Methods article template for submitting to *MethodsX*

Method Article – Title Page

Title	MIDD: An R Package for Multiple Imputation Discontinuity Designs
Authors	Masayoshi Takahashi*
Affiliations	School of Information and Data Sciences, Nagasaki University
Corresponding Author's email address	m-takahashi@nagasaki-u.ac.jp
Keywords	• Multiple imputation • Regression discontinuity • Local average treatment effect
Direct Submission or Co-Submission *Co-submissions are papers that have been submitted alongside an original research paper accepted for publication by another Elsevier journal*	Co-Submission • If Co-Submission, please provide a reference of your research article here This can be in any of the following forms: ○ DOI e.g. 10.1016/j.mex.1900.08.014 ○ PII e.g. S2215-0161(00)30217-1 ○ Elsevier production reference e.g. [MEX_00001]

ABSTRACT

Causal inference is often regarded as a missing data problem using the potential outcomes framework. The regression discontinuity design (RDD) is said to be one of the most credible causal inference techniques in non-experimental settings. On the other hand, it is reported that missing data in many fields are dealt with by multiple imputation. Nevertheless, multiple imputation has not been recognized as a method to estimate the local average treatment effect (LATE) around the cutoff point. The author proposed multiple imputation discontinuity designs (MIDDs) that can estimate the LATE at the cutoff as equally well as RDDs. This article will illustrate how to use R-Package MIDD.

- R-Package MIDD estimates the LATE based on RDDs and MIDDs.
- R-Package MIDD graphically diagnoses RDDs by comparing the results from MIDDs.
- R-Package MIDD is freely available and easy to use.

SPECIFICATIONS TABLE

Subject Area 1	Economics and Finance
Subject Area 2	Social Sciences
Subject Area 3	Statistics
More specific subject area	Causal Inference
Method name	Multiple Imputation Discontinuity Designs
Name and reference of original method	M. Takahashi: Multiple Imputation Discontinuity Designs: Alternative to Regression Discontinuity Designs to Estimate the Local Average Treatment Effect at the Cutoff, Journal of Statistical Planning and Inference.
Resource availability	R-package described in this article is freely available at https://github.com/m-takahashi123/MIDD.

Introduction

In the Rubin Causal Model (Rubin 1974), causal inference is regarded as a missing data problem using the potential outcomes framework. The regression discontinuity design (RDD) is one of the most credible causal inference techniques in non-experimental settings (Lee and Lemieux 2015), using the potential outcomes framework. An application of RDDs to the data on the incumbency advantage in U.S. House elections (Lee 2008) has been cited as an exemplar by many scholars (Angrist and Pischke 2009; Imbens and Kalyanaraman 2012; Calonico et al. 2014; Branson et al. 2019).

On the other hand, Rubin (1987) proposed multiple imputation to deal with missing data by independently and randomly drawing simulated values for missing data based on the predictive posterior distribution. Indeed, missing data in many fields are dealt with by multiple imputation (van Buuren 2018). Nevertheless, multiple imputation has not been recognized as a method to estimate the local average treatment effect (LATE) around the cutoff point.

A novel application of multiple imputation was presented in Takahashi (2021a), where multiple imputation is used to estimate the LATE around the cutoff point. This new method is named multiple imputation discontinuity designs (MIDDs). Evidence from Monte Carlo simulations in Takahashi (2021a) shows that MIDDs can indeed estimate the LATE at the cutoff as equally well as RDDs. Also, Takahashi (2021b) provides an easy-to-use software program, R-Package MIDD. This article will illustrate how to use R-Package MIDD. This is useful in diagnosing the validity of RDD analyses.

Method details

R-Package MIDD estimates the LATE based on RDDs and MIDDs, and graphically diagnoses RDDs by comparing the results from MIDDs (Takahashi 2021b). To use this package, click “Code and Download ZIP” at https://github.com/m-takahashi123/MIDD. After downloading the package, set the working directory in R, and read R-Package MIDD using R-function source as follows.

```
> setwd("C:/Folder")
> source("MIDD.R")
```

R-Package MIDD provides a real dataset "lee2008.csv" on the incumbency advantage in U.S. House elections as an example in the csv format. This dataset was originally used in Lee (2008) and is also available in Olivares and Sarmiento-Barbieri (2020). This dataset contains 6,558 observations on the following three variables. y1 is the variable of interest (the dependent variable), which is Democrat vote share in election at t+1. x1 is the running variable, which is the difference in vote share between the Democratic and Republican parties in election at t. The cutoff point is at x1=0. x2 is an additional covariate, which is Democrat vote share in election at t-1. In this example, we investigate whether there is an advantage for a candidate if the candidate’s party won the seat in the previous U.S. House election (Angrist and Pischke 2009, p.257). After downloading this dataset, use R-function read.csv as follows and attach it in R.

```
> data1<-read.csv("lee2008.csv", header=TRUE)
> attach(data1)
```

R-function MIdiagRDD computes the LATE by RDDs and MIDDs. Also, it graphically diagnoses the validity of RDDs based on MIDDs. At a minimum, the user needs to specify y, x, and cut, where y is the variable of interest (the dependent variable), x is the running variable (forcing variable) that determines the cutoff point, and cut specifies the RDD cutoff point in x. In the example of lee2008, the variable of interest is y1, the forcing variable is x1, and the cutoff point is 0.

```
> MIdiagRDD(y=y1, x=x1, cut=0)
```

Full Arguments in R-function MIdiagRDD

The full arguments using R-function MIdiagRDD are as follows. The first three options (y, x, and cut) need to be explicitly set by the user. In our case, y=y1, x=x1, and cut=0. The other options have default settings as follows.

```
> MIdiagRDD(y, x, cut, seed=1, M1=100, M2=5, M3=1, p2s1=1, emp=0, bw="mserd",
> ker="triangular", bwidth=1, p1=1, conf=95, upper=1, covs1=NULL)
```

The first part of the options is related to multiple imputation. Option seed sets the seed value for random numbers, where the default is 1. Since MIDD is based on the simulated technique of multiple imputation, setting the seed is important to exactly reproduce the same result. Option M1 is the number of imputed datasets to create, where the default is 100. Option M2 is the number of densities based on imputed datasets to display in graphs 3 and 4, and the number of estimated slopes in graphs 9 and 10. The default for M2 is 5. These datasets are the subsets of M1 imputed datasets; thus, M2 cannot be larger than M1. Option M3 is the number of imputed datasets to display in graphs 5 to 10, where the default is 1. These datasets are also the subsets of M1 imputed datasets; thus, M3 cannot be larger than M1. Option p2s1 is an integer value taking either 0 or 1, where 0 for no screen output and 1 for screen printing of multiple imputation process. The default is 1. This allows the user to monitor how long the chains of the EM algorithm would be. Long chains signal that the imputation model may need to be modified, such as transformations of variables (Honaker et al. 2011, p.9). Option emp is the number indicating level of the empirical (ridge) prior. The default is 0. When the subsample size is too small, this option may stabilize the imputation model. A reasonable upper bound is 0.1. Essentially, the ridge prior adds artificial observations to the dataset with the same means and variances as the original data with zero covariances (Honaker et al. 2011, p.20).

The second part of options is related to regression discontinuity designs. Option bw specifies the bandwidth selection procedure for the RDD based on Calonico et al. (2020). Choice is mserd, msesum, cerrd, and cersum, where mserd is one common MSE-optimal bandwidth selector, msesum is one common MSE-optimal bandwidth selector for the sum of regression estimates, cerrd is one common CER-optimal bandwidth selector, and cersum is one common CER-optimal bandwidth selector for the sum of regression estimates. MSE is Mean Squared Error, and CER is Coverage Error Rate. The default is mserd. Option ker is the kernel function used to construct the local-polynomial estimator for the RDD based on Calonico et al. (2020). Options are triangular, epanechnikov, and uniform. The default is triangular. Option bwidth is a number to adjust the size of the chosen bandwidth. The default is 1. If the user wants a narrower bandwidth, set this value less than 1. If the user wants a wider bandwidth, set this value larger than 1. Option p1 specifies the order of the local-polynomial used to construct the point-estimator for the RDD and the MIDD. p1 can take either 1 (local linear regression) or 2 (local quadratic regression), where the default is p1=1. When specified larger than 2, it will be considered 2.

Option conf is the confidence level for the confidence interval. The default is set to 95. Option upper specifies which part of the running variable is the treatment group. If the upper part is the treatment group, upper=1. If the lower part is the treatment group, upper=0. The default is set to 1. Option covs1 specifies additional covariates to be used for estimation and inference in RDDs and MIDDs. Adding covariates is a bit tricky, by specifying covs1=data.frame(V1, V2, V3), where V is a covariate.

Example 1

As noted above, the following codes will read in the example data of lee2008. Remember that, in the example of lee2008, the variable of interest is y1, the forcing variable is x1, and the cutoff point is 0. The following codes will replicate the results reported in Section 7.2 of Takahashi (2021a).

```
> setwd("C:/Folder")
> source("MIDD.R")
> data1<-read.csv("lee2008.csv", header = TRUE)
> attach(data1)
> MIdiagRDD(y=y1, x=x1, cut=0, bwidth=2.2)
```

Example 2

Using the same dataset, we may change the settings as follows, where the bandwidth is half-size and we have an additional covariate x2 in the model. The outputs are reported in the next section.

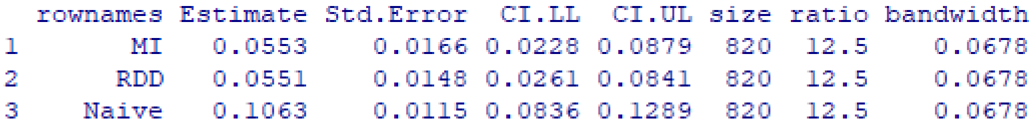
```
> MIdiagRDD(y=y1, x=x1, cut=0, bwidth=0.5, covs1=data.frame(x2))
```

Note that the cutoff in this example is 0, but this value depends on data in hand. Although this dataset has only three variables in total, adding more covariates is also straightforward, by covs1=data.frame(x2, x3, x4) if additional covariates are named x3 and x4.

Outputs in R-function MIdiagRDD

R-function MIdiagRDD produces the following values in the output: Estimate, Std.Error, CI.LL, CI.UL, size, ratio, and bandwidth. Estimate is the point estimate of the LATE at the cutoff. Std.Error is the standard error of the estimate. CI.LL is the lower limit of the confidence interval. CI.UL is the upper limit of the confidence interval. Size is the subsample size to estimate the LATE at the cutoff. Ratio is the ratio of the subsample to the original sample size. Bandwidth is the length of the bandwidth used for RDD analysis.

The output is reported in Figure 1, where Naive is the naïve estimator that simply takes a difference between the mean of observed Y(0) and the mean of observed Y(1), where Y(0) and Y(1) are the pair of potential outcomes. MI is the proposed method based on multiple imputation discontinuity designs explained in Takahashi (2021a). RDD is the regression discontinuity design.



```
  rownames Estimate Std.Error  CI.LL  CI.UL size ratio bandwidth
1       MI   0.0553    0.0166 0.0228 0.0879  820  12.5    0.0678
2      RDD   0.0551    0.0148 0.0261 0.0841  820  12.5    0.0678
3    Naive   0.1063    0.0115 0.0836 0.1289  820  12.5    0.0678
```

Figure 1. Numerical Output of the Analyses

Along with these numerical outputs, R-function MIdiagRDD produces twelve diagnostic plots. Figure 2 presents diagnostic Plots 1 and 2. Plot 1 (MI, RDD, Naive) is a diagnostic plot to visualize the relationship among the three estimators. Red vertical line is RDD, black solid line is naïve, and histogram is MI. Plot 2 (MI and RDD) is a diagnostic plot to visualize the relationship between the two estimators. Red vertical line is RDD and histogram is MI.

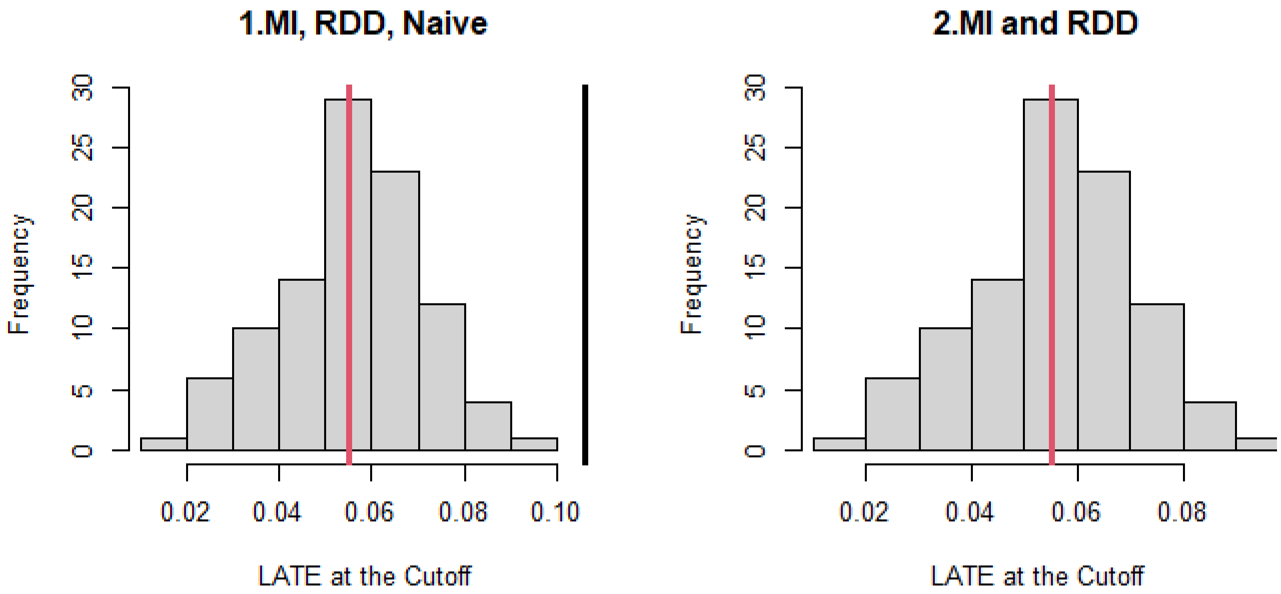


Figure 2. Diagnostic Plots 1 and 2.

Figure 3 presents diagnostic Plots 3 and 4. Plot 3 (Densities: Control) is a diagnostic plot to visualize the densities of observed and imputed data. Gray solid curve is the density of observed data in the control group. Blue solid curve is the density of observed data in the treatment group. Red dashed lines are the densities of imputed data in the control group. Plot 4 (Densities: Treatment) is a diagnostic plot to visualize the densities of observed and imputed data. Gray solid curve is the density of observed data in the control group. Blue solid curve is the density of observed data in the treatment group. Red dashed lines are the densities of imputed data in the treatment group.

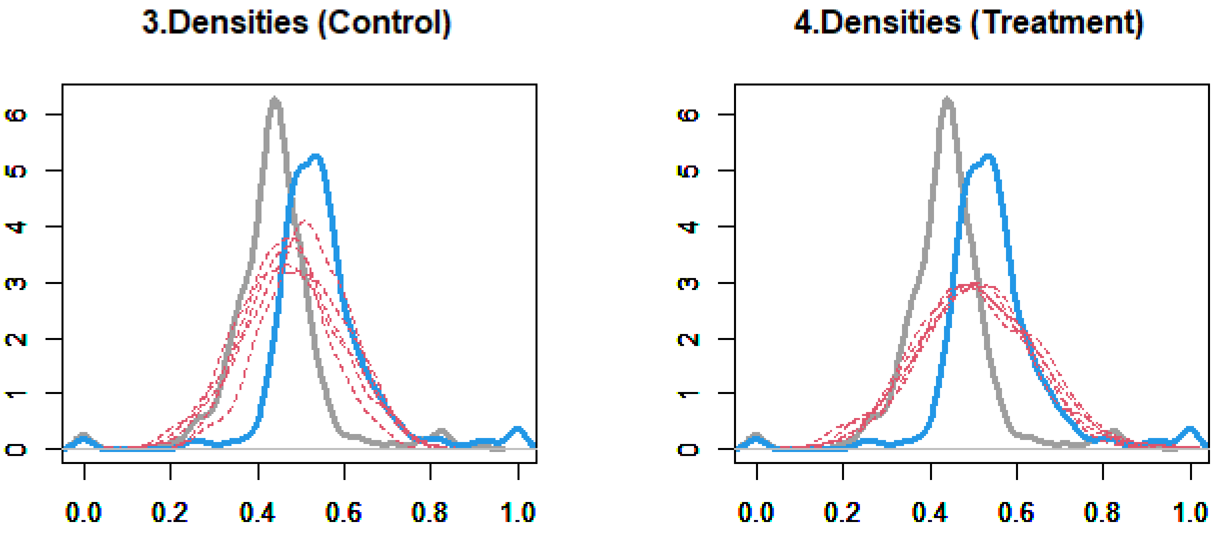


Figure 3. Diagnostic Plots 3 and 4.

Figure 4 presents diagnostic Plots 5 and 6. Plot 5 (Observed Values) is a diagnostic plot to visualize the scatterplot of observed data. Gray circles are observed data in the control group. Blue triangles are observed data in the treatment group. Plot 6 (Observed & Imputed Values) is a diagnostic plot to visualize the scatterplot of observed and imputed data. Red circles are imputed data in the control group. Red triangles are imputed data in the treatment group. These imputed data are overlaid on the observed data in Plot 5.

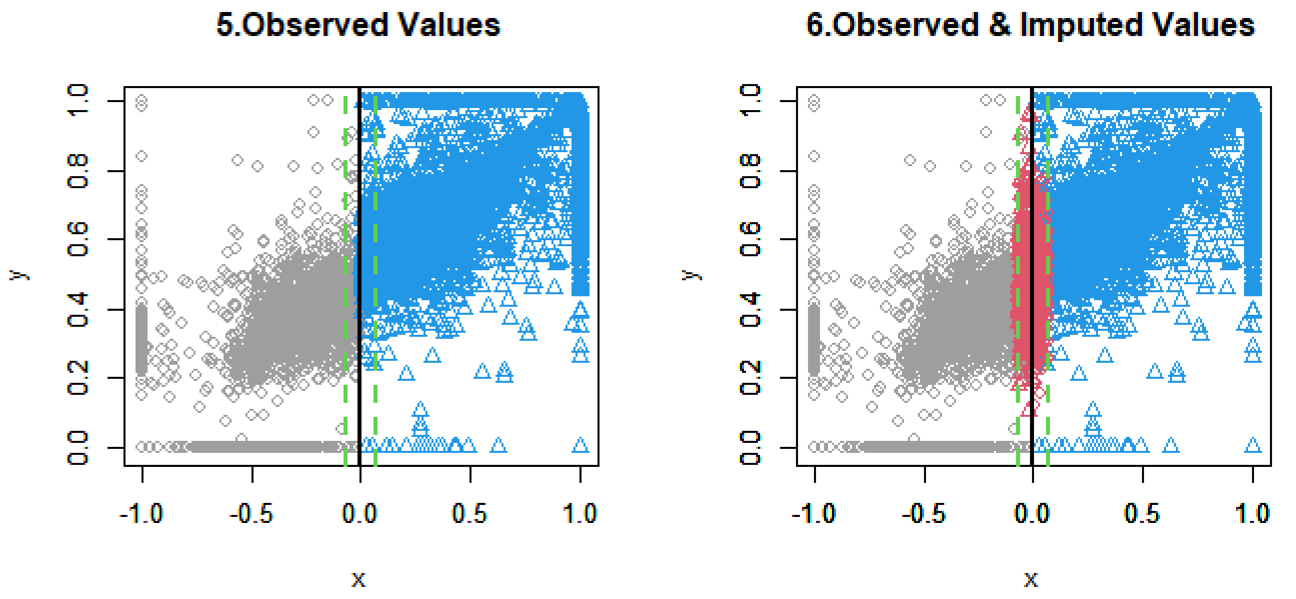


Figure 4. Diagnostic Plots 5 and 6.

Figure 5 presents diagnostic Plots 7 and 8. Plot 7 (Observed & Imputed: Control) is a diagnostic plot to clearly visualize the scatterplot of observed and imputed data in the control group only. Plot 8 (Observed & Imputed: Treatment) is a diagnostic plot to clearly visualize the scatterplot of observed and imputed data in the treatment group only.

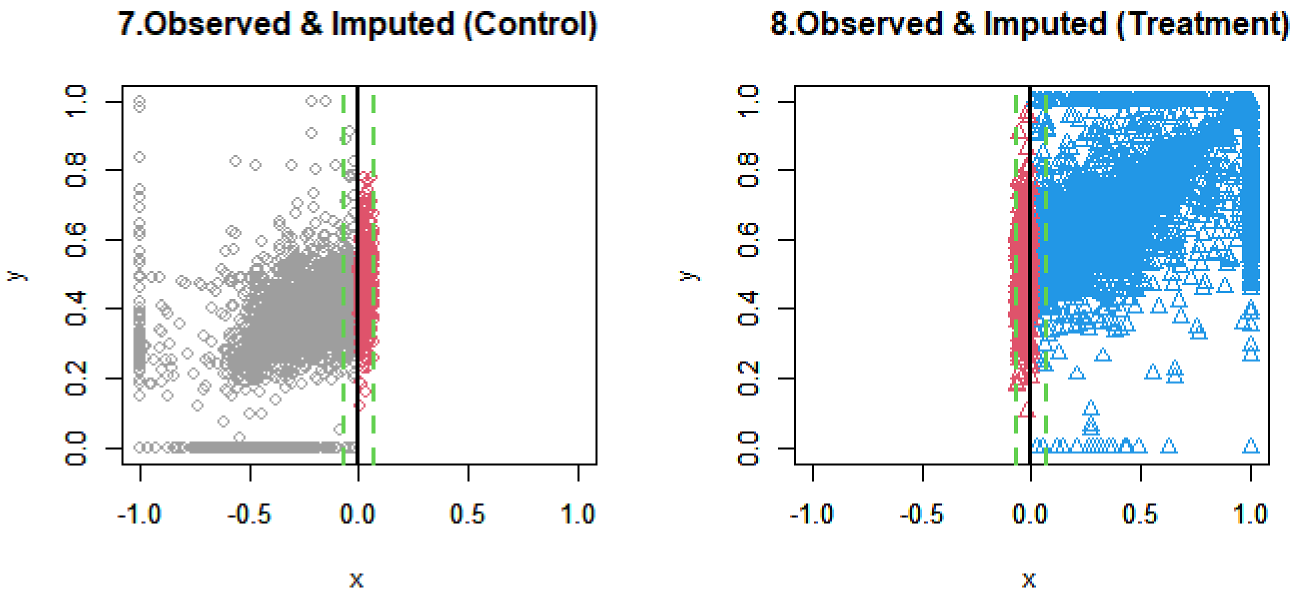


Figure 5. Diagnostic Plots 7 and 8.

Figure 6 presents diagnostic Plots 9 and 10. Plot 9 (Around Cutoff: Control) is a diagnostic plot of observed and imputed data in the control group only, to clearly visualize the scatterplot, around the cutoff point. Plot 10 (Around Cutoff: Treatment) is a diagnostic plot of observed and imputed data in the treatment group only, to clearly visualize the scatterplot, around the cutoff point. Five solid lines are the estimated linear regression lines based on multiply imputed data.

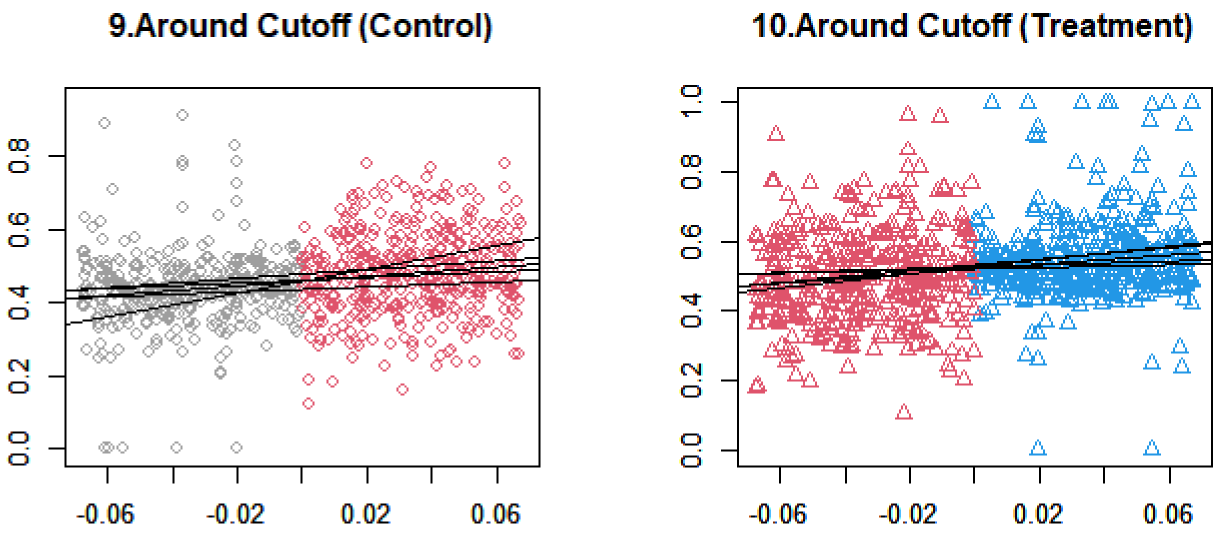


Figure 6. Diagnostic Plots 9 and 10.

Figure 7 presents diagnostic Plots 11 and 12. Plot 11 (Local Slope: Control) is a diagnostic plot to visualize the distribution of the coefficients of the estimated linear regression models around the cutoff point in the control group. Plot 12 (Local Slope: Treatment) is a diagnostic plot to visualize the distribution of the coefficients of the estimated linear regression models around the cutoff point in the treatment group.

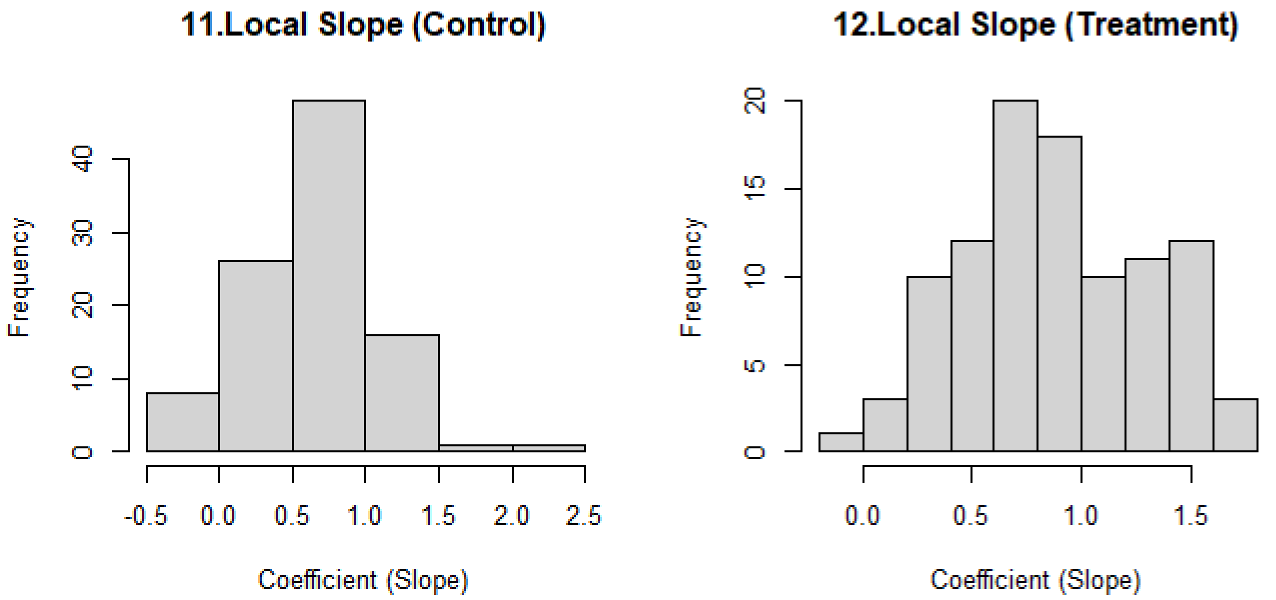


Figure 7. Diagnostic Plots 11 and 12.

Acknowledgements

The author wishes to thank the editors and the anonymous reviewers.

Declaration of interests

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Angrist, J. D., and Pischke, J. S. (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton: Princeton University Press.

[2] Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019), "A Nonparametric Bayesian Methodology for Regression Discontinuity Designs," Journal of Statistical Planning and Inference, 202, 14-30.

[3] Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," Econometrica, 82 (6), 2295-2326.

[4] Calonico, S., Cattaneo, M. D., and Titiunik, R. (2020), "Package 'rdrobust'," The Comprehensive R Archive Network. Available at https://cran.r-project.org/web/packages/rdrobust/rdrobust.pdf.

[5] Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), pp.1-47.

[6] Imbens, G., and Kalyanaraman, K. (2012), “Optimal Bandwidth Choice for the Regression Discontinuity Estimator,” The Review of Economic Studies, 79 (3), 933-959.

[7] Lee, D. S. (2008). Randomized Experiments from Non-Random Selection in U.S. House Elections. Journal of Econometrics, 142, pp.675-697.

[8] Lee, D. S., and Lemieux, T. (2015), “Regression Discontinuity Designs in Social Sciences,” in The Sage Handbook of Regression Analysis and Causal Inference, eds. H. Best and C. Wolf, Thousand Oaks: Sage Publications.

[9] Olivares, M., and Sarmiento-Barbieri, I. (2020). “Package ‘RATest’,” The Comprehensive R Archive Network. Available at https://cran.r-project.org/web/packages/RATest/RATest.pdf.

[10] Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” Journal of Educational Psychology, 66 (5), 688-701.

[11] Rubin, D. B. (1987), Multiple Imputation for Nonresponse in Surveys, New York, NY: John Wiley & Sons.

[12] Takahashi, M. (2021a). Multiple Imputation Discontinuity Designs: Alternative to Regression Discontinuity Designs to Estimate the Local Average Treatment Effect at the Cutoff. Currently under review.

[13] Takahashi, M. (2021b). “R-Package ‘MIDD’,” GitHub repository. Available at https://github.com/m-takahashi123/MIDD.

[14] van Buuren, S. (2018), Flexible Imputation of Missing Data (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.