

# Notes for Matlab and Stata Regression Discontinuity Software

Daisuke Fuji,<sup>\*</sup> Guido Imbens,<sup>†</sup>  
and Karthik Kalyanaraman<sup>‡</sup>

August 4, 2009

## 1 Introduction

These notes describe the implementation of the Imbens-Kalyanaram bandwidth selection in Regression Discontinuity settings using Matlab and Stata. There are two data sets to illustrate the programs. The first is an artificial data set for the sharp regression discontinuity design, posted as `art_sharp_rd.txt`. This file contains data on 1,000 units, with for each unit the values for three variables an outcome  $Y_i$ , the forcing variable  $X_i$ , and four additional covariates  $Z_i$ . The value of the threshold is  $c = 0.2990$ .

The second is an artificial data set for the fuzzy regression discontinuity design, posted as `art_fuzzy_rd.txt`. This file contains data on 2,000 units, with for each unit the values for four variables an outcome  $Y_i$ , a treatment indicator  $W_i$ , the forcing variable  $X_i$ , and three additional covariates  $Z_i$ . The value of the threshold is  $c = 0.5$ .

## 2 Matlab

There are two Matlab programs, `rd_matlab_09aug4.m`, and `rd_optbandwidth.m`. The program `rd_matlab_09aug4.m` reads in the data and calls the routine `rd_optbandwidth` with five inputs: the vector of outputs `y`, the vector of treatment indicators `w`, the vector of forcing variables `x`, a matrix of the threshold `c`, a matrix of additional covariates `z`, a binary indicator `output`, and a binary indicator for the presence of additional covariates `add_cov`. If the indicator `output` is equal to 1, intermediate output will be reported, otherwise only the point estimate of the rd estimand, its standard error, and the optimal bandwidth will be reported. If the indicator

---

<sup>\*</sup>Electronic correspondence: [fujii@fas.harvard.edu](mailto:fujii@fas.harvard.edu).

<sup>†</sup>Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: [imbens@harvard.edu](mailto:imbens@harvard.edu), <http://www.economics.harvard.edu/faculty/imbens/imbens.html>.

<sup>‡</sup>Harvard University, Cambridge, MA 02138. Electronic correspondence: [kalyanar@harvard.edu](mailto:kalyanar@harvard.edu).

`add_cov` is equal to 1, the additional covariates will be used, otherwise they will be ignored. If there are no additional covariates, the value of `z` does not matter, but it always has to be entered as an input.

The output from `rd_matlab_09aug4.m` is in the file `output_09aug4.txt`.

### 3 Stata

The program `rd_stata_09aug4.do` reads in the data and calls the routine `rdob.ado`. This general form of the call is

```
y x z1 z2 z3, c(0.5) fuzzy(w) detail
```

`y` is the output. `x` is the forcing variable. These two inputs are followed by an optional set of covariates, denoted here by `z1`, `z2`, and `z3`. This is followed by a comma, followed by `c(threshold)`, where `threshold` is the value of the threshold for the regression discontinuity design. `fuzzy(w)` is an option if the design is fuzzy. In that case `w` is the treatment indicator. `detail` is an option to report intermediate output.

The output from `rd_stata_09aug4.m` is in the file scml file `rd_log_09aug4.log`.

### 4 Optimal Bandwidth

IK propose estimating the optimal bandwidth as

$$\hat{h}_{\text{opt}} = C_K \cdot \left( \frac{2 \cdot \hat{\sigma}^2(c) / \hat{f}(c)}{\left( \hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-)} \right)^{1/5} \cdot N^{-1/5}. \quad (4.1)$$

In this formula  $\hat{f}(c)$  is an estimator for the density of the forcing variable at the threshold  $c$ ,  $\hat{\sigma}^2(c)$  is an estimator for the conditional variance of  $Y$  given the forcing, evaluated at the threshold,  $\hat{m}_+^{(2)}(c)$  and  $\hat{m}_-^{(2)}(c)$  are estimators for the second derivatives of the regression function from the left and the right, as a function of the forcing variable, evaluated at the threshold. The remaining components,  $\hat{r}_+$  and  $\hat{r}_-$  are regularization parameters to avoid instabilities associated with low values of the difference in the second derivatives from the left and the right. For the edge kernel we use in the calculations, with  $K(u) = \mathbf{1}_{|u| \leq 1}(1 - |u|)$  (e.g., Cheng, Fan and Marron, 1997), the multiplicative constant is  $C_K \approx 3.4375$ .

### 5 IK algorithm on artificial data

In this section we implement our proposed bandwidth on the `srd` artificial dataset. For clarity we gave all the intermediate steps. There are 1000 observations, 617 with  $X_i < c$ , and 383 with

$X_i \geq c$ , where  $c = 0.2990$ .

Step 1: Estimation of density  $f(c)$  and conditional variance  $\sigma^2(c)$

We start with the modified Silverman bandwidth,

$$h_1 = 1.84 \cdot S_X \cdot N^{-1/5} = 0.4765,$$

leading to

$$\hat{f}(c) = 0.3715,$$

and

$$\hat{\sigma}^2(c) = 1.9761^2.$$

Step 2: Estimation of second derivatives  $\hat{m}_+^{(2)}(c)$  and  $\hat{m}_-^{(2)}(c)$

To estimate the curvature at the threshold, we first need to choose bandwidths  $h_{2,+}$  and  $h_{2,-}$ . We choose these bandwidths based on an estimate of  $\hat{m}^{(3)}(c)$ , obtained by fitting a global cubic with a jump at the threshold. We estimate this global cubic regression function by dropping observations with covariate values below the median of the covariate for observations with covariate values below the threshold, and dropping observations with covariate values above the median of the covariate for observations with covariate values above the threshold. For the 617 (383) observations with  $X_i < c$  ( $X_i > c$ ), the median of the forcing variable is -0.5418 (0.8224). Next, we estimate, using the data with  $X_i \in [-0.5418, 0.8224]$ , the polynomial regression function of order three, with a jump at the threshold of  $c = 0.2990$ :

$$Y_i = \gamma_0 + \gamma_1 \cdot X_i + \gamma_2 \cdot X_i^2 + \gamma_3 \cdot X_i^3 + \gamma_4 \cdot 1_{X_i \geq 0.3} + \varepsilon_i.$$

The least squares estimate for  $\gamma_3$  is  $\hat{\gamma}_3 = -2.0717$ , and thus the third derivative at  $c$  is estimated as  $\hat{m}^{(3)}(0.2990) = 6 \cdot \hat{\gamma}_3 = -12.4302$ . This leads to the two bandwidths

$$h_{2,-} = 0.9685, \quad \text{and} \quad h_{2,+} = 1.0367.$$

The two pilot bandwidths are used to fit two quadratics. The quadratic to the right (left) of  $c = 0.2990$  is fitted on  $[0.2990, 1.3357]$  ( $[-0.6695, 0.2990]$ ), yielding

$$\hat{m}_-^{(2)}(c) = 1.8604, \quad \text{and} \quad \hat{m}_+^{(2)}(c) = -5.7224.$$

Step 3: Calculation of Regularization Terms  $\hat{r}_-$  and  $\hat{r}_+$ , and Calculation of  $\hat{h}_{\text{opt}}$

Next, the regularization terms are calculated. We obtain

$$\hat{r}_- = \frac{720 \cdot \hat{\sigma}^2(c)}{N_{2,+} \cdot h_{2,+}^4} = 9.0033 \quad \text{and} \quad \hat{r}_+ = \frac{720 \cdot \hat{\sigma}^2(c)}{N_{2,-} \cdot h_{2,-}^4} = 8.2225.$$

Now we have all the ingredients to calculate the optimal bandwidth under different kernels and the corresponding RD estimates. Using the edge kernel, with  $C_K = 3.4375$ , we obtain

$$\hat{h}_{\text{opt}} = C_K \cdot \left( \frac{2 \cdot \hat{\sigma}^2(c) / \hat{f}(c)}{\left( \hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-)} \right)^{1/5} \cdot N^{-1/5} = 0.6700.$$

## 6 Estimation and Inference Given the Bandwidth

In the previous section we described an algorithm for choosing the bandwidth  $h$ . This value of the optimal bandwidth depends on the choice of kernel  $K(\cdot)$ . In this section we discuss estimation and inference given the kernel and the bandwidth. We use local linear estimation, with a triangular kernel. We allow for additional covariates, denoted by  $Z_i$ .

### 6.1 Estimation and Inference for the Sharp Regression Discontinuity Design

Define the weight for observation  $i$ ,

$$\lambda_i = K\left(\frac{X_i - c}{h}\right) = \left(1 - \frac{|X_i - c|}{h}\right) \cdot \mathbf{1}_{|X_i - c| \leq h},$$

and let  $\bar{Z}_\lambda$  be the weighted average of the covariates:

$$\bar{Z}_\lambda = \frac{\sum_{i=1}^N \lambda_i \cdot Z_i}{\sum_{i=1}^N \lambda_i}.$$

Then

$$\left(\hat{\alpha}_l, \hat{\beta}_l, \hat{\gamma}_l\right) = \arg \min_{\alpha, \beta, \gamma} \sum_{i: X_i < c} \lambda_i \cdot \left(Y_i - \alpha - \beta \cdot (X_i - c) - \gamma'(Z_i - \bar{Z}_\lambda)\right)^2,$$

and

$$\left(\hat{\alpha}_r, \hat{\beta}_r, \hat{\gamma}_r\right) = \arg \min_{\alpha, \beta, \gamma} \sum_{i: X_i \geq c} \lambda_i \cdot \left(Y_i - \alpha - \beta \cdot (X_i - c) - \gamma'(Z_i - \bar{Z}_\lambda)\right)^2.$$

For the artificial data in `art_sharp_rd.txt`, we find: that the mean and standard deviation

of  $\lambda_i$  are 0.3527 (0.3409), and  $\bar{Z}_\lambda = 0.0364$ .

$$\begin{pmatrix} \hat{\alpha}_l \\ \hat{\beta}_l \\ \hat{\gamma}_{l1} \\ \hat{\gamma}_{l2} \\ \hat{\gamma}_{l3} \\ \hat{\gamma}_{l4} \end{pmatrix} = \begin{pmatrix} 0.7368 & (0.0382) \\ 1.1476 & (0.1154) \\ 0.9802 & (0.0208) \\ 1.0024 & (0.0196) \\ 1.0225 & (0.0215) \\ 0.9342 & (0.0235) \end{pmatrix}, \quad \begin{pmatrix} \hat{\alpha}_r \\ \hat{\beta}_r \\ \hat{\gamma}_{r1} \\ \hat{\gamma}_{r2} \\ \hat{\gamma}_{r3} \\ \hat{\gamma}_{r4} \end{pmatrix} = \begin{pmatrix} 1.7364 & (0.0366) \\ 0.8156 & (0.1177) \\ 0.9847 & (0.0200) \\ 0.9969 & (0.0223) \\ 1.0232 & (0.0183) \\ 1.0234 & (0.0210) \end{pmatrix}.$$

Then

$$\hat{\tau}_{\text{RD}} = \hat{\alpha}_r - \hat{\alpha}_l.$$

The point estimate for the RD estimand is then

$$\hat{\tau}_{\text{RD}} = 1.0094 \quad (\text{s.e. } 0.0400).$$

Let  $\hat{e}_i$  be the residual,

$$\hat{e}_i = \begin{cases} Y_i - \hat{\alpha}_l - \hat{\beta}_l \cdot (X_i - c) - \hat{\gamma}'_l (Z_i - \bar{Z}_\lambda) & \text{if } X_i < c, \\ Y_i - \hat{\alpha}_r - \hat{\beta}_r \cdot (X_i - c) - \hat{\gamma}'_r (Z_i - \bar{Z}_\lambda) & \text{if } X_i \geq c, \end{cases}$$

and

$$A_i = \begin{pmatrix} 1 & (X_i - c) & (Z_i - \bar{Z}_\lambda)' \\ (X_i - c) & (X_i - c)^2 & (X_i - c) \cdot (Z_i - \bar{Z}_\lambda)' \\ (Z_i - \bar{Z}_\lambda) & (X_i - c) \cdot (Z_i - \bar{Z}_\lambda) & (Z_i - \bar{Z}_\lambda) \cdot (Z_i - \bar{Z}_\lambda)' \end{pmatrix}.$$

The variance for  $(\hat{\alpha}_l, \hat{\beta}_l, \hat{\gamma}'_l)'$  and  $(\hat{\alpha}_r, \hat{\beta}_r, \hat{\gamma}'_r)'$  are estimated as

$$\hat{\mathbb{V}}_l = \hat{\Gamma}_l^{-1} \hat{\Delta}_l \hat{\Gamma}_l^{-1}, \quad \text{and} \quad \hat{\mathbb{V}}_r = \hat{\Gamma}_r^{-1} \hat{\Delta}_r \hat{\Gamma}_r^{-1},$$

where,

$$\begin{aligned} \hat{\Delta}_l &= \sum_{i: X_i < c} \lambda_i^2 \cdot \hat{e}_i^2 \cdot A_i, & \hat{\Delta}_r &= \sum_{i: X_i > c} \lambda_i^2 \cdot \hat{e}_i^2 \cdot A_i \\ \hat{\Gamma}_l &= \sum_{i: X_i < c} \lambda_i \cdot A_i, & \text{and } \hat{\Gamma}_r &= \sum_{i: X_i > c} \lambda_i \cdot A_i \end{aligned}$$

This leads to standard errors for  $(\hat{\alpha}_l, \hat{\beta}_l, \hat{\gamma}'_l)'$  and  $(\hat{\alpha}_r, \hat{\beta}_r, \hat{\gamma}'_r)'$  given above.

The variance of  $\hat{\tau}_{\text{RD}}$  is

$$\mathbb{V}(\hat{\tau}_{\text{RD}}) = \mathbb{V}(\hat{\alpha}_l) + \mathbb{V}(\hat{\alpha}_r),$$

estimated as

$$\hat{\mathbb{V}}(\hat{\tau}_{\text{RD}}) = \hat{\mathbb{V}}_{l,(1,1)} + \hat{\mathbb{V}}_{r,(1,1)} = 0.0400^2.$$

## 6.2 Estimation and Inference for the Fuzzy Regression Discontinuity Design

In the fuzzy regression discontinuity design we use the same bandwidth based on the outcome and forcing variable as before, ignoring the data on the value of the treatment. Given the same bandwidth, we calculate the same weights  $\lambda_i$ . Then we do four local linear regressions:

$$\begin{aligned} (\hat{\alpha}_{Y,l}, \hat{\beta}_{Y,l}, \hat{\gamma}_{Y,l}) &= \arg \min_{\alpha, \beta, \gamma} \sum_{i: X_i < c} \lambda_i \cdot (Y_i - \alpha - \beta \cdot (X_i - c) - \gamma'(Z_i - \bar{Z}_\lambda))^2, \\ (\hat{\alpha}_{Y,r}, \hat{\beta}_{Y,r}, \hat{\gamma}_{Y,r}) &= \arg \min_{\alpha, \beta, \gamma} \sum_{i: X_i \geq c} \lambda_i \cdot (Y_i - \alpha - \beta \cdot (X_i - c) - \gamma'(Z_i - \bar{Z}_\lambda))^2, \\ (\hat{\alpha}_{W,l}, \hat{\beta}_{W,l}, \hat{\gamma}_{W,l}) &= \arg \min_{\alpha, \beta, \gamma} \sum_{i: X_i < c} \lambda_i \cdot (W_i - \alpha - \beta \cdot (X_i - c) - \gamma'(Z_i - \bar{Z}_\lambda))^2, \end{aligned}$$

and

$$(\hat{\alpha}_{W,r}, \hat{\beta}_{W,r}, \hat{\gamma}_{W,r}) = \arg \min_{\alpha, \beta, \gamma} \sum_{i: X_i \geq c} \lambda_i \cdot (W_i - \alpha - \beta \cdot (X_i - c) - \gamma'(Z_i - \bar{Z}_\lambda))^2.$$

We then estimate the fuzzy rd estimand as

$$\hat{\tau}_{\text{FRD}} = \frac{\hat{\alpha}_{Y,r} - \hat{\alpha}_{Y,l}}{\hat{\alpha}_{W,r} - \hat{\alpha}_{W,l}}.$$

Next we estimate the variance. The main issue is keeping track of the covariance between the estimators in the numerator and the denominator. Define the residuals

$$\hat{e}_{Y,i} = \begin{cases} Y_i - \hat{\alpha}_{Y,l} - \hat{\beta}_{Y,l} \cdot (X_i - c) - \hat{\gamma}'_{Y,l}(Z_i - \bar{Z}_\lambda) & \text{if } X_i < c, \\ Y_i - \hat{\alpha}_{Y,r} - \hat{\beta}_{Y,r} \cdot (X_i - c) - \hat{\gamma}'_{Y,r}(Z_i - \bar{Z}_\lambda) & \text{if } X_i \geq c. \end{cases}$$

and

$$\hat{e}_{W,i} = \begin{cases} W_i - \hat{\alpha}_{W,l} - \hat{\beta}_{W,l} \cdot (X_i - c) - \hat{\gamma}'_{W,l}(Z_i - \bar{Z}_\lambda) & \text{if } X_i < c, \\ W_i - \hat{\alpha}_{W,r} - \hat{\beta}_{W,r} \cdot (X_i - c) - \hat{\gamma}'_{W,r}(Z_i - \bar{Z}_\lambda) & \text{if } X_i \geq c. \end{cases}$$

Define

$$\begin{aligned} \hat{\Delta}_{Y,l} &= \sum_{i: X_i > c} \lambda_i^2 \cdot \hat{e}_{Y,i}^2 \cdot A_i, & \hat{\Delta}_{W,l} &= \sum_{i: X_i > c} \lambda_i^2 \cdot \hat{e}_{W,i}^2 \cdot A_i \\ \hat{\Delta}_{Y,r} &= \sum_{i: X_i > c} \lambda_i^2 \cdot \hat{e}_{Y,i}^2 \cdot A_i, & \hat{\Delta}_{W,r} &= \sum_{i: X_i > c} \lambda_i^2 \cdot \hat{e}_{W,i}^2 \cdot A_i \\ \hat{\Delta}_{YW,l} &= \sum_{i: X_i < c} \lambda_i^2 \cdot \hat{e}_{Y,i} \cdot \hat{e}_{W,i} \cdot A_i & \text{and } \hat{\Delta}_{YW,r} &= \sum_{i: X_i > c} \lambda_i^2 \cdot \hat{e}_{Y,i} \cdot \hat{e}_{W,i} \cdot A_i. \end{aligned}$$

Finally, define

$$\hat{\Delta} = \begin{pmatrix} \hat{\Delta}_{Y,l} & \hat{\Delta}_{YW,l} & 0 & 0 \\ \hat{\Delta}_{YW,l} & \hat{\Delta}_{W,l} & 0 & 0 \\ 0 & 0 & \hat{\Delta}_{Y,r} & \hat{\Delta}_{YW,r} \\ 0 & 0 & \hat{\Delta}_{YW,r} & \hat{\Delta}_{W,r} \end{pmatrix}$$

and, with  $\hat{\Gamma}_l$  and  $\hat{\Gamma}_r$  estimated as before in the sharp rd case,

$$\hat{\Gamma} = \begin{pmatrix} \hat{\Gamma}_l & 0 & 0 & 0 \\ 0 & \hat{\Gamma}_l & 0 & 0 \\ 0 & 0 & \hat{\Gamma}_r & 0 \\ 0 & 0 & 0 & \hat{\Gamma}_r \end{pmatrix}.$$

Then we estimate the covariance matrix of  $\hat{\theta} = (\hat{\alpha}_{Y,l}, \hat{\beta}_{Y,l}, \hat{\gamma}'_{Y,l}, \hat{\alpha}_{W,l}, \hat{\beta}_{W,l}, \hat{\gamma}'_{W,l}, \hat{\alpha}_{Y,r}, \hat{\beta}_{Y,r}, \hat{\gamma}'_{Y,r}, \hat{\alpha}_{W,r}, \hat{\beta}_{W,r}, \hat{\gamma}'_{W,r})'$  as  $\hat{V} = \hat{\Gamma}^{-1} \hat{\Delta} \hat{\Gamma}^{-1}$ . Let  $K_Z$  be the dimension of the additional covariates  $Z_i$ , so that the dimension of  $\hat{\theta}$  is  $4(K_Z + 2)$ . Then the estimator that is the focus of the analysis is

$$\hat{\tau}_{\text{FRD}} = \frac{\hat{\alpha}_{Y,r} - \hat{\alpha}_{Y,l}}{\hat{\alpha}_{W,r} - \hat{\alpha}_{W,l}} = \frac{\hat{\theta}_{2K_Z+5} - \hat{\theta}_1}{\hat{\theta}_{3K_Z+7} - \hat{\theta}_{K_Z+3}}.$$

Hence by the delta method the variance of  $\hat{\tau}_{\text{FRD}}$  is estimated as

$$\hat{V}_\tau = g' \hat{V} g,$$

where  $g$  is a column vector with dimension  $4(K_Z + 2)$ , with all elements equal to zero other than those corresponding to the four  $\hat{\alpha}$ 's:

$$\begin{aligned} g_1 &= -1/(\hat{\theta}_{3K_Z+7} - \hat{\theta}_{K_Z+3}) & g_{K_Z+3} &= (\hat{\theta}_{2K_Z+5} - \hat{\theta}_1)/(\hat{\theta}_{3K_Z+7} - \hat{\theta}_{K_Z+3})^2 \\ g_{2K_Z+5} &= 1/(\hat{\theta}_{3K_Z+7} - \hat{\theta}_{K_Z+3}) & g_{3K_Z+7} &= -(\hat{\theta}_{2K_Z+5} - \hat{\theta}_1)/(\hat{\theta}_{3K_Z+7} - \hat{\theta}_{K_Z+3})^2. \end{aligned}$$

## REFERENCES

- CHENG, M.-Y., FAN, J. AND MARRON, J.S., (1997), "On Automatic Boundary Corrections," *The Annals of Statistics*, 25, 1691-1708.
- HAHN, J., TODD, P., AND VAN DER KLAUW, W., (2001), "Regression discontinuity," *Econometrica*, 69(1), 201-209.
- IMBENS, G., AND K. KALYANARAMAN, (2009), "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," NBER Working Paper # 14726.
- LEE, D., AND T. LEMIEUX, (2009), "Regression Discontinuity Designs in Economics," Working Paper, Dept of Economics, Princeton University.
- VAN DER KLAUW, W., (2008), "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2): 219-245.