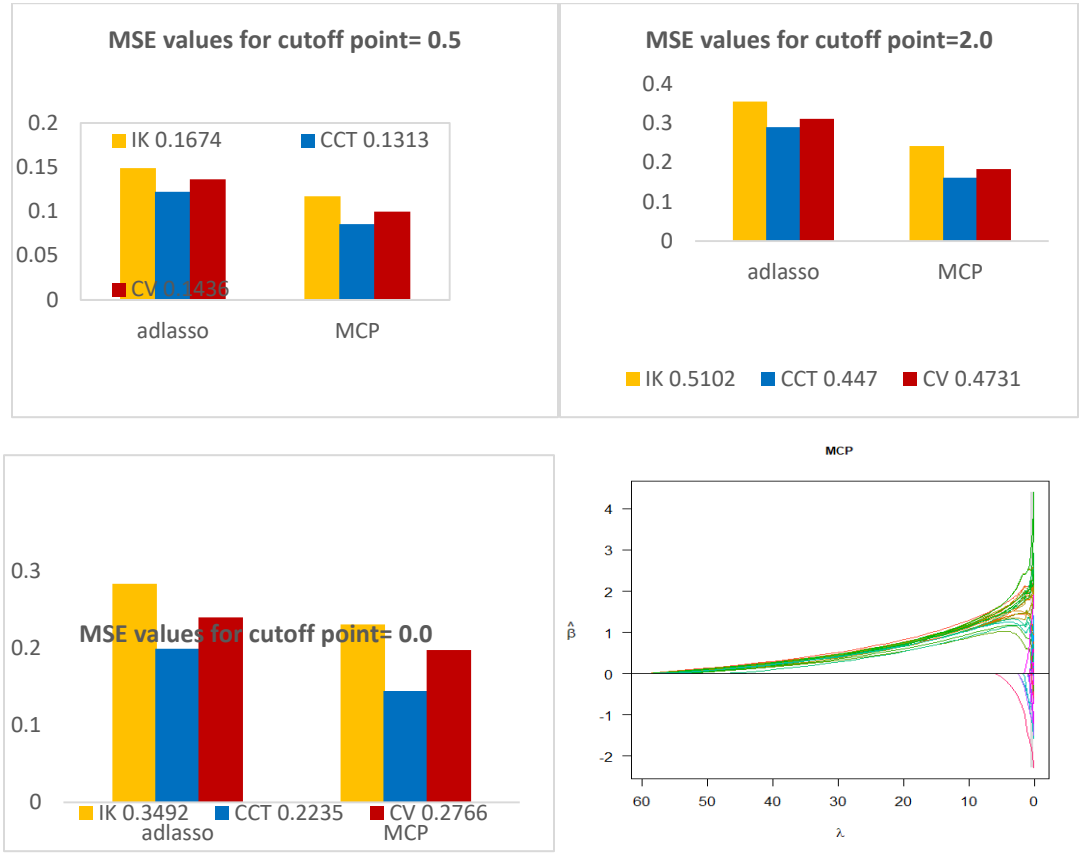

Figure 4: MSE values for methods of study for n=100, p= 15, s=5 and ρ =0.75 .

Figure (4) shows the preference of our suggested method over the adaptive lasso method in RDD model. According to the (MSE) criterion . Although both methods have the advantage that they apply less shrinkage to non-zero operands, this indicates less bias , we note that the number of important variables appeared far from zero, as is true coefficients that assumed by the simulation and the figure that shows the features in the method MCP.

7. Conclusions:

Model (RDD) is used in many economic, social, medical, and other applications, and when this model is combined with one of selecting variable method, its performance and results are acceptable and satisfactory. One of these methods is the MCP, which gave good results through (MSE) in the simulation study. The results also indicate that the MCP method applies less shrinkage to non-zero coefficients, which indicates bias reduction We conclude that the MCP method, is a fast, continuous and almost unbiased method. MCP provides for maximally scattered scattered loss convexity given certain thresholds for variable selection and unbiasedness Through the conclusions of this papre, statisticians are assisted by the presence of the technique of organization methods in statistics, using this new technique to ensure accurate and useful results for correct prediction .

Also, the best bandwidth method used is the CCT method, followed by the CV bandwidth method Bandwidth and then IK by comparing with the value of the mean square error . We recommend employing some other variable selection methods such as group Lasso, SCAD, PACS, and others with the (RDD) model.

References:

Anastasopoulos, J. (2019). Principled estimation of regression discontinuity designs with covariates: a machine learning approach. arXiv preprint arXiv:1910.06381.

Ali, O. A., Naji, M. Q., & Ismaeel, M. M. (2020). Kernel estimation of returns of retirement funds of employers based on monetary earnings (subscriptions and compensation) via regression discontinuity in Iraq. Periodicals of Engineering and Natural Sciences *(PEN)*, *8*(3), 1752-1766.

Bloniarz, A., Liu, H., Zhang, C. H., Sekhon, J. S., & Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. Proceedings of the National Academy of Sciences, 113(27), 7383-7390

Breheny, P. (2016). Adaptive lasso, MCP, and SCAD

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82(6), 2295-2326.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2018). A Practical Introduction to Regression Discontinuity Designs: Volume II. Cambridge Elements: Quantitative and Computational Methods for Social Science, II, 113.

Choon, C. L. (2012). Minimax concave bridge penalty function for variable selection.

Frölich, M., & Huber, M. (2019). Including covariates in the regression discontinuity design. Journal of Business & Economic Statistics, 37(4), 736-748.

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

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. The Review of economic studies, 79(3), 933-959.

Jin, B., Lorenz, D. A., & Schiffler, S. (2009). Elastic-net regularization: error estimates and active set methods. Inverse Problems, 25(11), 115022.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of economic literature, 48(2), 281-355.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in e conomics. Journal of economic literature, 48(2), 281-355.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. The Quarterly journal of economics, 122(1), 159-208.

Mutair, Hafez Muhammad, 2011 "Comparison of Some Non-Parametric Regression Methods Using Simulations", Master's Message, College of Computer Science and Mathematics, University of Qadisiyah.

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

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational psychology, 51(6), 309.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320 .