ECON 7710
Homework 6: Regression Discontinuity Designs

General instructions

The `.Rmd` source for this document will be the template for your homework submission. You must submit your completed assignment as a single html document saved as a PDF file, uploaded to eLC by **11:59p on April 10** using the filename `sectiontime_teamnumber_hw1.html` (e.g. `935_1_hw6.html`).

Notes:

- Include the name of each teammate under `author` in the `yaml` .
- Do not alter the formatting code in this template.
- For questions requiring analytical solutions, you can type them in using markdown math code. Or, you can submit handwritten solutions, embedding them in the knitted document as clearly readable images.
- For questions requiring computation, some or all of the required code is included in associated chunks. Modify chunks where and how you are directed.
- For (almost) all questions about R Markdown, consult The Definitive Guide (https://bookdown.org/yihui/rmarkdown/).
- The `setup` chunk above indicates the packages required for this assignment.
- You will find a description of the variables in the referenced dataset through the Help tab in the Plot pane of RStudio.
- **Switch `eval` to `TRUE` in the global options command to execute code chunks**.

Ride-share platforms like Uber and Lyft use surge pricing to “clear the market” when passenger demand outstrips driver supply. Uber applies a “surge multiplier” to its base fare during peak hours. Higher prices turn away some passengers and attract more drivers, thereby balancing demand and supply.

Uber’s surge pricing operates in increments of 0.1. An algorithm assesses supply and demand conditions, produces a continuous surge index, and then rounds it to the nearest tenth of a point (the minimum multiplier is 1). For example, the algorithm may come up with a surge index of 2.483, which is then rounded to 2.5. Drivers only see the rounded multiplier.

One caveat with surge pricing is that many drivers set a daily income target. Surge pricing may allow such drivers to hit their target sooner and end their driving sessions earlier, counteracting its intended purpose of increasing driver supply. The effect of surge pricing on driver supply is thus an open question.

Your task is to evaluate whether drivers are more likely to stop a driving session when surge pricing is on. The dataset uberrides.csv contains data on two weeks of rides. In the data, a “session” is a set of consecutive trips a driver completed one after another (i.e., if the driver takes a break for a few hours and then starts driving again, this is coded as a new session). For simplicity, I have restricted the data to sessions with a maximum surge multiplier of 1.1. Import Data

```
uber <- read.csv("uberrides.csv")
str(uber)
```

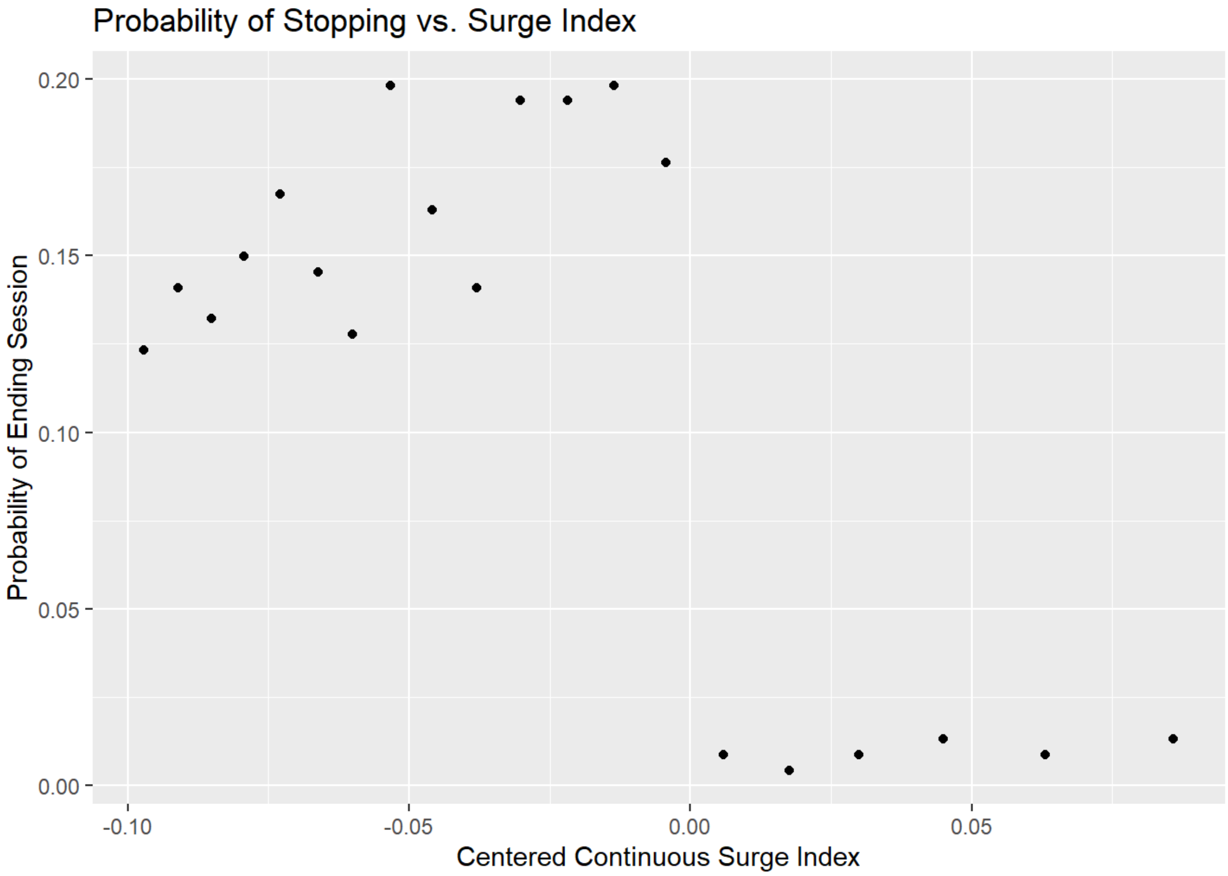
```
## 'data.frame':    46264 obs. of  11 variables:
##  $ sessionid            : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ driverid             : int  422 422 422 422 422 422 422 422 422 422 ...
##  $ ride_partialid       : int  1 2 3 4 5 6 7 8 9 1 ...
##  $ session_cumltime     : num  24.2 58.6 77.1 99 119.8 ...
##  $ num_rides            : int  9 9 9 9 9 9 9 9 9 8 ...
##  $ ride_length          : num  21 30.1 14.9 19.3 19.8 ...
##  $ dayofweek            : int  5 5 5 5 5 5 5 5 5 6 ...
##  $ precip               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ continuous_surge_index: num  0.725 0.716 0.671 0.351 0.38 ...
##  $ surge_mult           : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ lastride             : int  0 0 0 0 0 0 0 0 1 0 ...
```

1. What is the threshold above which the continuous surge index results in a multiplier of 1.1?

Being that the surge is rounded up to the nearest tenth, the threshold above which the continuous surge index results in a multiplier of 1.1 is 1.05. Values from 1.05 up to (but not including) 1.15 will round to 1.1.

2. Create a binned scatter plot looking at the average probability of stopping driving across the threshold.

```
#Variable construction: dummy for being above the threshold
uber$surge <- as.numeric(uber$continuous_surge_index>=1.05)

# Recenter running variable at threshold
uber$multiplier_centered <- uber$continuous_surge_index - 1.05

# create a binned scatter plot
data_subset <- subset(uber, (continuous_surge_index>=.95
                          & continuous_surge_index <=1.15))

data_subset$multiplier_splits <- ntile(data_subset$multiplier_centered, 20)

collapsed_data <- summaryBy(.  ~ multiplier_splits, data=data_subset)

ggplot(collapsed_data, aes(multiplier_centered.mean, lastride.mean)) + geom_point() +
  labs(title = "Probability of Stopping vs. Surge Index") + xlab("Centered Continuous Surge Index") + yl
ab("Probability of Ending Session")
```



3. Build a regression discontinuity model to estimate the effect of surge pricing on drivers' choice of whether to stop driving.

a. What's the effect of a 1.1 multiplier on drivers' probability of ending a ride?
b. How does the probability of ending a ride change with the continuous surge index away from the threshold? Briefly comment on whether the slopes make sense to you.

```
rd_baseline <- lm_robust(lastride ~ surge + multiplier_centered + surge * multiplier_centered, data=ube
r,
                  subset=(continuous_surge_index>=.95
                          & continuous_surge_index <=1.15))

summary(rd_baseline)
```

```
##
## Call:
## lm_robust(formula = lastride ~ surge + multiplier_centered +
##     surge * multiplier_centered, data = uber, subset = (continuous_surge_index >=
##     0.95 & continuous_surge_index <= 1.15))
##
## Standard error type:  HC2
##
## Coefficients:
##                            Estimate Std. Error t value  Pr(>|t|) CI Lower
## (Intercept)                  0.1987    0.01463  13.583 3.196e-41   0.1701
## surge                       -0.1919    0.01517 -12.645 4.841e-36  -0.2216
## multiplier_centered          0.6930    0.22958   3.018 2.554e-03   0.2429
## surge:multiplier_centered  -0.6279    0.24597  -2.553 1.072e-02  -1.1101
##                            CI Upper   DF
## (Intercept)                  0.2274 4531
## surge                       -0.1621 4531
## multiplier_centered          1.1431 4531
## surge:multiplier_centered  -0.1457 4531
##
## Multiple R-squared:  0.04994 ,   Adjusted R-squared:  0.04931
## F-statistic: 157.1 on 3 and 4531 DF,  p-value: < 2.2e-16
```

Answers

a. The formula is:

$$\text{Last Ride} = \beta_0 + \beta_1 \times \text{Surge} + \beta_2 \times \text{Multiplier Centered} + \beta_3 \times \ textSurge * MultiplierCentered + \epsilon$$

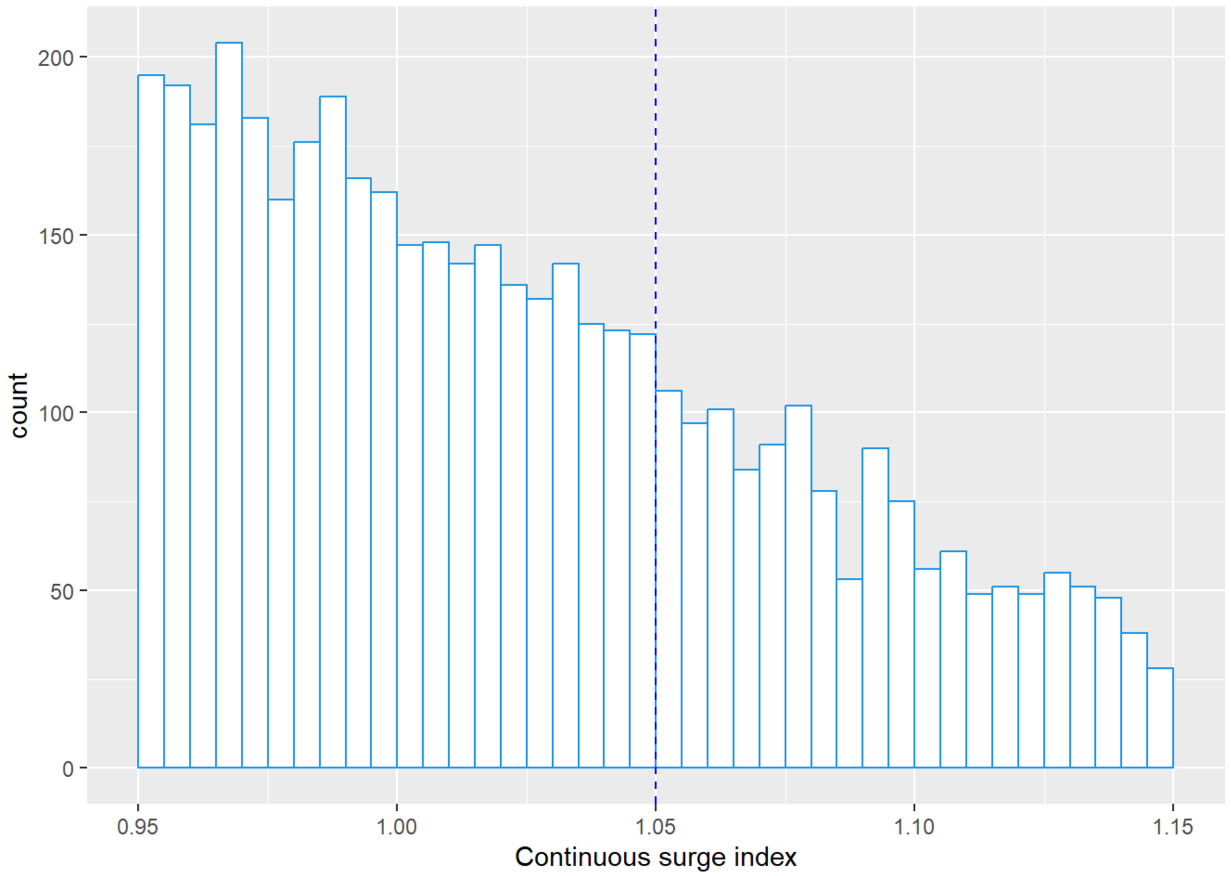
In this model, β_1 represents the causal impact of crossing the 1.05 threshold on a driver’s likelihood of ending their session. The estimated value of β_1 is -0.1919, indicating that when the surge multiplier reaches 1.1, the probability that a driver ends their session drops by approximately 19.2 percentage points.

b. As the continuous surge index approaches the threshold from below, the likelihood of a driver ending their ride session decreases. The coefficient on multiplier_centered is -0.0679, which means that for each one-unit increase in the surge index below the threshold, the probability of ending a session drops by about 6.8 percentage points. Above the threshold, the slope becomes the sum of the multiplier_centered coefficient and the interaction term: -0.0679 + (-0.6279) = -0.6958. This indicates a much steeper decline in the probability of ending a session as the surge index increases beyond the threshold. These results suggest that drivers are significantly less likely to end

4. Check if the RD design is valid

a. Plot the histogram of the continuous surge index. Is there evidence of manipulation at the threshold?
b. Are there other changes at the threshold? Check this for one variable that you consider the most important to check. Why do you choose this variable?
c. Is your result robust to bandwidth selection? Try at least one other bandwidth and comment on your result.

```
# a: histogram of the continuous surge index
ggplot(uber, aes(x=continuous_surge_index)) +
geom_histogram(colour = 4, fill = "white", breaks = seq(from=0.95, to=1.15, by=0.005)) +
  geom_vline(aes(xintercept=1.05), color="blue", linetype="dashed") +
  xlab("Continuous surge index")
```



```
# b: other changes at the threshold
rd_check1 <- lm_robust(ride_length ~ multiplier_centered + surge * multiplier_centered, data = uber,
                  subset=(continuous_surge_index>=.95
                          & continuous_surge_index <=1.15))
summary(rd_check1)
```

```
##
## Call:
## lm_robust(formula = ride_length ~ multiplier_centered + surge *
##     multiplier_centered, data = uber, subset = (continuous_surge_index >=
##     0.95 & continuous_surge_index <= 1.15))
##
## Standard error type:  HC2
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|) CI Lower
## (Intercept)                   19.395     0.2979 65.0984   0.0000   18.811
## multiplier_centered            2.258     5.0190  0.4500   0.6528   -7.581
## surge                         -0.570     0.4672 -1.2200   0.2225   -1.486
## multiplier_centered:surge     -2.194     8.6229 -0.2544   0.7992  -19.099
##                             CI Upper   DF
## (Intercept)                   19.979 4531
## multiplier_centered           12.098 4531
## surge                          0.346 4531
## multiplier_centered:surge     14.712 4531
##
## Multiple R-squared:  0.0007598 , Adjusted R-squared:  9.82e-05
## F-statistic: 1.207 on 3 and 4531 DF,  p-value: 0.3057
```

```
# c: bandwidth selection
rd_check2 <- lm_robust(lastride ~ surge * multiplier_centered + surge * multiplier_centered, data=uber,
                  subset=(continuous_surge_index>=1.00
                          & continuous_surge_index <=1.10))
summary(rd_check2)
```

```
##
## Call:
## lm_robust(formula = lastride ~ surge * multiplier_centered +
##     surge * multiplier_centered, data = uber, subset = (continuous_surge_index >=
##     1 & continuous_surge_index <= 1.1))
##
## Standard error type:  HC2
##
## Coefficients:
##                            Estimate Std. Error t value  Pr(>|t|) CI Lower
## (Intercept)                  0.2025    0.02161  9.3709 1.698e-20   0.1601
## surge                       -0.2003    0.02229 -8.9875 5.248e-19  -0.2441
## multiplier_centered          0.9393    0.70799  1.3267 1.847e-01  -0.4491
## surge:multiplier_centered   -0.6394    0.75251 -0.8497 3.956e-01  -2.1151
##                            CI Upper   DF
## (Intercept)                  0.2448 2237
## surge                       -0.1566 2237
## multiplier_centered          2.3277 2237
## surge:multiplier_centered    0.8363 2237
##
## Multiple R-squared:  0.06961 ,   Adjusted R-squared:  0.06836
## F-statistic: 84.16 on 3 and 2237 DF,  p-value: < 2.2e-16
```

Answers

a. The histogram shows a smooth distribution of the continuous surge index around the 1.05 threshold. There is no visible drop-off at the threshold of 1.05, which suggests no evidence of manipulation at the threshold. This supports the validity of the RD design.

b. I chose `ride_length` because longer trips might affect a driver’s likelihood to stop driving, so it’s important to ensure it doesn’t change discontinuously at the threshold. The regression output shows no significant jump in ride length at the threshold, meaning the discontinuity in ending sessions is not confounded by ride length — supporting the validity of the RD assumptions.

c. When using a narrower bandwidth (1.00 to 1.10), the estimated treatment effect remains negative and significant, though the magnitude may slightly vary. This consistency suggests our RD estimate is robust to reasonable bandwidth changes, which adds credibility to our causal interpretation.