A Regression Discontinuity Design on Uber's surge pricing model based on temperature levels

Fjolle Gjonbalaj

May 2021

Abstract

This paper explores the topic of Uber Surge pricing based on the temperature levels above and below zero degrees celcius. A special focus is provided towards covariates such as the time of the day the uber ride occurred, trip time on Uber, driver's gender, trip type based on uber service used, precipitation on the day of the ride, distance in kilometers with uber, the exchange rate betweed the Russian Ruble and the American US dollar as well as whether or not the ride occurred in the city of Saint Petersburg. A Regression Discontinuity Design is used in order to formalize the difference in uber's surging prices below and above the cutoff point. While most covariates prove to be balanced at the cutoff point based on the covariate balance table, other variables such as time of the day, and the exchange rate aren't. This allows for the isolation of the effect of the temperature value on Uber's surge pricing.

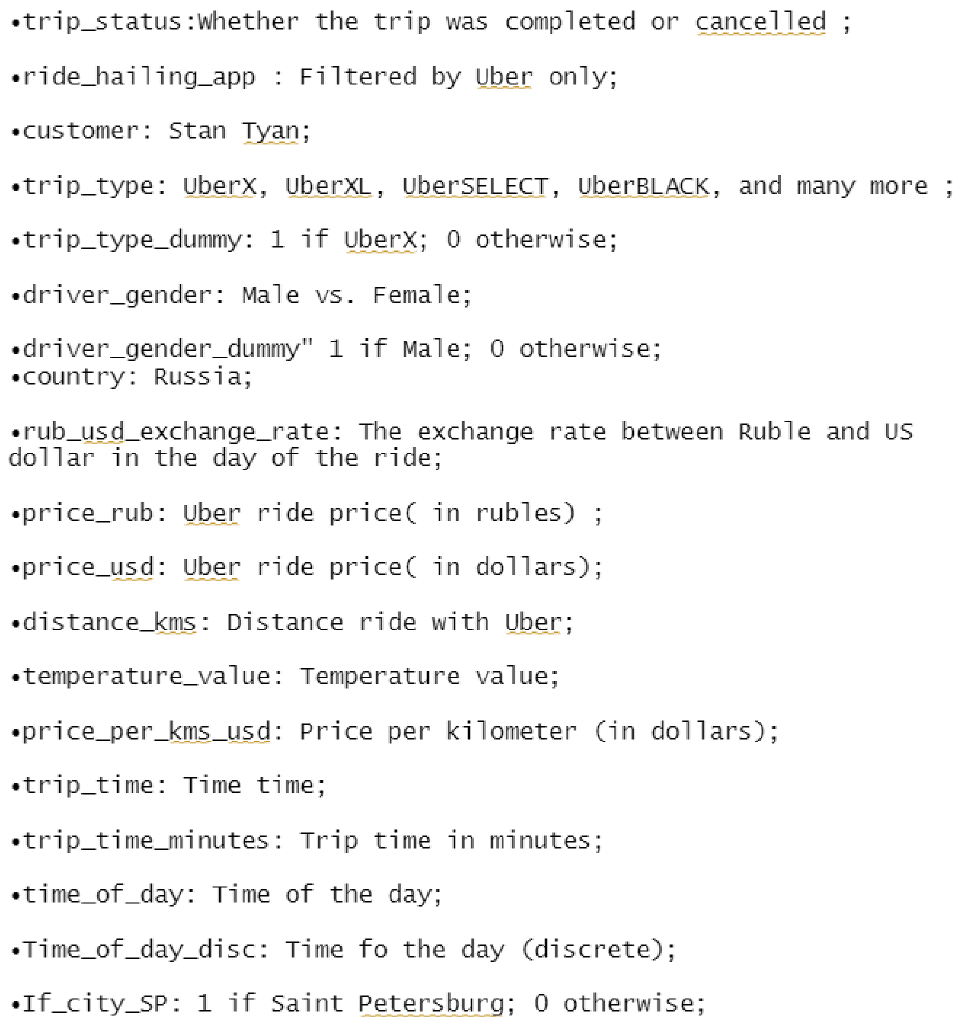
Keywords: Manipulation, Surge Pricing, Regression Discontinuity Design

1 Introduction

Around 10 years ago, in San Francisco, CA, a company named UberCap started operating by enabling riders to hail a car with a smart phone. Since then, the company has extended its operation in 65 countries and more than 700 cities around the world (Uber, 2021).Whether you are doing grocery shopping or visiting a new city, the uber app is a quick and convenient way to connect you with nearby drivers. All the rider needs to do is create an account with just an email address and a phone number. Once that step is done, you click on the app and enter the destination, check for nearby drivers and the price for the ride, confirm your pickup location and wait for your driver to show up(Uber, 2021). In 2014 Uber has announced its operation in Moscow, and today it has a presence in 16 cities in Russia.(3) Weather conditions in Russia tend to be rather extreme with temperature levels significantly below freezing levels during winter to very high temperatures in the summer. That being said, it might be of interest to know whether or not temperature level matters when it comes to pricing of uber rides. Weather condition is not the only factor that affects surge pricing for uber. Multiple other covariates such as time of the day, trip time, trip type, city, distance and exchange rate also have an effect on surge pricing. Since we want to isolate the effect of temperature on prices and need to make sure that the results are robust from endogeneity, we need to control for other covariates having an impact on surge pricing.

This paper is organized as follows: In the first section I describe my data methodology by briefing the data set and the methods used to perform a regression discontinuity design. In the second section, in order to see the effect of temperature levels below and above the freezing point, I begin by creating an indicator variable as a cutoff point for a temperature level of zero degrees celcius. I continue by implementing a manipulation test for the dataset using the rddensity with the local polynomial density estimators implemented by Cattaneo, Jansson and Ma (2020). This command constructs a graphical picture with valid confidence bands of any manipulation(”sorting on the running variable”) in the data below or after the cutoff point. This is a density estimation technique that prevents the data from pre-binning and enables restrictions on other model features (Cattaneo, Janson, 2018). I also use the companion command rdplotdensity in order to visually construct the density plot. Through a histogram plot I show that there seems to be some evidence of non-random heaping that might lead to a biased Regression Discontinuity Design. In the third section I check for covariate balance to find out whether or not covariates are balanced at the cutoff. I pay attention to both the sign and the magnitude of the estimates in order to draw any conclusions. Section four presents and discusses the main results, and section five summarizes the main conclusions of the paper.

The dataset used in this project is personal data from Stan Tyan on his 642 uber rides in Russia between years 2015 to 2018. The Uber rides dataset was downloaded from the Kaggle platform (Kaggle, 2021). However, this data set has some caveats. One major caveat is that the data set is completely based on one individual. Often times Uber price discriminates between regular and non-regular customers by offering special deals to individuals riding uber frequently. This may make the causal effect of temperature on uber prices charged more challenging. That being said, the dataset is sufficient to enable me to show regression discontinuity results. The following variables are used for the analysis that follows:



•trip_status:Whether the trip was completed or cancelled ;

•ride_hailing_app : Filtered by Uber only;

•customer: Stan Tyan;

•trip_type: UberX, UberXL, UberSELECT, UberBLACK, and many more ;

•trip_type_dummy: 1 if UberX; 0 otherwise;

•driver_gender: Male vs. Female;

•driver_gender_dummy" 1 if Male; 0 otherwise;
•country: Russia;

•rub_usd_exchange_rate: The exchange rate between Ruble and US dollar in the day of the ride;

•price_rub: Uber ride price(in rubles) ;

•price_usd: Uber ride price(in dollars);

•distance_kms: Distance ride with Uber;

•temperature_value: Temperature value;

•price_per_kms_usd: Price per kilometer (in dollars);

•trip_time: Time time;

•trip_time_minutes: Trip time in minutes;

•time_of_day: Time of the day;

•Time_of_day_disc: Time fo the day (discrete);

•If_city_SP: 1 if Saint Petersburg; 0 otherwise;

Figure 1: Variables

2 Test for Manipulation in the Running Variable

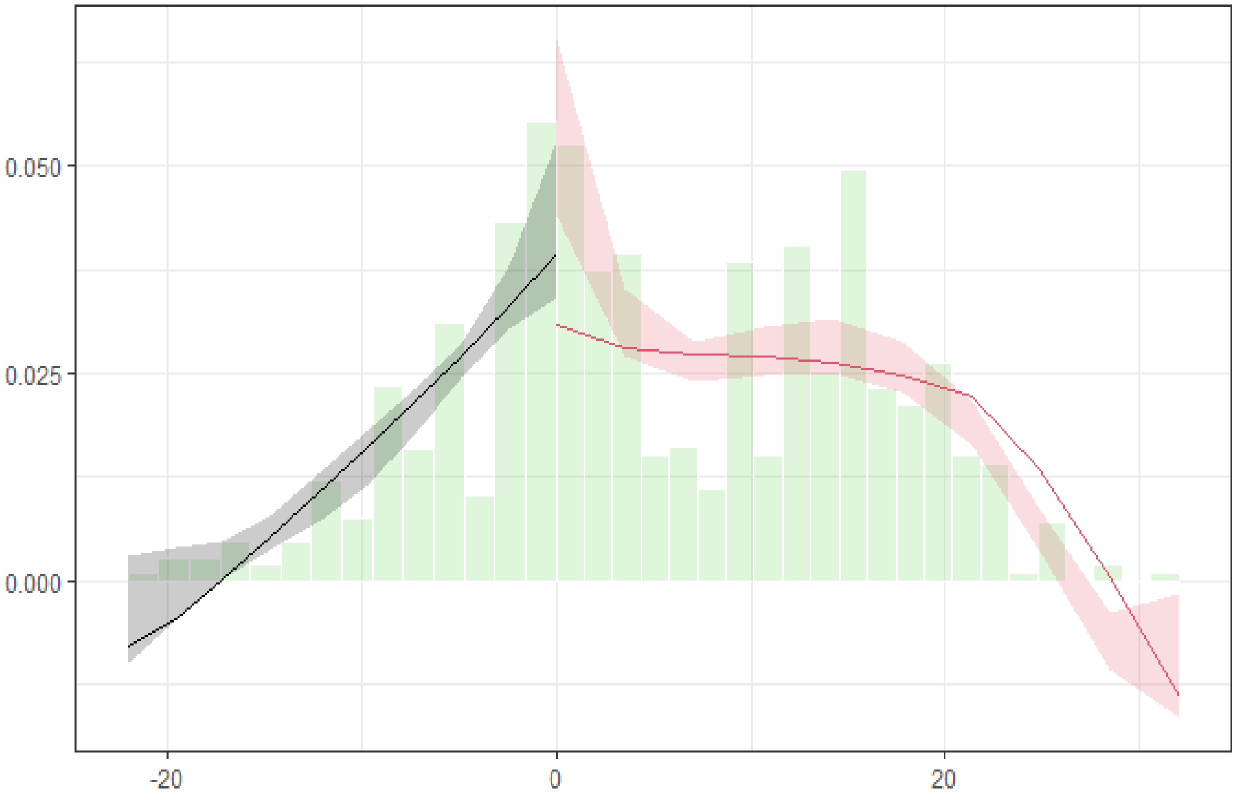


Figure 2: Manipulation Test

The rdplotdensity is used to construct density plots. Its usage is based on the local polynomial density estimator proposed in Cattaneo, Jansson and Ma (2020, 2021a). It is used for manipulation checks in the running variable. It is useful to run such tests when it is otherwise impossible to observe whether Uber prices per kilometer in US dollars differ based on the temperature value. In the regression discontinuity density we can see some clear evidence of manipulation on the running variable above and below the 0 cutoff point. There is a clear jump and a discontinuous pattern around the cutoff point. The McCrary density line falls outside of the confidence interval, further proving the point of manipulation in the running variable of 'Temperature Value'.

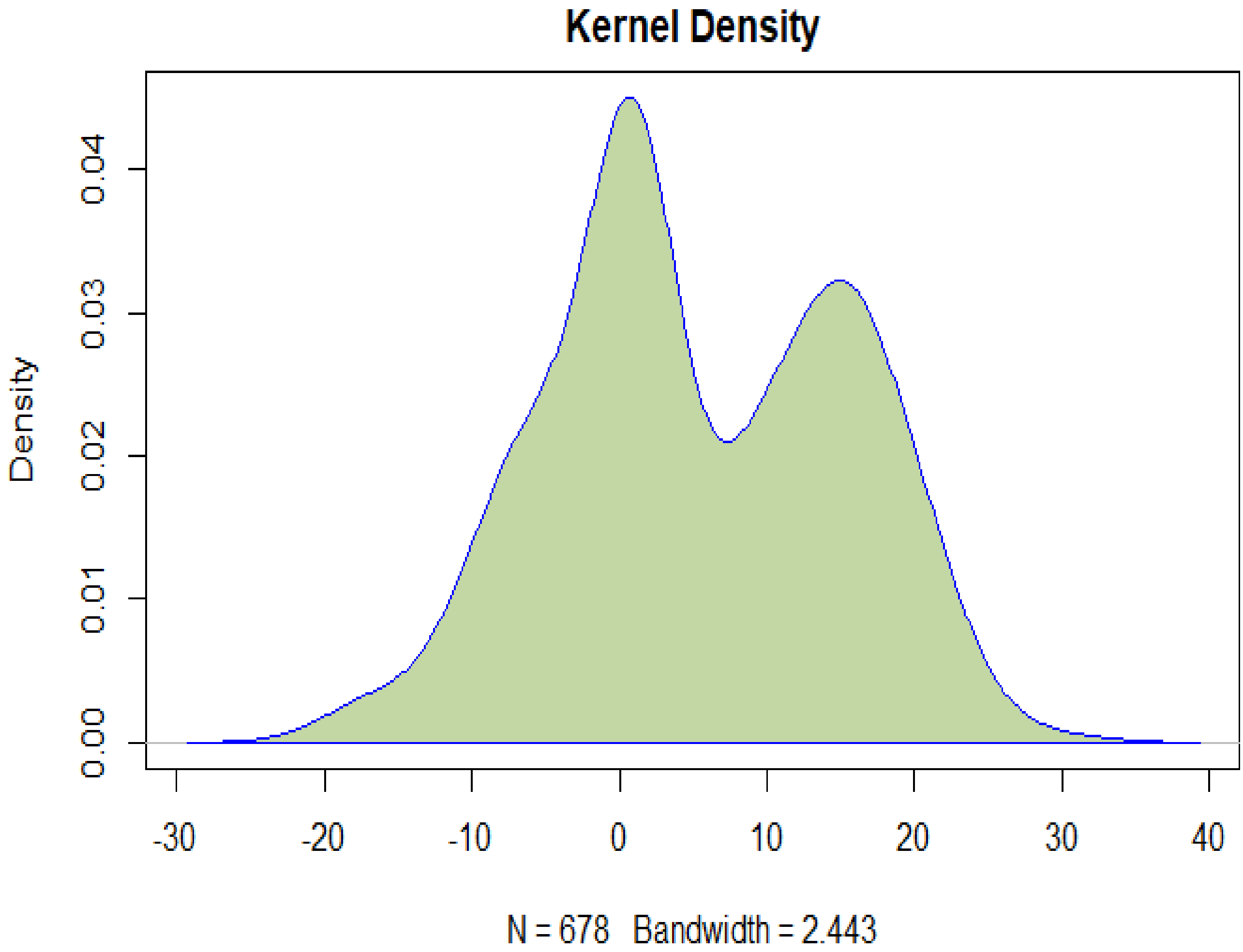
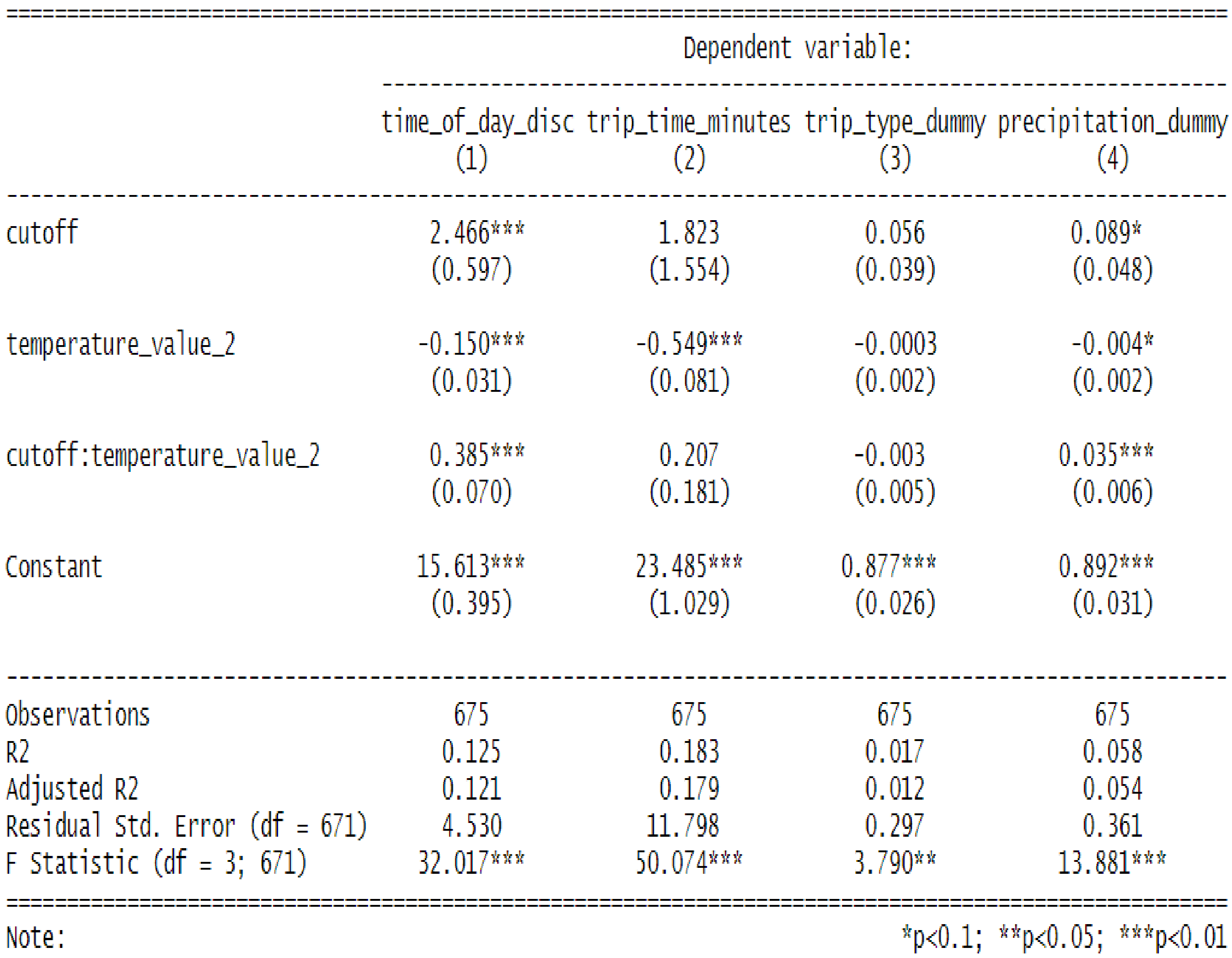


Figure 3: Kernel Density

To show the above results in a density format we plot the Kernel density function. What this figure shows is that there is heaping near the 0 degrees celcius cutoff, which provides further evidence that there is manipulation in the running variable 'Temperature Value'. Intuitively this is indicative that on colder winter days when the temperature value is below 0 degrees celcius, the uber rides are significantly more frequent. From the law of supply and demand, this implies that below 0 degrees celcius there will be less drivers available, and as a result prices will go up. This indicates that Uber's surge pricing becomes effective once the temperature values cross the freezing point of 0 degrees celcius.

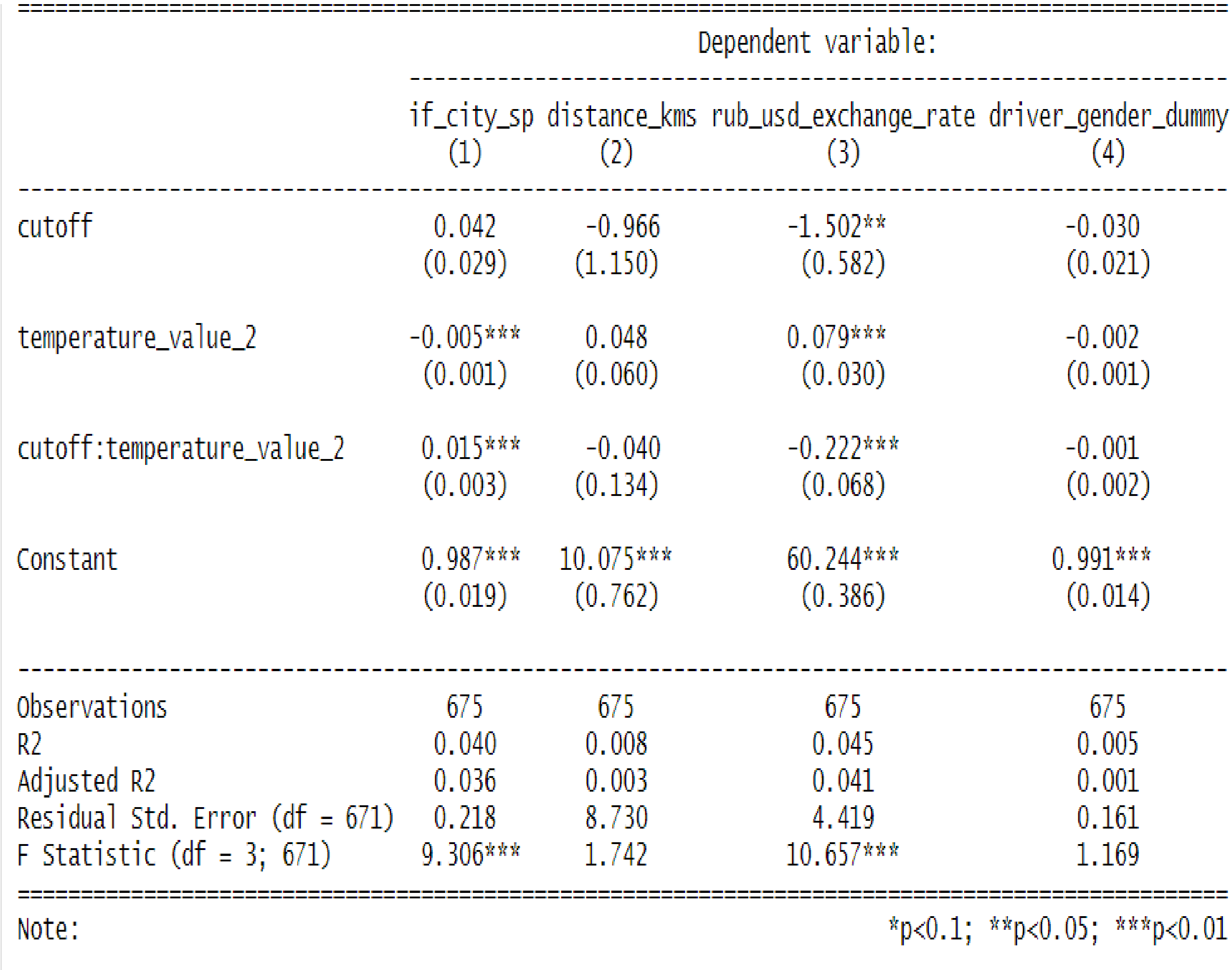
3 Covariate Balance



```
==========================================================================================
                                                   Dependent variable:
                              ------------------------------------------------------------
                              time_of_day_disc trip_time_minutes trip_type_dummy precipitation_dummy
                                    (1)               (2)              (3)              (4)
------------------------------------------------------------------------------------------
cutoff                            2.466***           1.823            0.056            0.089*
                                  (0.597)           (1.554)          (0.039)          (0.048)

temperature_value_2              -0.150***         -0.549***         -0.0003          -0.004*
                                  (0.031)           (0.081)          (0.002)          (0.002)

cutoff:temperature_value_2        0.385***           0.207            -0.003          0.035***
                                  (0.070)           (0.181)          (0.005)          (0.006)

Constant                         15.613***         23.485***         0.877***         0.892***
                                  (0.395)           (1.029)          (0.026)          (0.031)

------------------------------------------------------------------------------------------
Observations                        675               675              675              675
R2                                 0.125             0.183            0.017            0.058
Adjusted R2                        0.121             0.179            0.012            0.054
Residual Std. Error (df = 671)     4.530             11.798           0.297            0.361
F Statistic (df = 3; 671)        32.017***         50.074***         3.790**         13.881***
==========================================================================================
Note:                                                         *p<0.1; **p<0.05; ***p<0.01
```

Figure 4: Covariate Balance

The covariate balance is used to test the extent to which the distribution of covariates is similar across levels of the treatment. There are three main roles that this test performes: 1) a method to optimize the matching, 2) a method to quantify the quality of the found matches, and 3) as a verification that that estimated effects are close to the true effect. Once the covariate balance has been achieved we are able to say more confidently that the estimated effect is less sensitive to any model misspecifications and preferable close to the true treatment effect. The covariate balance test ran test the null hypothesis that the covariate is balanced. We can see from the p-value of the test that the variable 'time of day disc' is highly statistically significant at the *alpha* level of 0.01. This means there is substantial evidence against the null hypothesis against the null hypothesis that this variable is balanced. Similarly the indicator variable of 'Precipitation' is statistically significant at the *alpha* level of 0.1, but not at the *alpha* level of 0.05 of 0.01. Variable 'Trip time in minutes' as well as the indicator variable of 'Trip time' are suggested to be balanced by this covariate balance test.



```
=============================================================================================
                                              Dependent variable:
                               --------------------------------------------------------------
                               if_city_sp distance_kms rub_usd_exchange_rate driver_gender_dummy
                                  (1)          (2)             (3)                  (4)
---------------------------------------------------------------------------------------------
cutoff                           0.042       -0.966         -1.502**              -0.030
                                (0.029)      (1.150)         (0.582)              (0.021)

temperature_value_2            -0.005***      0.048          0.079***             -0.002
                                (0.001)      (0.060)         (0.030)              (0.001)

cutoff:temperature_value_2      0.015***     -0.040         -0.222***             -0.001
                                (0.003)      (0.134)         (0.068)              (0.002)

Constant                        0.987***    10.075***       60.244***             0.991***
                                (0.019)      (0.762)         (0.386)              (0.014)

---------------------------------------------------------------------------------------------
Observations                      675          675             675                  675
R2                               0.040        0.008           0.045                0.005
Adjusted R2                      0.036        0.003           0.041                0.001
Residual Std. Error (df = 671)   0.218        8.730           4.419                0.161
F Statistic (df = 3; 671)       9.306***      1.742          10.657***             1.169
=============================================================================================
Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Figure 5: Covariate Balance

Similarly, the variable 'Rub USD Exchange Rate' is significant at the *alpha* value of 0.05. Hence, the null hypothesis that the covariate is balanced can be rejected. On the other hand, covariates 'If city SP', 'Distance kms' as well as 'Driver gender dummy' are suggested to be balanced by the covariate balance test.

4 Regression Discontinuity plots

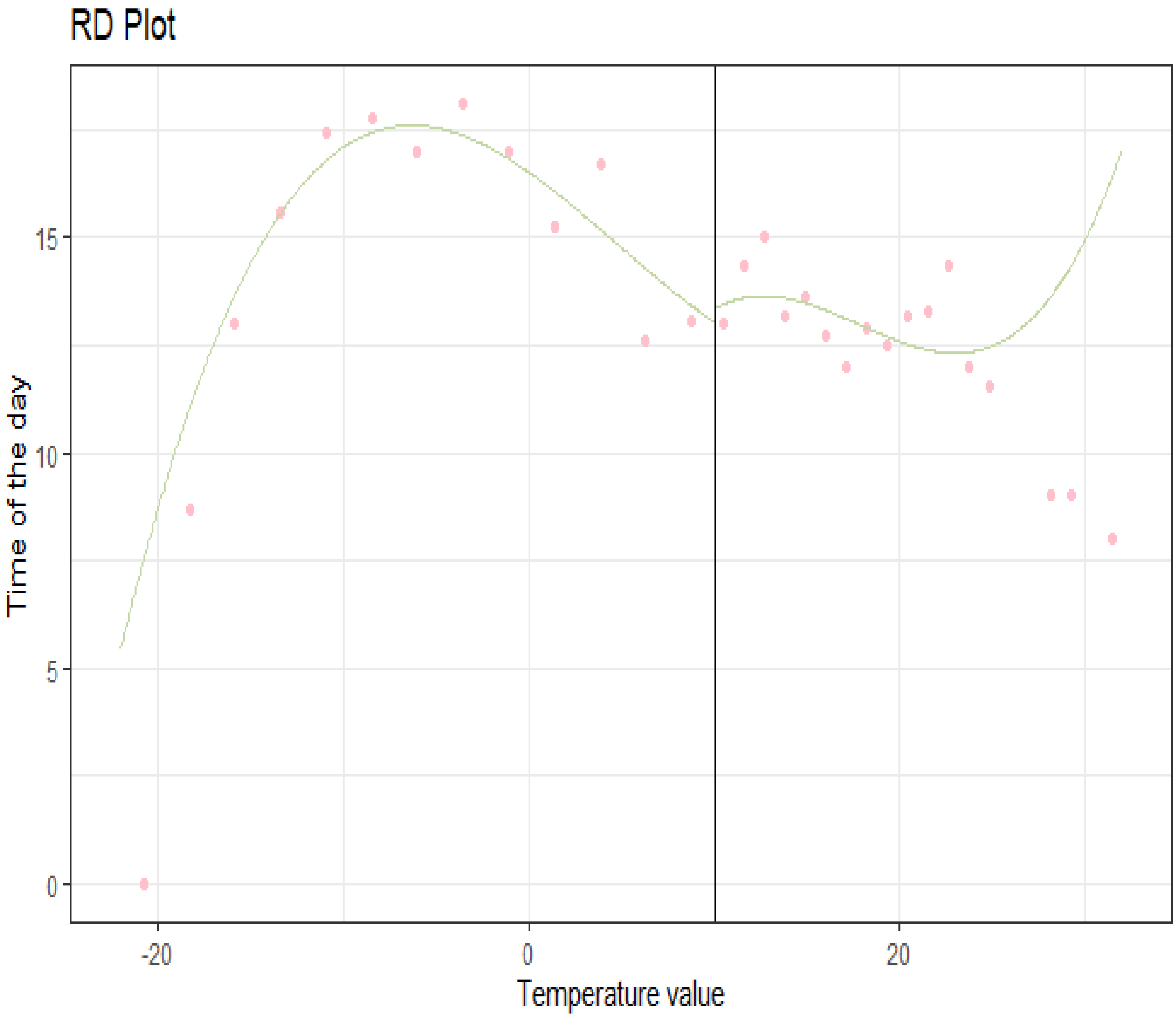


Figure 6: Manipulation in the Exogenous Variable

Figure 6. shows the RD plot for the variable Time of the day and its distribution below and above the cutoff. Although there is a slight jump observed, it does not seem to be of high significance. Thus, it is rather difficult to conclude whether or not the variable Time of day is balance by simply looking at the graph. Moreover, the results from the covariate balance table conflict the results from the RD plot. Treating the variable Trip type, hence, would probably not affect the results we get from the regression discontinuity model.

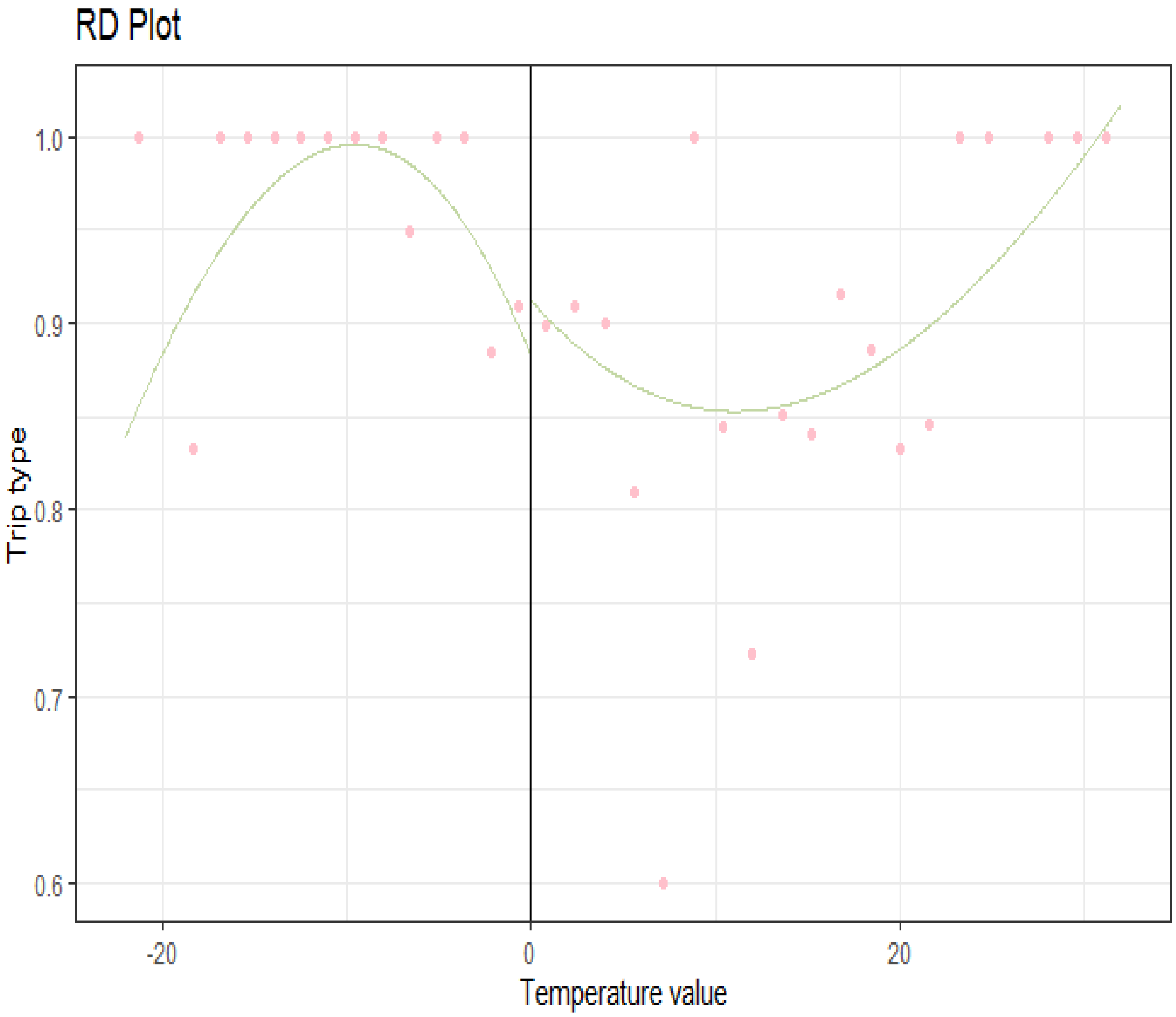


Figure 7: Manipulation in the Exogenous Variable

Figure 7. shows the RD plot for the Trip type variable. It can be observed from the plot that there is a slightly bigger jump below and above the cutoff point. Again, while the covariate balance suggests the variable 'Trip type' is balanced, the RD plot suggests the opposite. Intuitively, this means that it is difficult to argue that the temperature value is the only variable having an effect on Uber's surge pricing below and above the cutoff. Instead, variables such as Trip type also have an effect. Therefore, treating the trip type variable as exogenous would lead to biases in the regression discontinuity model.

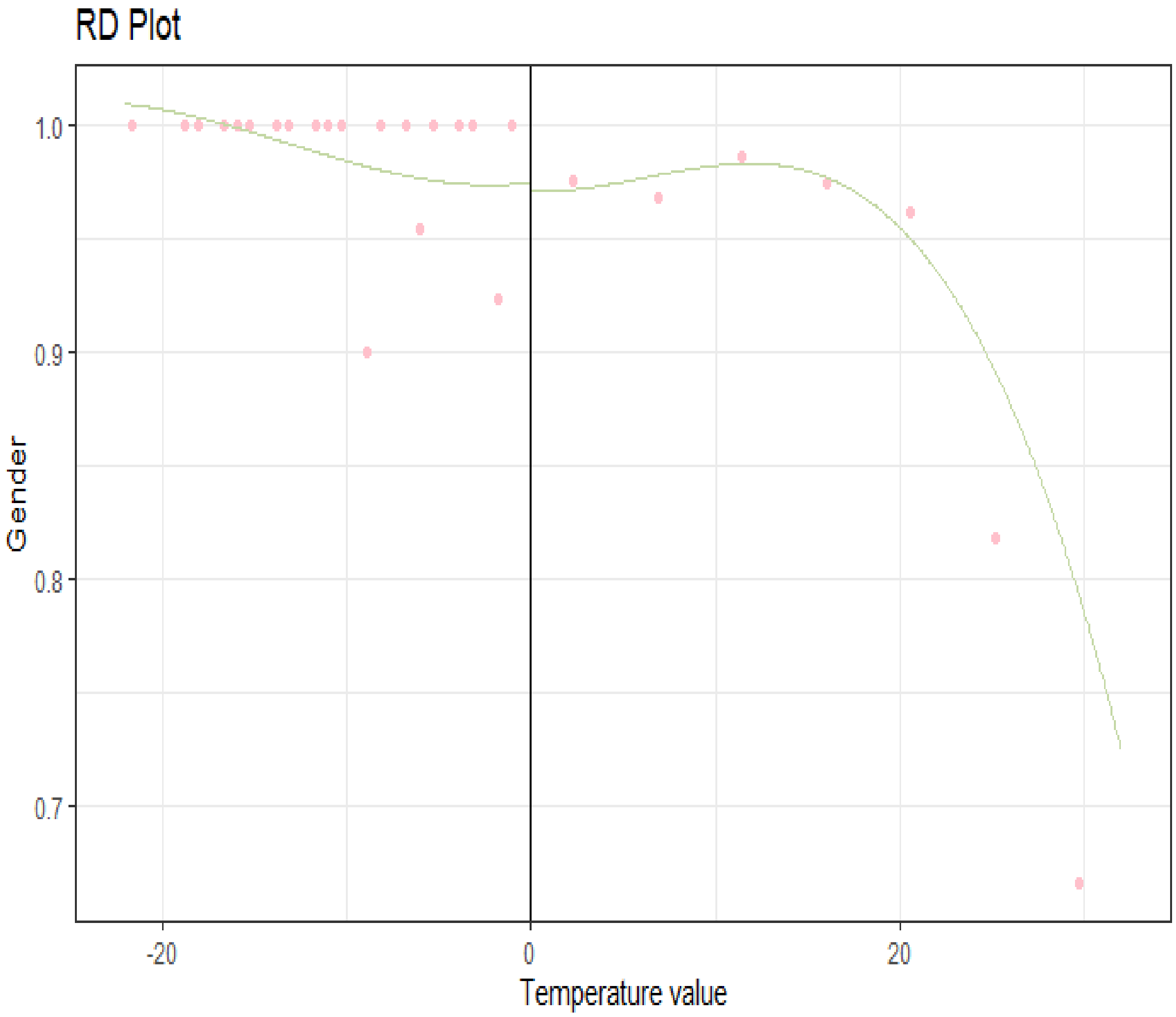


Figure 8: Manipulation in the Exogenous Variable

Figure 8. shows the RD plot for the variable 'Gender'. From the figure we can see that there is almost no jump in the plot below and above the 0 degrees celcius cutoff point. In this case, both the covariate balance table and the RD plot lead us to the same conclusion that the variable 'Gender' is balanced at the cutoff. As such, treating the Gender variable as an exogenous variable would most likely not lead to any biases in the regression discontinuity model.

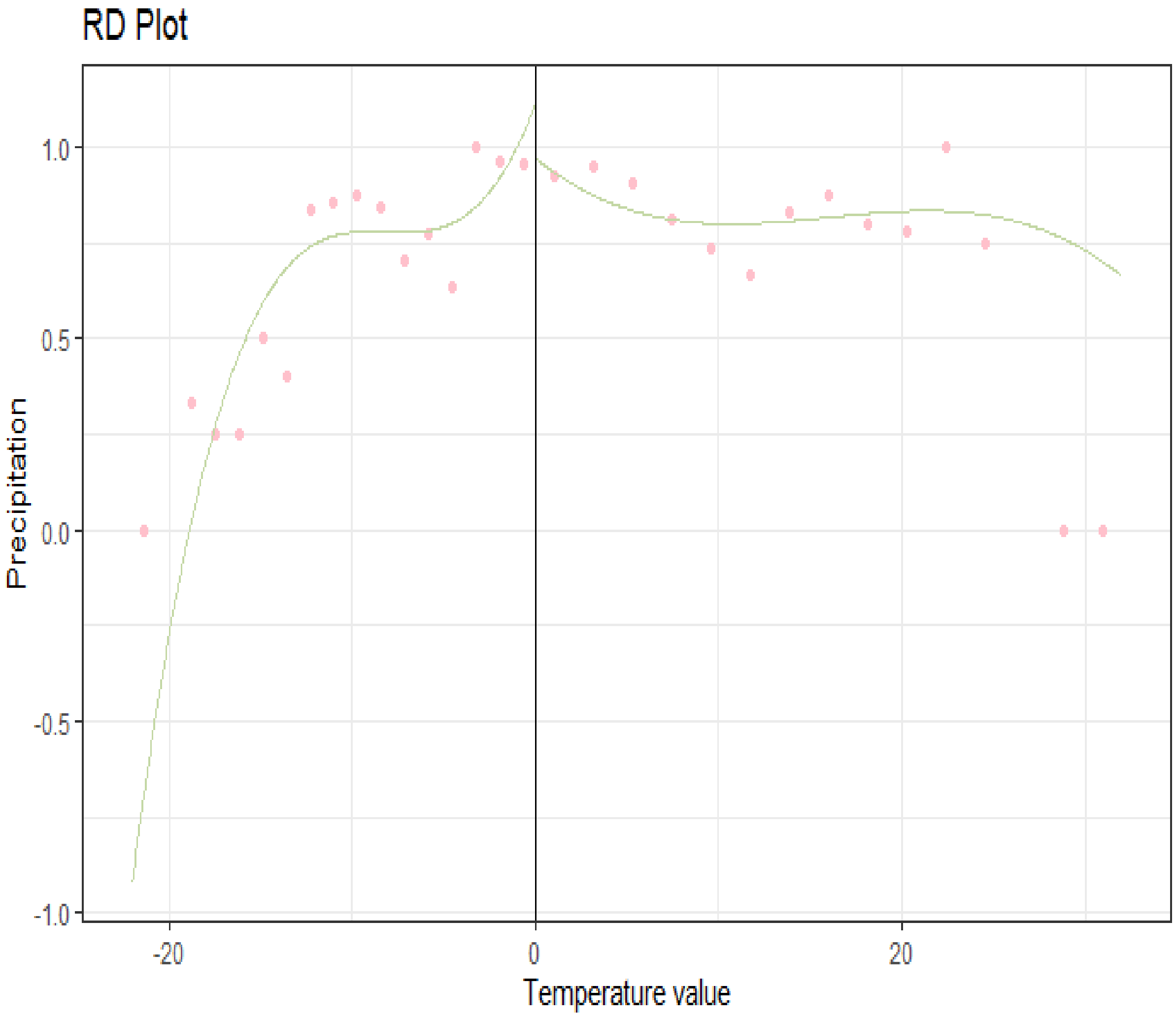


Figure 9: Manipulation in the Exogenous Variable

Figure 9. is the RD plot for the variable 'Precipitation'. This plot indicated there does exist a jump below and above the temperature level of 0 degrees celcius. Although the covariate balance table agrees to this result at the 10 percent significance level, it does not do so for the 1 percent and 5 percent significance levels. Additional investigation in this variable would be necessary to conclude whether or not the covariate is really balanced. Intuitively, this means that it might be difficult to argue that the temperature value is the only variable having an effect on Uber's surge pricing below and above the cutoff. Instead, variables such as Precipitation also have an effect. Therefore, treating Precipitation as exogenous would lead to biases in the regression discontinuity model.

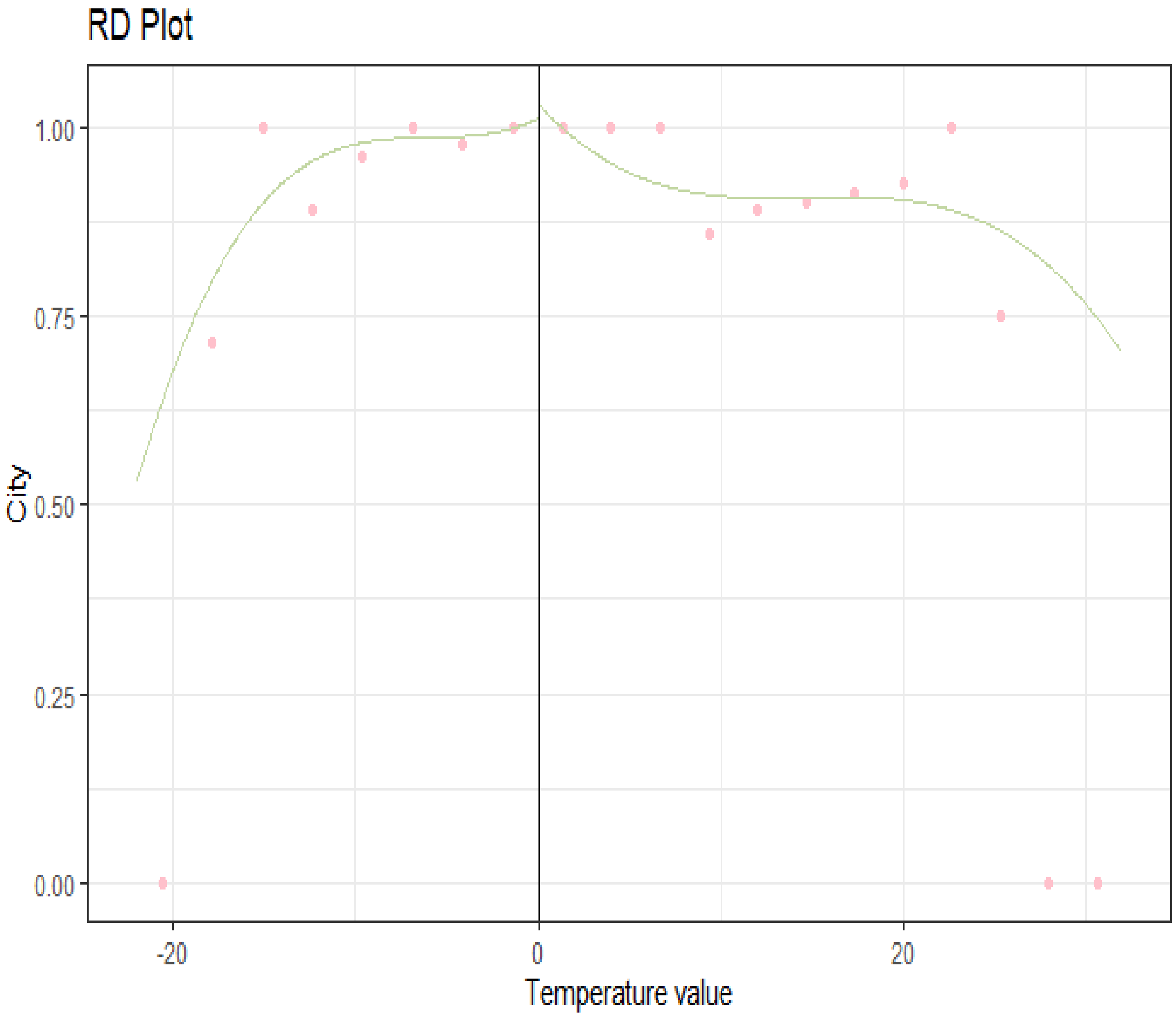


Figure 10: Manipulation in the Exogenous Variable

Figure 10. shows the RD plot for the variable City based on the temperature value. Although it can be observed there is a slight jump on the plot for the covariate, the jump is not highly significant. This result is consistent with the result obtained from the covariate balance table above. As such,it can be said that City would not lead to any significant biases if left outside of the regression discontinuity model.

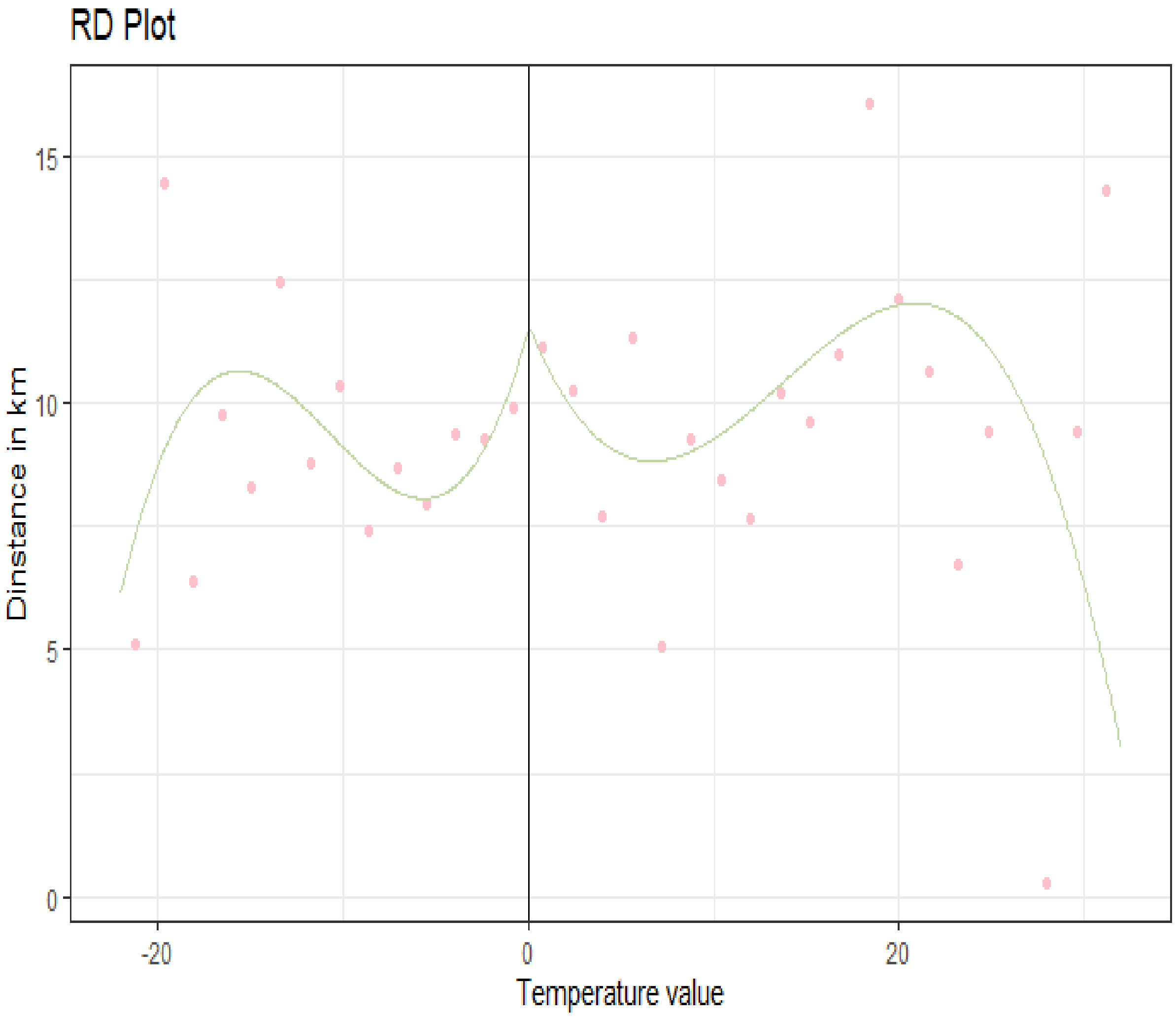


Figure 11: Manipulation in the Exogenous Variable

Figure 11. shows the distance in kilometers based on the temperature value through an RD plot to see whether or not there are significant jumps. From the plot is can be observed that the two sides of the graph are almost mirror reflections of one-another. This result accompanied by the fact that no jump around the zero degrees celcius value, is indicative that the covariate is balanced and consistent with the results from the covariate balance table. It, hence, can be said that the distance in kilometers would not lead to any significant biases if left outside of the regression discontinuity model.

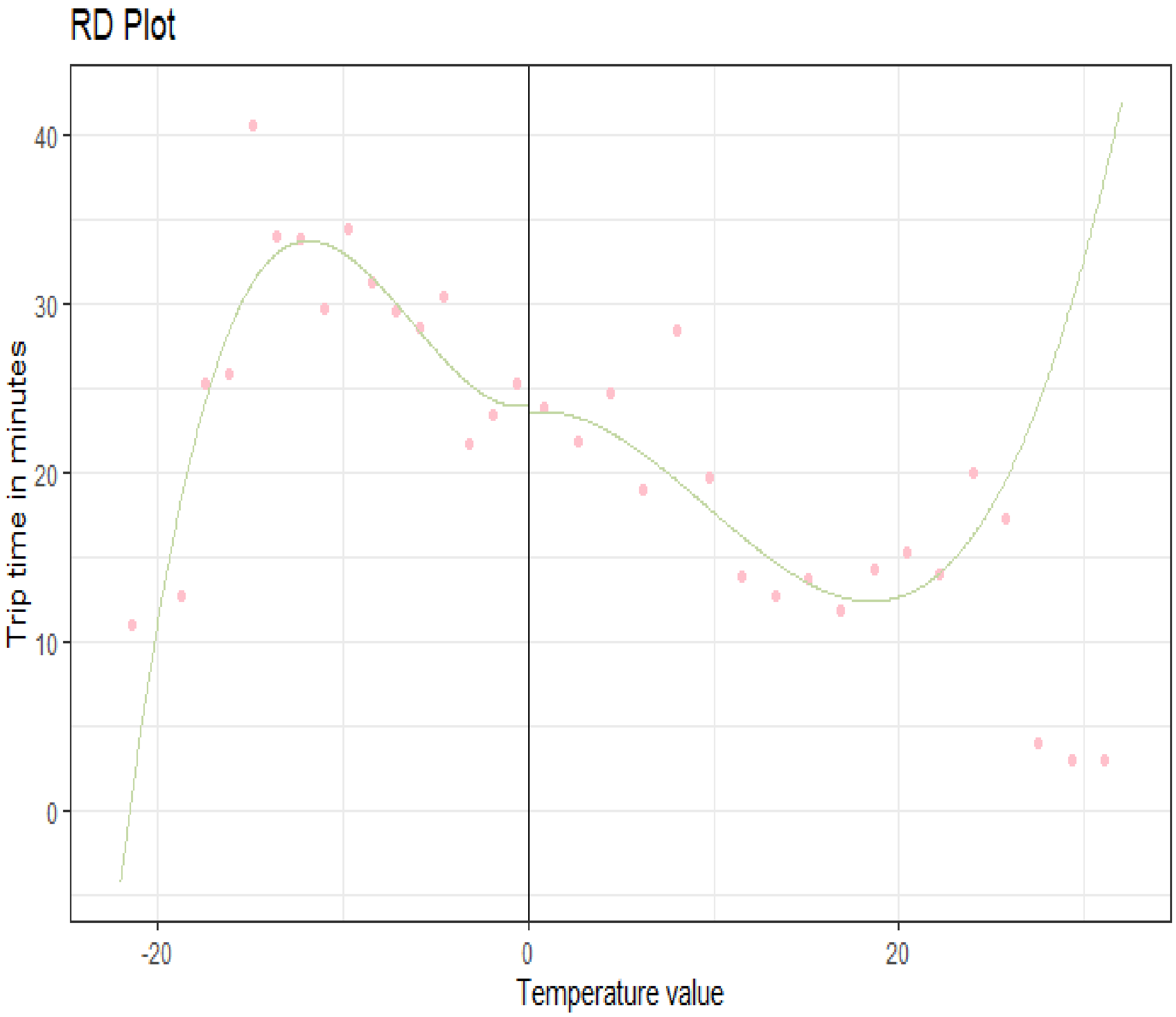


Figure 12: Manipulation in the Exogenous Variable

Figure 12. on the other hand, shows the RD plot for the Trip time in minutes. Although there is a slight jump around the cutoff point, this jump does not prove to be significant. The results from the graph are consistent with the covariate balance table. This is indicative that the Trip Time variable does not have a significant effect on Uber's surge pricing. Therefore, leaving it outside of the model would not lead to any biases in the regression discontinuity model.

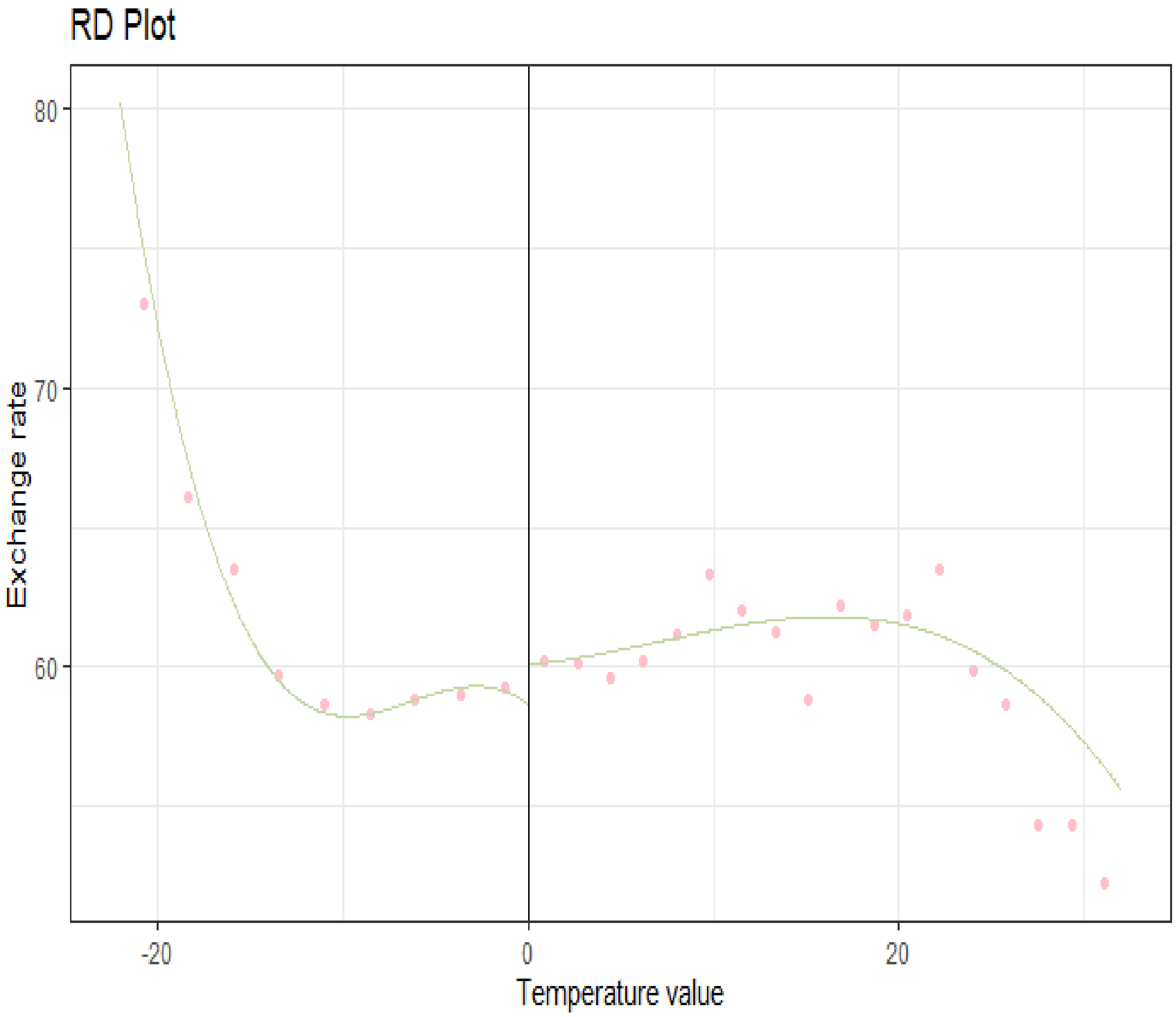


Figure 13: Manipulation in the Exogenous Variable

Figure 13. is the Exchnage Rate's RD plot against the temperature value. Here we can observe a slightly more significant jump in the covariate below and above the zero degrees temperature value. It is indicative that that the covariate is not balanced, but rather endogenous to the model. The result is consistent with the covariate balance table. Intuitively, this means that it is difficult to argue that the temperature value is the only variable having an effect on Uber's surge pricing below and above the cutoff. Instead, variables such as the Exchange rate also have an effect. Therefore, treating the exchange rate variable as exogenous would lead to biases in the regression discontinuity model.

5 Conclusion

To conclude, Uber's surge pricing can be affected by multiple components. One of the components is the temperature value, depending on whether there is freezing or not. In order to isolate this effect and make sure the only thing affecting Uber's surge pricing is temperature, checking for covariate balance on other variables is crucial for the regression discontinuity model to be valid. Both from the covariate balance and the RD plot we concluded that covariates such as trip time in minutes, trip type, If city Saint Petersburg, distance in kilometers as well as the driver's gender are balaced covariates. On ther other hand, however, we also found out that covariates such as time of the day, precipitation as well as the exchange rate between the Russian Ruble and the US dollar are not balanced. Whether or not this unbalance would jeopardize the results from the regression discontinuity model depends on whether we leave these endogenous variables outside of our model or not. It may be substantial to include those variables in the model itself. Doing so would increase the chance of getting robust results from the regression discontinuity model.

References

[1] Uber. *(2021, July 20). A guide for how to use Uber. Retrieved May 06, 2021, from https://www.uber.com/us/en/ride/how-it-works/*

[2] Kaggle *Your machine learning and data science community. (n.d.). Retrieved April 28, 2021, from https://www.kaggle.com/*

[3] Stan Tyan. *(2020, November 10). Data scientist @YOUSICIAN, Ex @uber and amp; @Wrike. Retrieved April 26, 2021, from https://stantyan.com/*

[4] Helling, B. *(2020, December 03). UberX vs. Uber: What's the difference? Retrieved April 24, 2021, from https://www.ridester.com/uberx-vs-uber/*

[5] Martin, N. *(2019, March 30). Uber charges more if they think you're willing to pay more. Retrieved April 24, 2021, from https://www.forbes.com/sites/nicolemartin1/2019/03/30/uber-charges-more-if-they-think-youre-willing-to-pay-more/?sh=536a77987365*

[6] Pope, K. *(2020, March 03). 5 ways to BEAT Uber surge pricing so you don't overpay for your NEXT RIDE. Retrieved April 24, 2021, from https://www.thepennyhoarder.com/save-money/beat-uber-surge-pricing/*

[7] Cattaneo, M. D., Jansson, M., and Ma, X. *(2018). Manipulation testing based on density discontinuity. The Stata Journal, 18(1), 234-261.*

[8] Cattaneo, M. D., Jansson, M., and Ma, X. *(2019). lpdensity: Local polynomial density estimation and inference. arXiv preprint arXiv:1906.06529.*

[9] Ong, T. *(2017, July 13). Uber is merging with Yandex in Russia a year after exiting China. Retrieved April 24, 2021, from https://www.theverge.com/2017/7/13/15963754/uber-yandex-merger-russia*

[10] Paige Okun, C. *(2019, November 08). Beyond Uber: Your guide to ridesharing apps around the world. Retrieved April 24, 2021, from https://www.cnbc.com/2019/11/08/top-ride-sharing-apps-in-europe-asia-south-america-africa-and-usa.html*

6 Appendix

```
\textbf{Set up chuck}

knitr::opts_chunk$set(echo = TRUE)

library(rddensity)
library(ggplot2)
library(tidyverse)
library(rmarkdown)
library(stargazer)
library(RDHonest)
library(rdrobust)
library(rdd)
library(ggthemes)

New_uber<-read_csv(’https://raw.githubusercontent.com/Fjolle/
Causal-Inference-Project/main/Data/New_uber.csv’)

New_uber <- New_uber %>%
  as_tibble() %>%
  janitor::clean_names() %>%
  glimpse()


 \textbf{ RD Density Plot}

New_uber$cutoff=ifelse(New_uber$temperature_value<=0,1,0)

rdplotdensity(rddensity(New_uber$temperature_value,
c = 0), New_uber$temperature_value)

\textbf{ Kernel Density Plot}

d=density(New_uber$temperature_value)
plot(d, main="Kernel Density")
polygon(d, col="#C3D7A4", border="blue")


\textbf{Covariate Balance}

New_uber2 <- New_uber %>%
  mutate(temperature_value_2 = temperature_value - 0)
New_uber2 <- New_uber2 %>%
  filter(temperature_value_2 >=-25 & temperature_value_2 <= 25)

Time_of_day = lm(time_of_day_disc ˜ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Time_of_day, type = "text")

Trip_time = lm(trip_time_minutes ˜ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Trip_time, type = "text")

Trip_type= lm(trip_type_dummy ˜ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Trip_type, type = "text")

City= lm(if_city_sp ˜ cutoff*temperature_value_2,
data=New_uber2)
stargazer(City, type = "text")
```

```
Distance= lm(distance_kms ~ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Distance, type = "text")

Exchange_Rate=
lm(rub_usd_exchange_rate ~ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Exchange_Rate, type = "text")

Precip= lm(precipitation_dummy ~ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Exchange_Rate, type = "text")

Gender= lm(driver_gender_dummy ~ cutoff*temperature_value_2,
data=New_uber2)
stargazer(Exchange_Rate, type = "text")

stargazer(Time_of_day,Trip_time,Trip_type,Precip,
type = "text")

stargazer(City,Distance,Exchange_Rate,Gender,
type = "text")

 \textbf{RD plot for Time of day disc}
rdplot(New_uber$time_of_day_disc,
New_uber$temperature_value, p=3, c=10,
y.label='Time of the day',
x.label='Temperature value',
kernel="triangular",
ci=NULL,col.dots = 'pink',
col.lines = "#C3D7A4",
masspoints="adjust", x.lim=NULL,
y.lim=NULL)
```

```
\textbf{ RD plot for Trip time dummy}
```

```
rdplot(New_uber$trip_type_dummy,
New_uber$temperature_value, p=3,
c=0, y.label='Trip type',
x.label='Temperature value',
kernel="triangular",
ci=NULL,col.dots = 'pink',
col.lines = '#C3D7A4' ,
masspoints="adjust", x.lim=NULL,
y.lim=(NULL))
```

```
\textbf{RD plot for Driver's gender dummy}
```

```
rdplot(New_uber$driver_gender_dummy,
New_uber$temperature_value,p=3, c=0,
y.label='Gender', x.label='Temperature value',
kernel="triangular" , ci=NULL,col.dots = 'pink',
col.lines = '#C3D7A4',masspoints="adjust",
x.lim=NULL, y.lim=NULL)
```

```
\textbf{RD plot for Precipitation dummy}
```

```
rdplot(New_uber$precipitation_dummy,
New_uber$temperature_value, p=3, c=0,
y.label=’Precipitation’, x.label=’Temperature value’,
kernel="triangular", ci=NULL,col.dots = ’pink’,
col.lines = ’#C3D7A4’,masspoints="adjust",
x.lim=NULL , y.lim=NULL)
```

```
\textbf{RD plot for If city SP}
rdplot(New_uber$if_city_sp,
New_uber$temperature_value,
p=3, c=0, y.label=’City’, x.label=’Temperature value’,
kernel="triangular",
ci=NULL,col.dots = ’pink’, col.lines = ’#C3D7A4’,
masspoints="adjust", x.lim=NULL, y.lim=NULL )
```

```
\textbf{RD plot for Distance kms}

rdplot(New_uber$distance_kms,
New_uber$temperature_value,
p=3, c=0, y.label=’Dinstance in km’,
x.label=’Temperature value’, kernel="triangular",
ci=NULL,col.dots = ’pink’, col.lines = ’#C3D7A4’,
masspoints="adjust" , x.lim=NULL, y.lim=NULL)
```

```
\textbf{RD plot for Trip time minutes}

rdplot(New_uber$trip_time_minutes,
New_uber$temperature_value,
p=3, c=0, y.label=’Trip time in minutes’,
x.label=’Temperature value’, kernel="triangular",
ci=NULL,col.dots = ’pink’, col.lines = ’#C3D7A4’ ,
masspoints="adjust", x.lim=NULL, y.lim=NULL)


\textbf{RD plot Rub USD exchange rate}

rdplot(New_uber$rub_usd_exchange_rate,
New_uber$temperature_value, p=3, c=0,
y.label=’Exchange rate’, x.label=’Temperature value’,
kernel="triangular", ci=NULL,
col.dots = ’pink’,
col.lines = ’#C3D7A4’ ,masspoints="adjust",
x.lim=NULL, y.lim=NULL)


\textbf{ci 1}

library(rdd)
RDestimate(price_per_kms_usd ~ temperature_value,
cutpoint = 0, bw = 20, data=New_uber)
RDHonest(price_per_kms_usd ~ temperature_value,
cutoff = 0, h=20, kern="triangular", M=0.1, sclass="T",
order = 1, data=New_uber)


\textbf{ ci 2}
RDestimate(price_per_kms_usd ~ temperature_value |
temperature_value*cutoff, cutpoint = 0, bw = 20,
data=New_uber)
RDHonest(price_per_kms_usd ~ temperature_value +
temperature_value*cutoff, cutoff = 0, h=20,
kern="triangular", M=0.1, sclass="T", order = 1,
data=New_uber)


\textbf{ci 3}

RDestimate(price_per_kms_usd ~ temperature_value |
temperature_value*cutoff + (temperature_value^2)*cutoff,
cutpoint = 0, bw = 20, data=New_uber)
RDHonest(price_per_kms_usd ~ temperature_value +
temperature_value*cutoff + (temperature_value^2)*cutoff,
cutoff = 0, h=20, kern="triangular", M=0.1, sclass="T",
order = 1, data=New_uber)


\textbf{ci 4}

RDestimate(price_per_kms_usd ~ temperature_value,
cutpoint = 0, bw = 40, data=New_uber)
RDHonest(price_per_kms_usd ~ temperature_value,
cutoff = 0, h=40, kern="triangular", M=0.1,
sclass="T", order = 1, data=New_uber)


\textbf{ci 5}

RDestimate(price_per_kms_usd ~ temperature_value |
temperature_value*cutoff, cutpoint = 0, bw = 40,
data=New_uber)
RDHonest(price_per_kms_usd ~ temperature_value +
temperature_value*cutoff, cutoff = 0, h=40,
kern="triangular", M=0.1, sclass="T", order = 1,
data=New_uber)
```

```
\textbf{ci 6}
```

```
RDestimate(price_per_kms_usd ~ temperature_value |
temperature_value*cutoff + (temperature_value^2)*cutoff,
cutpoint = 0, bw = 40, data=New_uber)
RDHonest(price_per_kms_usd ~ temperature_value +
temperature_value*cutoff + (temperature_value^2)*cutoff,
cutoff = 0, h=40, kern="triangular", M=0.1, sclass="T",
order = 1, data=New_uber)
```