

Difference in differences is a statistical tool broadly used by empirical economists. In this problem, we are going to replicate the results of David Card and Alan Krueger's *"Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania."* The accompanying data ([fastfood.csv](#)) was used to study the effects of an increase in the minimum wage on unemployment. Here is the abstract of the study:

On April 1, 1992, New Jersey's minimum wage rose from \$4.25 to \$5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above \$5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment.

The data set contains the following variables:

Note: NJ refers to New Jersey and PA refers to Pennsylvania

- **chain:** 1=Burger King; 2=KFC; 3=Roy Rogers; 4=Wendy's
- **state:** 1 if NJ; 0 if PA
- **southj:** 1 if in southern NJ
- **centralj:** 1 if in central NJ
- **northj:** 1 if in northern NJ
- **shore:** 1 if on NJ shore
- **pa1:** 1 if in PA, northeast suburbs of Philadelphia

- **pa2:** 1 if in PA, all other areas besides the northeast suburbs of Philadelphia
- **empft:** number of full-time employees before the change in the minimum wage
- **emppt:** number of part-time employees before the change in the minimum wage
- **wage_st:** starting wage in the local (per hour) before the change in the minimum wage
- **empft2:** number of full-time employees after the change in the minimum wage
- **emppt2:** number of part-time employees after the change in the minimum wage
- **wage_st2:** starting wage in the local (per hour) after the change in the minimum wage

Load the data into R and run a linear model in which you compare whether there are differences between fast-food restaurants located in NJ and Pennsylvania prior to the change in the minimum wage in terms of the number of full-time employees and the starting wage.

Question 1

0.0/1.0 point (graded)

What is the average difference between fast-food restaurants located in NJ and Pennsylvania in terms of the number of full-time employees (before the change in minimum wage)?

Note: This is not an absolute difference. Please include signs in your answer and round to the nearest three decimal places.

Answer: -2.6006

Explanation

If we run the following code in R:

```
model1 <- lm(empft ~ state, data = fastfood)
model2 <- lm(emppt ~ state, data = fastfood)
model3 <- lm(wage_st ~ state, data = fastfood)
summary(model1)
summary(model2)
summary(model3)
```

This is the output that we get for model 1:

Call:

```
lm(formula = empft ~ state, data = fastfood)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.316	-5.715	-2.715	3.385	52.285

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.3158	0.9894	10.43	<2e-16 ***
state	-2.6006	1.1020	-2.36	0.0188 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.626 on 390 degrees of freedom

Multiple R-squared: 0.01408, Adjusted R-squared: 0.01155

F-statistic: 5.569 on 1 and 390 DF, p-value: 0.01877

Question 2

0.0/1.0 point (graded)

Is this difference statistically significant at the 1% level?

☐ Yes

☒ No ✓

Explanation

According to the model, the associated p-value for the null hypothesis that this difference is equal to zero is **0.019**. Since it is larger than 0.01, we can't reject this hypothesis at the 1% level, however we can reject it at the 5% level.

[Show answer](#)

Submit

You have used 0 of 1 attempt

i Answers are displayed within the problem

Now, let's look at the starting wage. According to the model that you've run, what is the average wage in Pennsylvania prior to the change?

Please round your answer to the nearest three decimal places.

Answer: 4.62863

Explanation

This corresponds to the intercept in the linear model. In this case we have the following output from R:

```
Call:
lm(formula = wage_st ~ state, data = fastfood)

Residuals:
    Min       1Q   Median       3Q      Max
-0.3786 -0.3574 -0.1074  0.3426  1.1426

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.62863    0.04044  114.460  <2e-16 ***
state        -0.02123    0.04509   -0.471    0.638
---

```


Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3455 on 371 degrees of freedom

(19 observations deleted due to missingness)

Multiple R-squared: 0.0005972, Adjusted R-squared: -0.002097

F-statistic: 0.2217 on 1 and 371 DF, p-value: 0.638

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 4

0.0/1.0 point (graded)

What is the average starting wage in New Jersey?

Please round your answer to the nearest three decimal places.

Answer: 4.6074

Can we reject the null hypothesis that the average starting wage is the same in NJ and Pennsylvania prior to the change in the minimum wage?

☐ Yes

☒ No ✓

Explanation

The p-value associated to the null hypothesis $\beta_1 = 0$ is 0.638. Thus, we can't reject this null hypothesis at the typical confidence levels.

[Show answer](#)

Submit

You have used 0 of 1 attempt

i Answers are displayed within the problem

Now assume that someone pointed out that the northeast suburbs of Philadelphia are very different from the rest of Pennsylvania. This person claims that the model that should be used to estimate the differences between fast food restaurants located in NJ and PA prior to the change is as follows:

$$\text{full time employment} = \beta_0 + \beta_1 \text{state} + \beta_2 \text{pa1} + \beta_3 \text{pa2} + \varepsilon$$

Question 6

0.0/1.0 point (graded)

According to this model above, what would be the average difference in full time employment between the restaurants located in the northeast suburbs of Philadelphia and the rest of Pennsylvania?

☐ $(\beta_2 + \beta_3) - \beta_1$

☐ $\beta_2 - \beta_3$

☐ $\beta_3 - \beta_2$

☐ β_1

☒ It is not possible to tell from this model ✓

Explanation

This model can't be estimated since its variables are perfectly collinear. In particular, all the observations that have the dummy of *state* equal to 0, have that either *pa1* or *pa2* is equal to one. Therefore, we can't identify this difference from this model.

Now, let's run the difference in differences model to see whether the change created a difference in the employment. According to what we saw in the lecture, the model to estimate should be the following:

$$empft_{it} = \beta_0 + \beta_1 state_i + \beta_2 post_t + \beta_3 state_i \times post_t \quad (\text{equation 1})$$

Where $post_t$ is a dummy variable that takes the value of 1 after the change takes place.

Question 7

0.0/1.0 point (graded)

What is the parameter that captures the differences between New Jersey and Pennsylvania prior to the implementation of the change?

☐ $\beta_3 + \beta_1 - \beta_2$

☐ β_2

☐ β_3

☒ β_1 ✓

Explanation

Algebraically we have that from equation (1):

$$\mathbb{E}[empft|state = 0 \& post = 0] = \beta_0$$

$$\mathbb{E}[empft|state = 1 \& post = 0] = \beta_0 + \beta_1$$

$$\mathbb{E}[empft|state = 0 \& post = 1] = \beta_0 + \beta_2$$

$$\mathbb{E}[empft|state = 1 \& post = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

The difference prior to the implementation of the change is given by:

$$\mathbb{E}[empft|state = 1 \& post = 0] - \mathbb{E}[empft|state = 0 \& post = 0] = \beta_1 + \beta_0 - \beta_0 = \beta_1$$

[Show answer](#)

Submit

You have used 0 of 2 attempts

Is this statement correct? In other words, is it true that α_1 in equation (2) is equivalent to β_3 in equation (1)?

☒ Yes ✓

☐ No

Explanation

Intuitively, by taking the difference over time and regressing it on the state, we are also running a difference in differences. We can also show that this is algebraically the same as doing it the "long" way. Recall that from equation (1), we have that:

$$\mathbb{E}[empft|state = 0 \& post = 0] = \beta_0$$

$$\mathbb{E}[empft|state = 1 \& post = 0] = \beta_0 + \beta_1$$

$$\mathbb{E}[empft|state = 0 \& post = 1] = \beta_0 + \beta_2$$

$$\mathbb{E}[empft|state = 1 \& post = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

This implies that:

$$(\mathbb{E}[empft|state = 1 \& post = 1] - \mathbb{E}[empft|state = 1 \& post = 0]) -$$

$$(\mathbb{E}[empft|state = 0 \& post = 1] - \mathbb{E}[empft|state = 0 \& post = 0]) =$$

$$\mathbb{E}[empft_{t=2}|state = 1] - \mathbb{E}[empft_{t=1}|state = 1] -$$

$$\mathbb{E}[empft_{t=2}|state = 0] - \mathbb{E}[empft_{t=1}|state = 0]) =$$

$$\beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 - \beta_1 + \beta_0 - \beta_2 - \beta_0 = \beta_3$$

Similarly from equation (2):

$$\mathbb{E}[empft_{t=2} - empft_{t=1}|state = 0] = \alpha_0$$

$$\mathbb{E}[empft_{t=2} - empft_{t=1}|state = 1] = \alpha_0 + \alpha_1$$

$$\mathbb{E}[empft_{t=2} - empft_{t=1}|state = 1] - \mathbb{E}[empft_{t=2} - empft_{t=1}|state = 0]$$

$$= \alpha_0 + \alpha_1 - \alpha_0 = \alpha_1$$

Thus, $\alpha_1 = \beta_3$.

[Show answer](#)

Now estimate model in equation (2) in R. What value do you obtain for the DiD estimate?

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

Answer: 3.443

Explanation

The following code in R:

```
fastfood$diff_empft <- fastfood$empft2-fastfood$empft  
model5 <- lm(diff_empft ~ state, data = fastfood)  
summary(model5)
```

Produces the following output:

Call:

```
lm(formula = diff_empft ~ state, data = fastfood)
```

Residuals:

Min	1Q	Median	3Q	Max
-32.257	-3.699	0.301	4.301	34.301

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.743	1.181	-2.323	0.02071 *

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 10

0.0/1.0 point (graded)

Assuming that we can interpret the estimate for α_1 as causal and that the minimum wage for these fast-food restaurants is binding, can you conclude that the NJ increase in the minimum wage had a negative effect on full-time employment?

☐ Yes

☒ No ✓

Explanation

Even though we can reject the null hypothesis that $\alpha_1 = 0$ at the 5% level, the estimate for this parameter is positive. Thus, we are finding that the increase in the minimum wage led to an increase in full-time employment.

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.

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

Answer: 0.3990

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

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

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

0.0/1.0 point (graded)

What is the difference in the log estimate in heights at the cutpoint? *Note: this should be a negative number.*

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.

Answer: -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 .

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 13

0.0/1.0 point (graded)

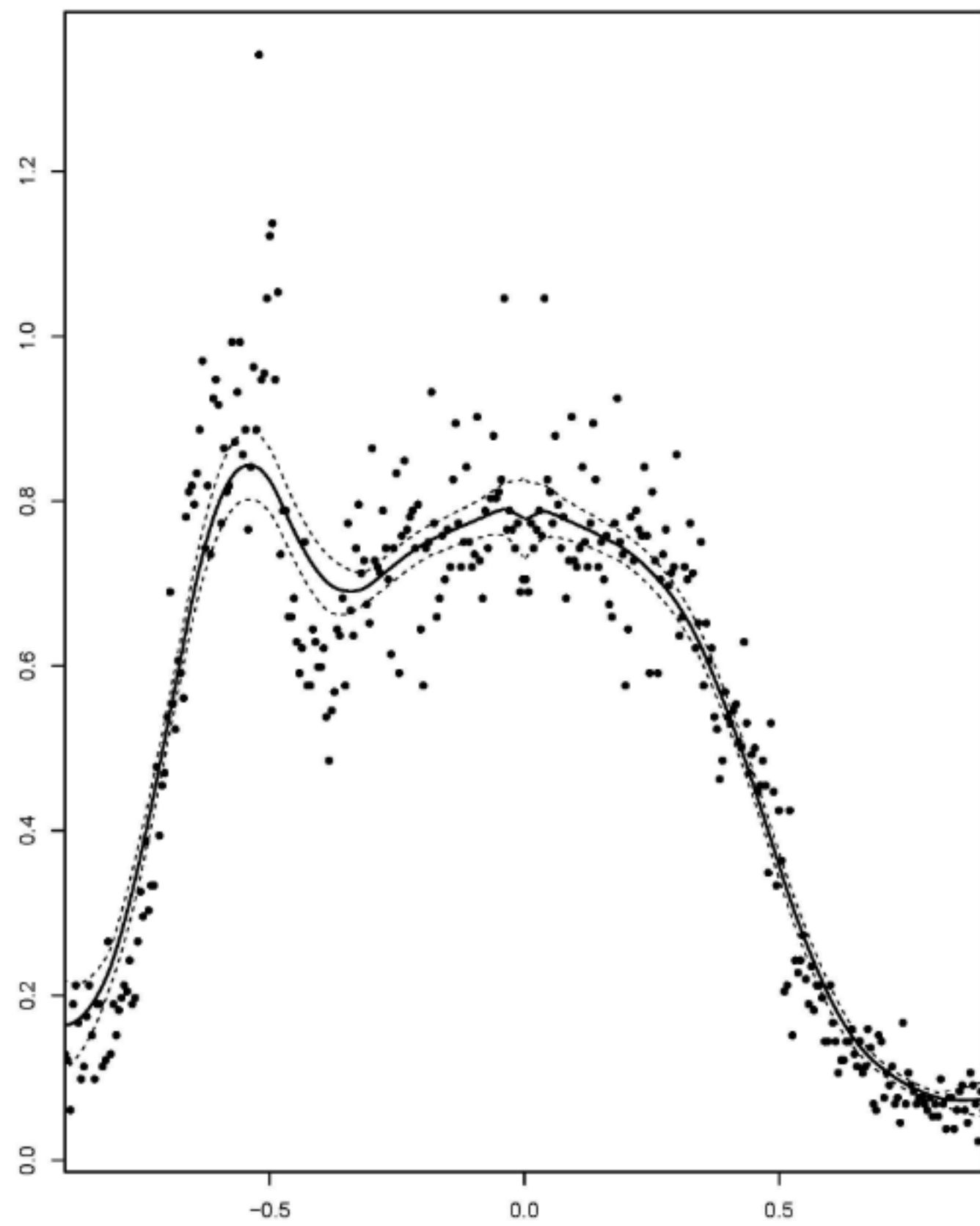
Can you reject the null hypothesis that this difference is equal to zero?

☐ Yes

☒ 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:



Now, keep only the observations within 50 percentage points of the cutoff (the absolute value of **difshare** is less than or equal to 0.5). Also, create in R the required variables to run the following models:

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

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

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

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

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

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

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

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.

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,] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 8.514610e-155 1.704866e-158
```

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,] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 8.514610e-155 1.704866e-158
```

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.

Answer: 0.4707

Explanation

Here is the code:

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

When we input this code into R, we find that the output is equal to 0.4707

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 17

0.0/1.0 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**?

☐ 0.5118950

☒ 0.4510954 ✓

☐ 0.4707463

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.

[Show answer](#)

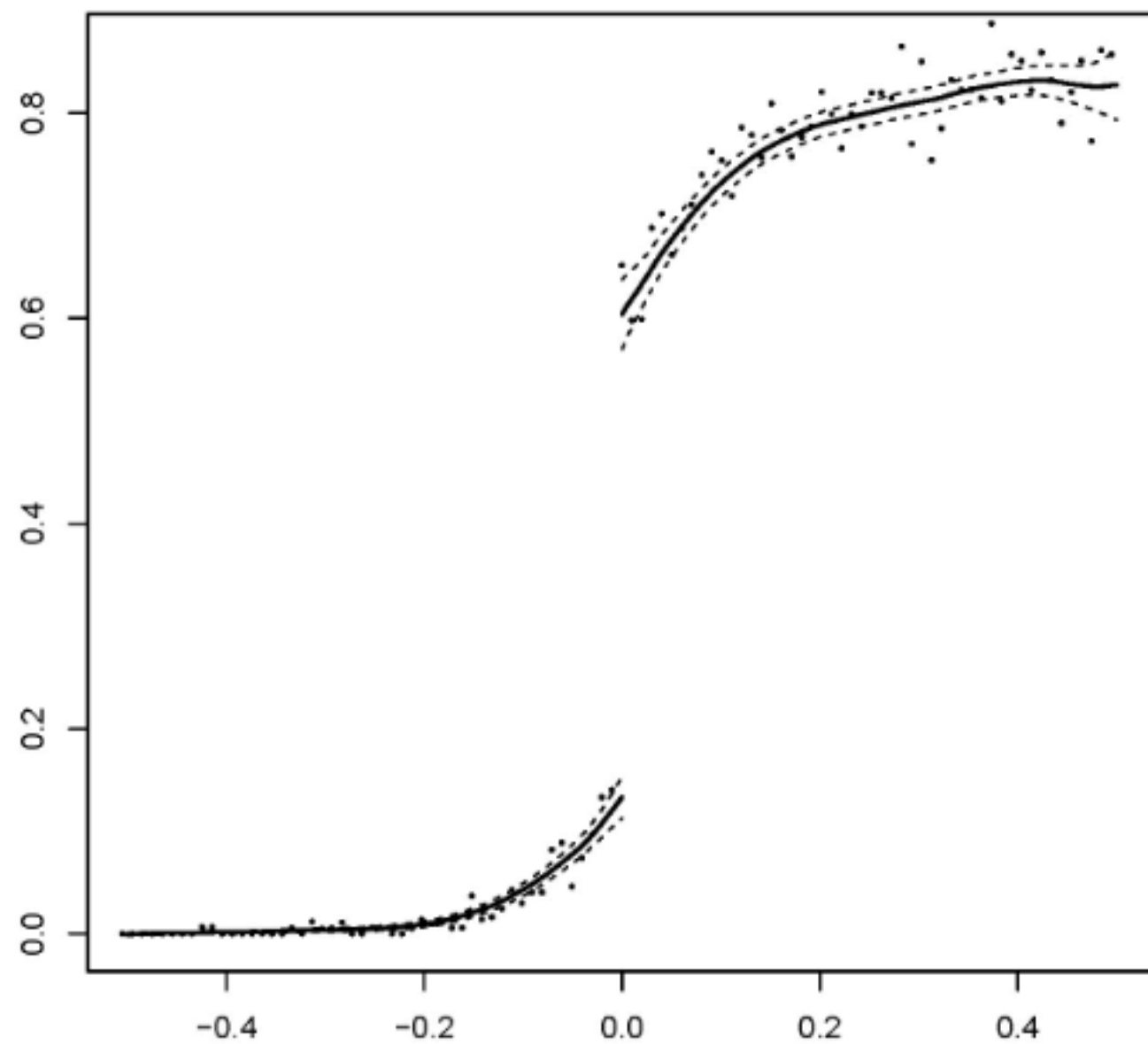
Submit

You have used 0 of 1 attempt

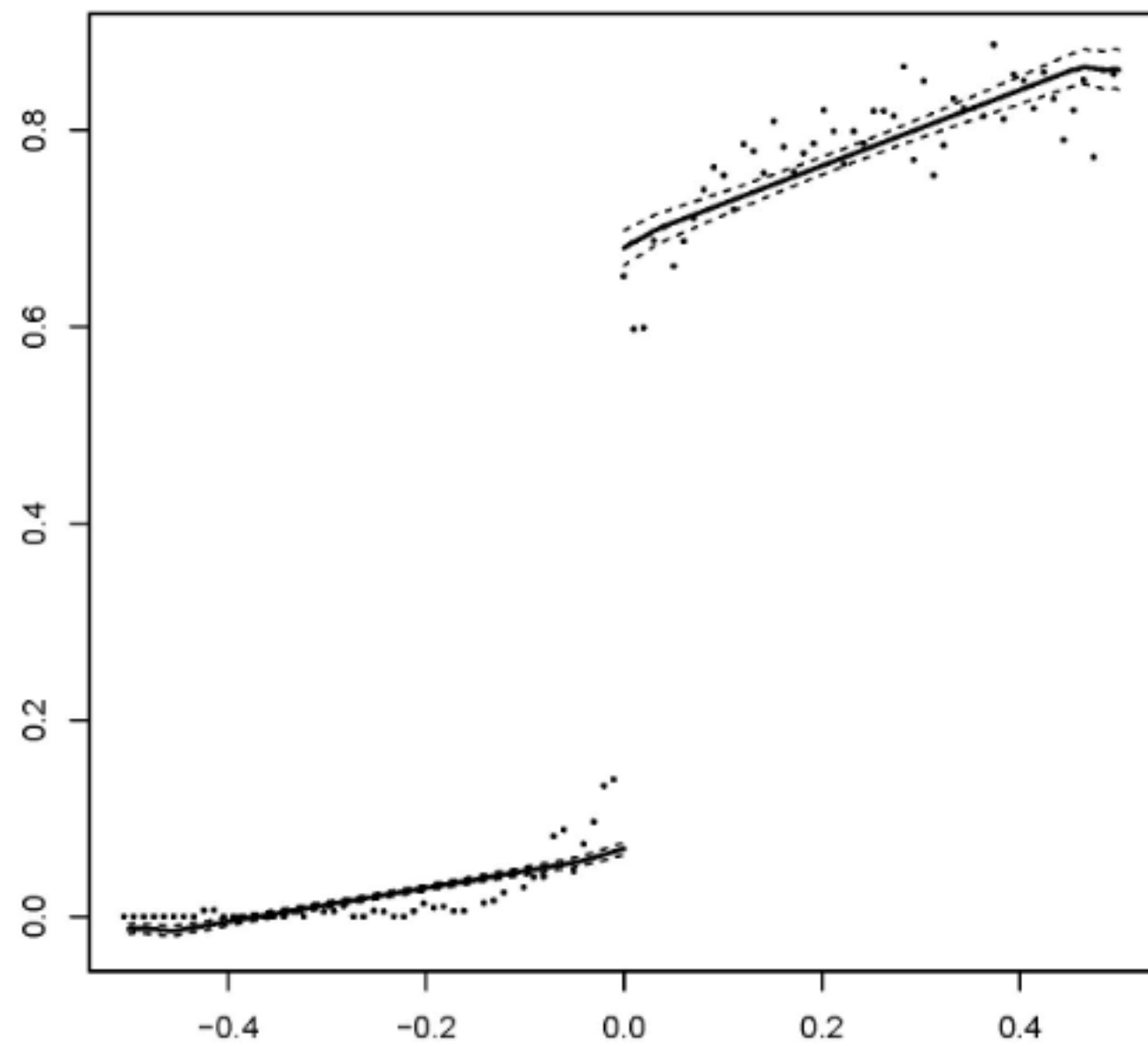
 Answers are displayed within the problem

Now take a look at the following plots:

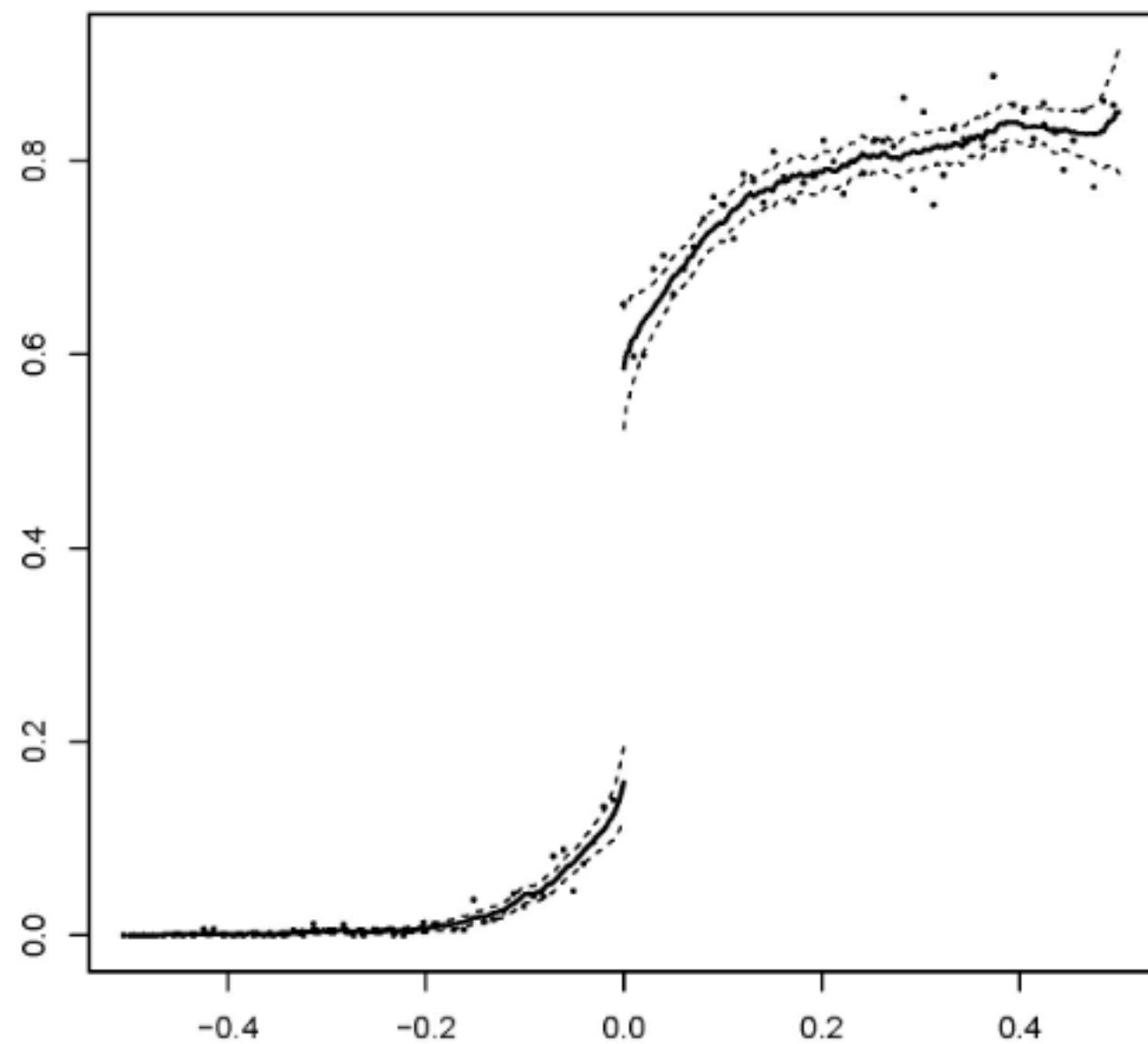
Plot A:



Plot B:



Plot C:



Question 18

0.0/1.0 point (graded)

One of them was done with the optimal bandwidth, another one with three times the optimal bandwidth and a third one with one-third ($1/3$) of the optimal bandwidth. Rank them based on the size of the bandwidth (from the largest to smallest). What is the ranking?

☐ B, C, A

☒ B, A, C ✓

☐ C, A, B

☐ A, B, C

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(myoutcome ~ 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)
```

And Plot C was produced from this code:

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