MITx: 14.310x Data Analysis for Social Scientists

Hel

Bookmarks

- Module 1: The Basics of R and Introduction to the Course
- Entrance Survey
- Module 2: Fundamentals of Probability, Random Variables, Distributions, and Joint Distributions
- Module 3: Gathering and Collecting Data, Ethics, and Kernel Density Estimates
- Module 4: Joint, Marginal, and Conditional Distributions & Functions of Random Variable

Module 10: Practical Issues in Running Regressions, and Omitted Variable Bias > Module 10: Homework > Regression Discontinuity: Questions 11 - 18

Regression Discontinuity: Questions 11 - 18

Bookmark this page

In this part of the homework we are going to replicate the results of David S. Lee paper, who has kindly provided his data to the Mostly Harmless Econometrics Data Archive

Lee (2008) studies the effect of party incumbency on reelection probabilities. In general, Lee is interested in whether a Democratic candidate for a seat in the U.S. House of Representatives has an advantage if his party won the seat last time. Here is the abstract of the working paper version of "The Electoral Advantage to Incumbency and Voters' Valuation of Politicians Experience: A Regression Discontinuity Analysis of Elections to the U.S. Houses"

> Using data on elections to the United States House of Representatives (1946-1998), this paper exploits a quasi-experiment generated by the electoral system in order to determine if political incumbency provides an electoral advantage - an implicit first-order prediction of principal-agent theories of politicians and voter behavior. Candidates who just barely won an election (barely became the incumbent) are likely to be ex ante comparable in all other ways to candidates who barely lost, and so their differential electoral outcomes in the next election should represent a true incumbency advantage. The regression discontinuity analysis provides striking evidence that incumbency has a significant *causal* effect of raising the probability of subsequent electoral success - by about 0.4 to 0.45. Simulations - using estimates from a structural model of individual voting behavior - imply that about two-thirds of the apparent electoral success of incumbents can be attributed to voters' valuation of politicians' experience. The quasi-experimental analysis also suggest that heuristic "fixed effects" and "instrumental variable" modeling approaches would have led to misleading inferences in this context.

We have provided you with the data set individ_final.csv. It contains the following variables:

- **yearel**: election year
- **myoutcomenext**: a dummy variable indicating whether the candidate of the incumbent party was elected
- **difshare**: a normalized running variable: *proportion of votes of the party in the previous election* - 0.5. If $difshare > 0$ then the candidate runs for the same party as the incumbent.

Load this data into R and install the package **rdd** to answer the following questions:

Question 11

1/1 point (graded)

Based on the information provided, create a variable for whether the party of the candidate is the same party as the incumbent. What is the proportion of these cases in your data set?

Please round your answer to the second decimal place, i.e. if your answer is 0.8982, round to 0.90 and if it is 0.8922, round to 0.89

0.40 ✓ **Answer:** 0.40

0.40

Explanation

Create a dummy variable on whether the variable $difshare > 0.$ Calculate the sample average of that variable. The answer you should receive is $0.3990,$ which rounds to 0.40

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

One of the main assumptions in RD designs is that there are no jumps in the density of the running variable around the cutoff. The package in R **rdd** has a command **DCdensity**. Run the command in R using *difshare* as the running variable. Refer to the documentation for the command if you have any questions.

Question 12

1/1 point (graded)

What is the difference in the log estimate in heights at the cutpoint?

Please round your answer to the fourth decimal place, i.e. if your answer is 1.03456, please round to 1.0346 and if it is 1.03451, round to 1.0345.

-0.0025 ✓ **Answer:** -0.0025

-0.0025

Explanation

When you run the command in R, you should run it with the option **ext.out = TRUE.** Then, the variable **theta** corresponds to a difference of $-0.002470001,$ which rounds to $-0.0025.$

Submit You have used 1 of 2 attempts

✓ Correct (1/1 point)

Question 13

1/1 point (graded)

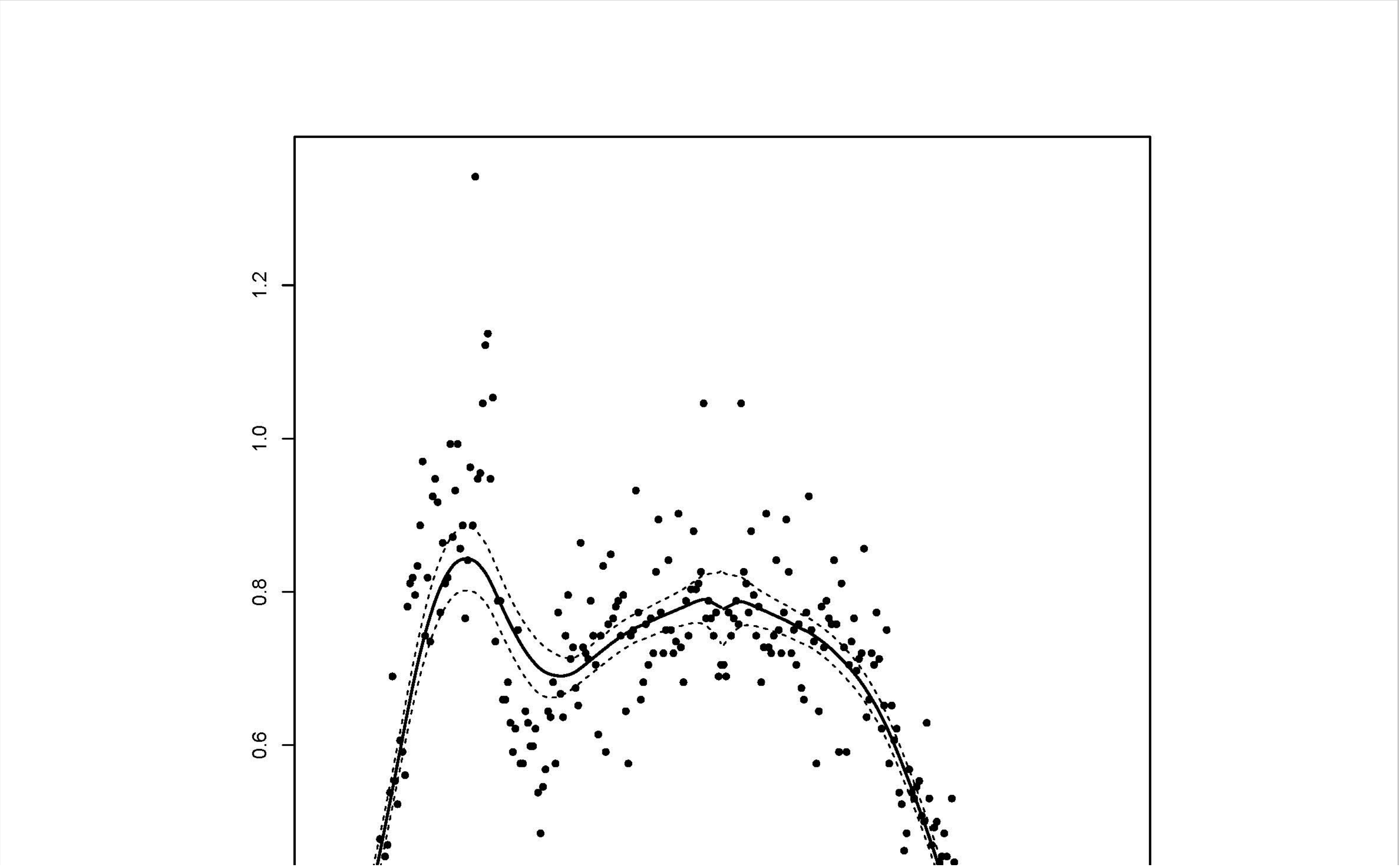
Can you reject the null hypothesis that this difference is equal to zero?

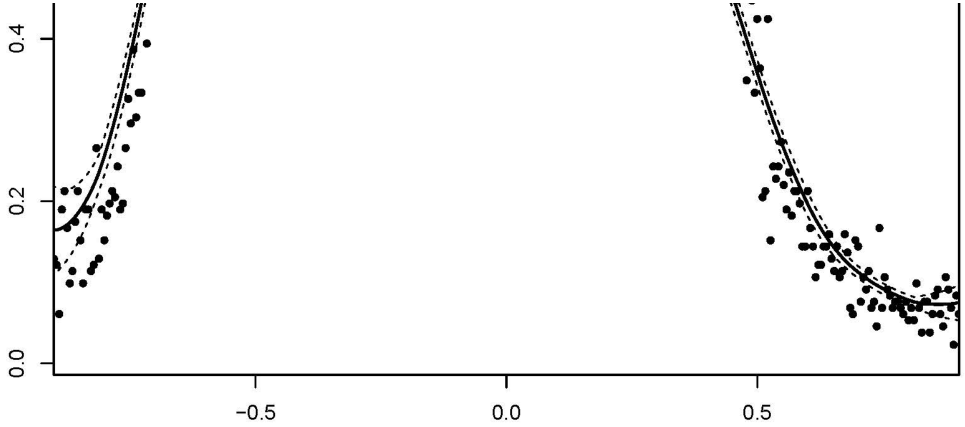
○ a. Yes

◉ b. No ✓

Explanation

According to the output the p-value associated to this test is $0.9620681.$ Thus, you can't reject the null hypothesis that this difference is equal to zero. This implies that the assumption that there is no differential density around the cutoff holds in this case. We can also see this in the plot produced by the R command and that is presented below:





Submit You have used 1 of 1 attempt

✔ Correct (1/1 point)

Now, keep only the observations within 50 percentage points of the cutoff. Also, create in R the required variables to run the following models:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0, i} + \varepsilon_i$$ **(model 1)**

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0, i} + \gamma_1 difshare_i + \varepsilon_i$$ **(model 2)**

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0, i} + \gamma_1 difshare_i + \delta_1 difshare_i \times \mathbf{1}_{difshare \geq 0, i} + \varepsilon_i$$ **(model 3)**

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + \varepsilon_i \qquad \textbf{(model 4)}$$

$$\begin{aligned} y_i \quad = \quad & \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 \\ & + \delta_1 difshare_i \times \mathbf{1}_{difshare \geq 0,i} + \delta_2 difshare_i^2 \times \mathbf{1}_{difshare \geq 0,i} + \varepsilon_i \end{aligned} \qquad \textbf{(model 5)}$$

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + + \gamma_3 difshare_i^3 + \varepsilon_i \qquad \textbf{(model 6)}$$

$$\begin{aligned} y_i \quad = \quad & \beta_0 + \beta_1 \mathbf{1}_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + \gamma_3 difshare_i^3 \\ & + \delta_1 difshare_i \times \mathbf{1}_{difshare \geq 0,i} + \delta_2 difshare_i^2 \times \mathbf{1}_{difshare \geq 0,i} \\ & + \delta_3 difshare_i^3 \times \mathbf{1}_{difshare \geq 0,i} + \varepsilon_i \end{aligned} \qquad \textbf{(model 7)}$$

Where y_i corresponds to the **myoutcomenext** variable in the data set, and $\mathbf{1}_{difshare \geq 0}$ to a dummy variable that indicates whether the party of the candidate won in the previous election.

Question 14

1/1 point (graded)

For which of the models do you find that the effects of party incumbency over re-election is greater than 0.6? (Select all that apply)

☑ Model 1

☑ Model 2

☑ Model 3

☑ Model 4

□ Model 5

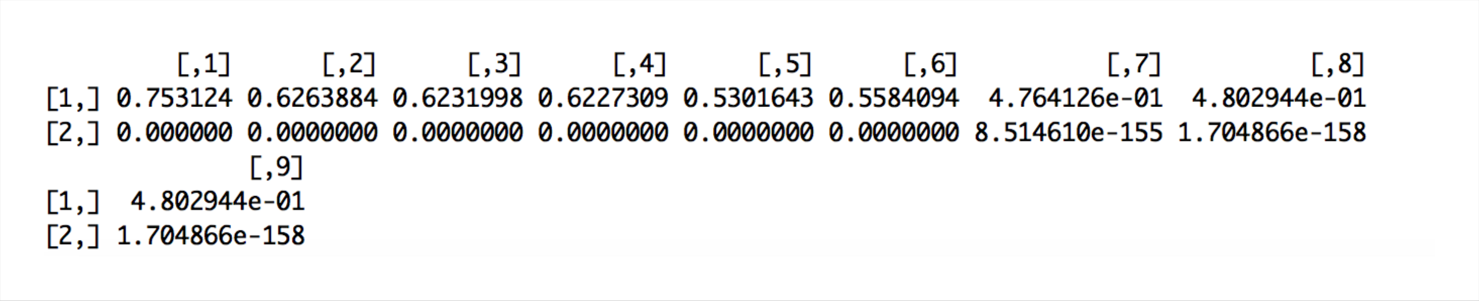
□ Model 6

□ Model 7

✓

Explanation

This code produces the following output in R:



```
         [,1]      [,2]      [,3]      [,4]      [,5]      [,6]          [,7]          [,8]
[1,] 0.753124 0.6263884 0.6231998 0.6227309 0.5301643 0.5584094  4.764126e-01  4.802944e-01
[2,] 0.000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 8.514610e-155 1.704866e-158
              [,9]
[1,]  4.802944e-01
[2,] 1.704866e-158
```

The first row shows the point estimates and the columns are ordered according to the models. The point estimate is larger than 0.6 in the first 4 models.

Submit You have used 1 of 2 attempts

✔ Correct (1/1 point)

Question 15

1/1 point (graded)

For which of the models can you reject the null hypothesis that the incumbent party has no advantage over re-election outcomes with a significance level of 99%? (Select all that apply)

☑ Model 1

☑ Model 2

☑ Model 3

☑ Model 4

☑ Model 5

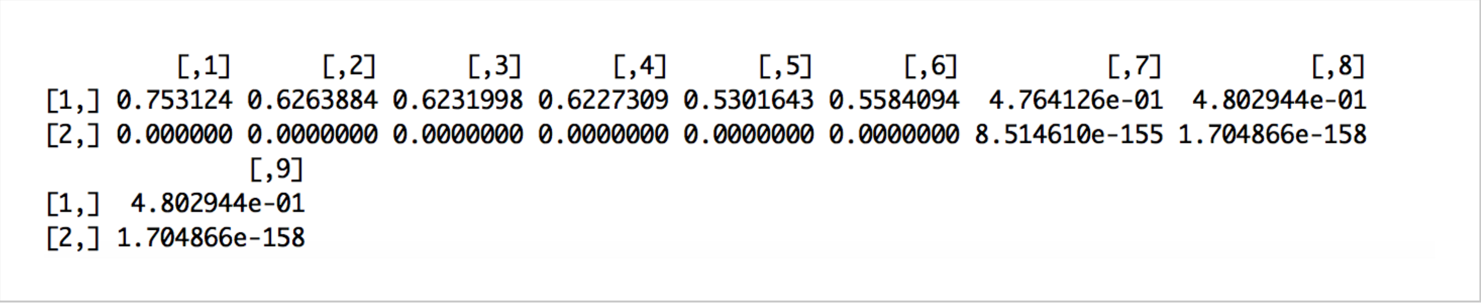
☑ Model 6

☑ Model 7

✓

Explanation

See the code from question 14 that produces the following output in R:



```
           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]          [,7]          [,8]
[1,] 0.753124 0.6263884 0.6231998 0.6227309 0.5301643 0.5584094  4.764126e-01  4.802944e-01
[2,] 0.000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 8.514610e-155 1.704866e-158
              [,9]
[1,]  4.802944e-01
[2,] 1.704866e-158
```

The second row shows the p-value of the null hypothesis $\beta_1 = 0$. In all cases, this is lower than 0.01. This implies that we can reject the null hypothesis in all cases.

Submit You have used 1 of 2 attempts

✓ Correct (1/1 point)

Question 16

1/1 point (graded)

Now use the **RDestimate** command in R to estimate the effect non-parametrically. What is the point estimate that you obtain using this command?

Do not round. Please input the answer exactly as it appears in your R output.

0.4707 ✓ **Answer:** 0.4707

0.4707

Explanation

```
model <− RDestimate(myoutcomenext   difshare, data = indiv, subset = abs(indiv$difshare)
<= 0.5)
```

and the output is equal to 0.4707

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 17

1/1 point (graded)

The command also returns the estimate with half and double of the optimal bandwidth, which one of the following values corresponds to the point estimate with **half of the bandwidth**?

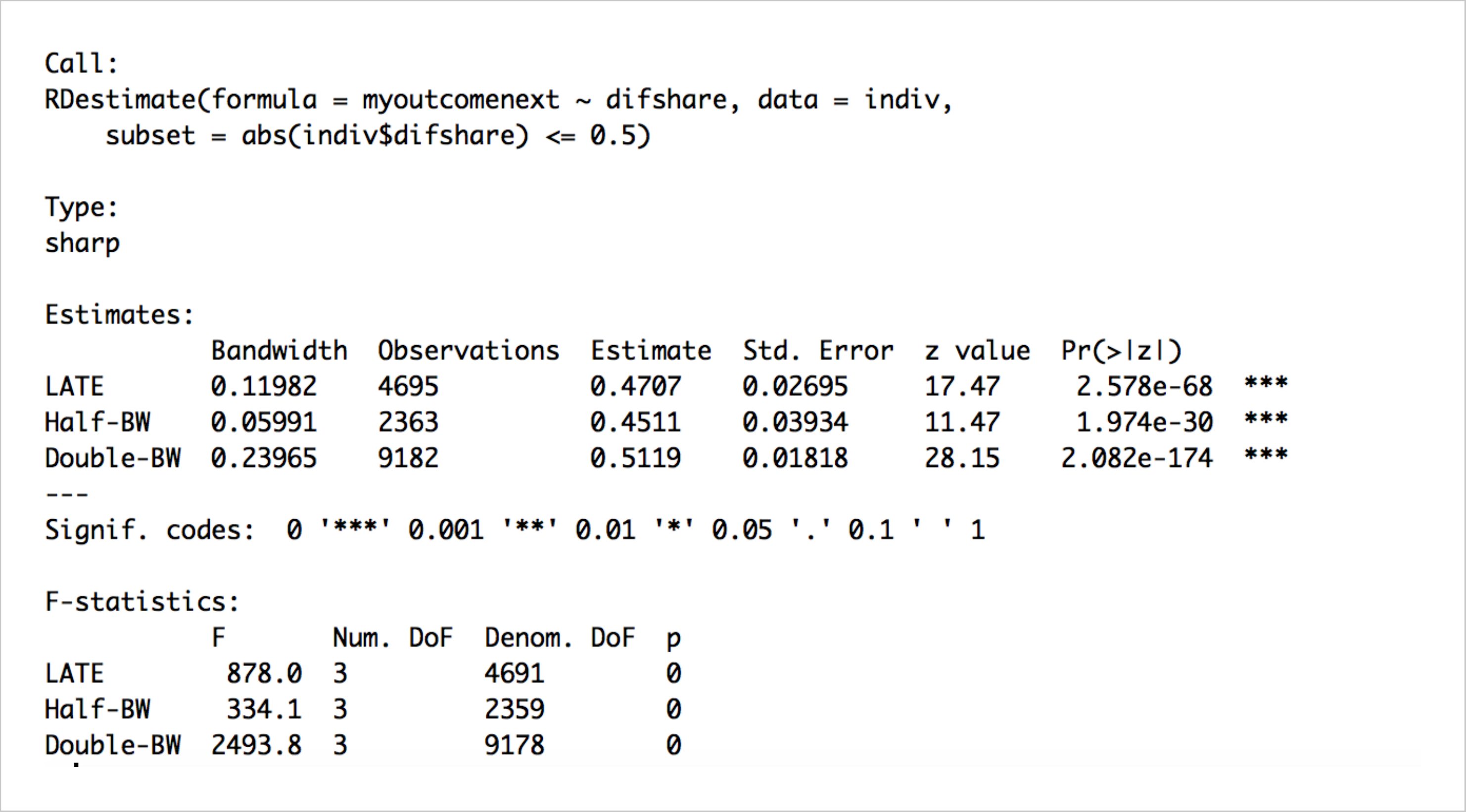
○ a. 0.4707463

● b. 0.4510954 ✓

○ c. 0.5118950

Explanation

This is the output that we get when we run the non-parametric model in R:



```
Call:
RDestimate(formula = myoutcomenext ~ difshare, data = indiv,
    subset = abs(indiv$difshare) <= 0.5)

Type:
sharp

Estimates:
           Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
LATE       0.11982    4695          0.4707    0.02695     17.47     2.578e-68  ***
Half-BW    0.05991    2363          0.4511    0.03934     11.47     1.974e-30  ***
Double-BW  0.23965    9182          0.5119    0.01818     28.15    2.082e-174  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F-statistics:
           F       Num. DoF  Denom. DoF  p
LATE        878.0  3         4691        0
Half-BW     334.1  3         2359        0
Double-BW  2493.8  3         9178        0
```

From there we have that the point estimate within half the size of the optimal bandwidth is 0.4510954.

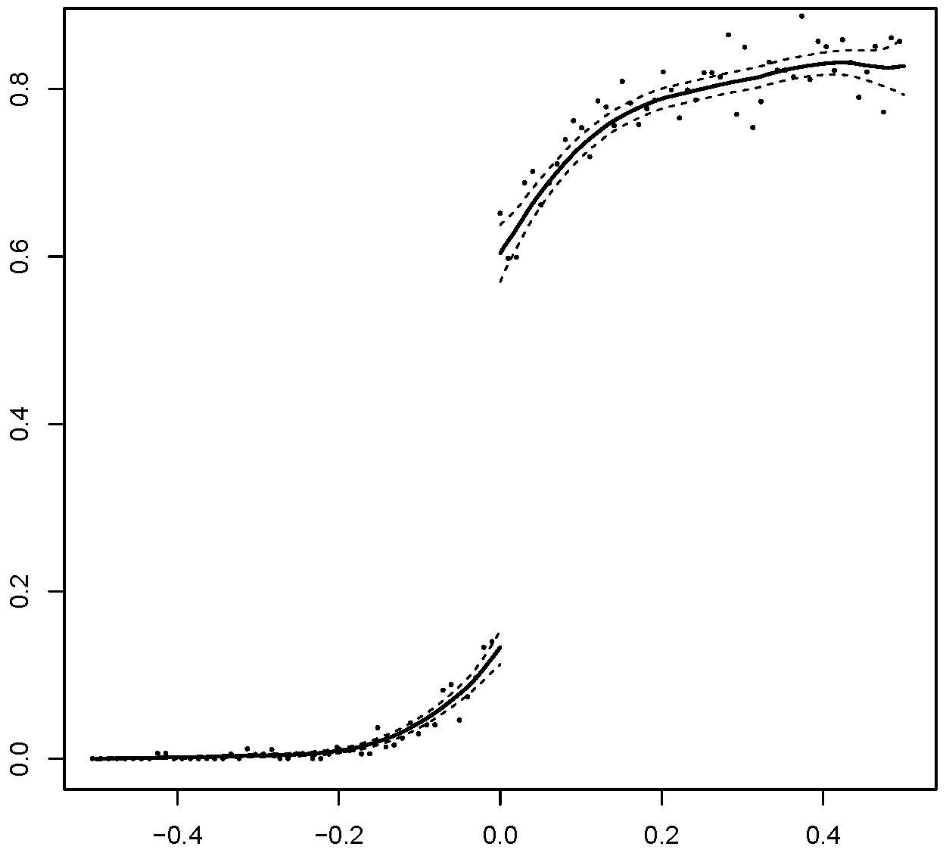
Submit

You have used 1 of 1 attempt

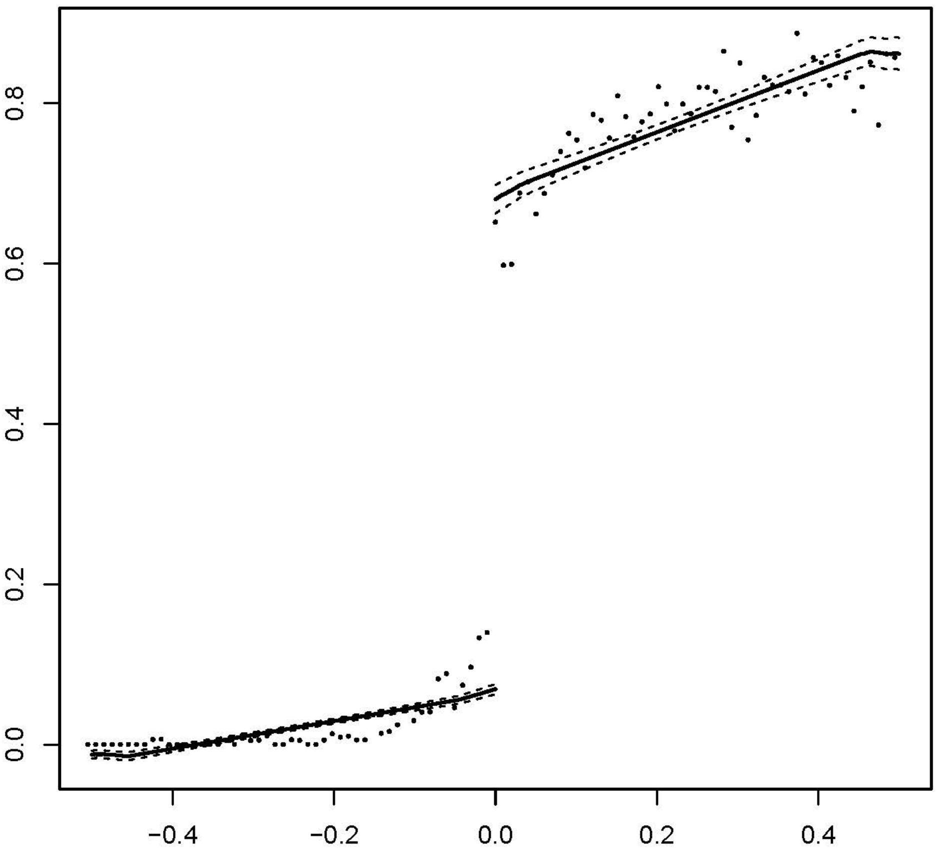
✓ Correct (1/1 point)

Now take a look at the following plots:

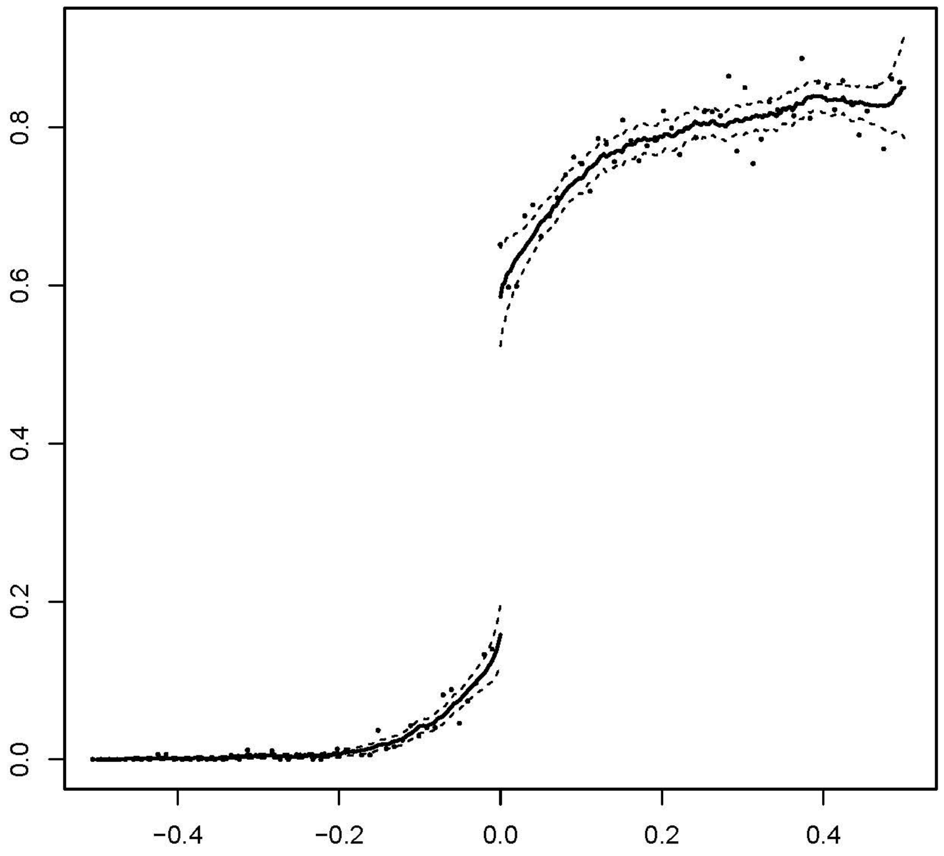
Plot A:



Plot B:



Plot C:



Question 18

0 points possible (ungraded)

One of them was done with the optimal bandwidth, the other one with three times the optimal bandwidth and a third one with one third of the optimal bandwidth. Rank them based on the size of the bandwidth (from the largest to smallest). What is the rank?

○ a. A, B, C

● b. B, A, C ✓

○ c. C, A, B

○ d. B, C, A

Explanation

We can see this visually, as A is very smooth and C is very squiggly. We can also take a look at the R code for each of the plots. Plot A was produced from this code:

```
  model1 <- RDestimate(myoutcomenext   difshare, data = indiv, subset = abs(indiv$difshare)
<= 0.5)
```

Plot B was produced from this code:

```
  model2 <- RDestimate(myoutcomenext   difshare, data = indiv, subset = abs(indiv$difshare)
<= 0.5, kernel = "rectangular", bw = 3*bandwidth)
```