[Features and model description 2](#_Toc9515732)

[Introduction & Analysis 3](#_Toc9515733)

[1.1 Features Initial Analysis and Grouping: 4](#_Toc9515734)

[First Model - City Features 5](#_Toc9515735)

[2.1 Preliminary analysis: 5](#_Toc9515736)

[2.2 Regression Model: 6](#_Toc9515737)

[2.3 Model validation 9](#_Toc9515738)

[Final Model – Crime Added 10](#_Toc9515739)

[3.1 Preliminary analysis: 10](#_Toc9515740)

[3.2 Regression Model: 12](#_Toc9515741)

[3.3 Model validation 14](#_Toc9515742)

[3.4 Model Improvement 15](#_Toc9515743)

[Interpretation 16](#_Toc9515744)

[4.1 Predictor interpretation 16](#_Toc9515745)

[4.2 Correlation interpretation 19](#_Toc9515746)

[Conclusion 21](#_Toc9515747)

MSIT 423

Project One

*Ray Liu,* *Jessica Qin, YangHong, Yifan Chen, Yunzi Zhang*

# Features and model description

We picked some of original variables considering their casual relationships, and finally get the fowling model based on the linear assumption with an adjusted R square of 79.24%.

|  |  |  |
| --- | --- | --- |
| **Predictors** | **Definition [(Former explanation)](#_4.1_Predictor_interpretation)** | **Transformation** |
| ASSAULT | A crime activity of threat, excluding bodily harm |  |
| BATTERY | A crime activity of battle | log |
| NARCOTICS | A crime activity of drug | Square root |
| DECEPTIVE\_PRACTICE | A crime activity of giving an appearance or impression different from the true one; misleading. | log |
| BURGLARY | A crime activity of unlawful entry into a building for the purpose of committing an offence |  |
| HOMICIDE | A crim activity of skill themselves | 5\*Log(x+2) |
| THEFT | Stealing, it increases the demand to our surprise. [(explanation)](#_4.1_Predictor_interpretation) | log |
| avgbf | The composite predictors by average the business licenses and food establishment. It is an indication about CBD | log |
| ROBBERY | A crime activity of taking property unlawfully from a person or place by force or threat of force. | log |
| CAPACITY | The maximum carrying amount |  |
| MINORITY | The group type of people in certain location | square |
| CBD | Dumpy predictors, where 1 is CBD, and 0 is not. |  |

# Introduction & Analysis

The goal of this project is to build a predictive model to predict bike demands based on major crime types and city features information. 8 major crime types, assault, battery, burglary, criminal trespass, deceptive practice, homicide, narcotics, robbery, theft, might affect demands of bikes. Increasing the number of crimes that cause loss of personal possession and direct harm of body could lead the decreasing demands of bikes because people will feel unsafe riding a bike on the road. On the other hand, crimes that happen less frequently and cause major social affect such as homicide will have less effect on the demand of bikes due to the fact that there is nothing people can do to prevent it from happening.

City’s city features information can also have profound effect on the demands of bikes. High number of bike routes means there are more demands on bikes. High number of retails stores and restaurant indicate this area is prosperous, therefore having more people, which leads to high demand of bikes.

## 1.1 Features Initial Analysis and Grouping:

The bike data has 45 predictor variables and 1 response variables. According to the business objective, we only select 8 crime types and city features information. Our initial thought is to composite 8 crimes as one variable since they are highly correlated. Also, we want to group some city features variables that are highly related. However, after grouping, R square has decreased. Also, we want to analyze bike demands based on each type of crime, so we decide not to include composite variables in our model.

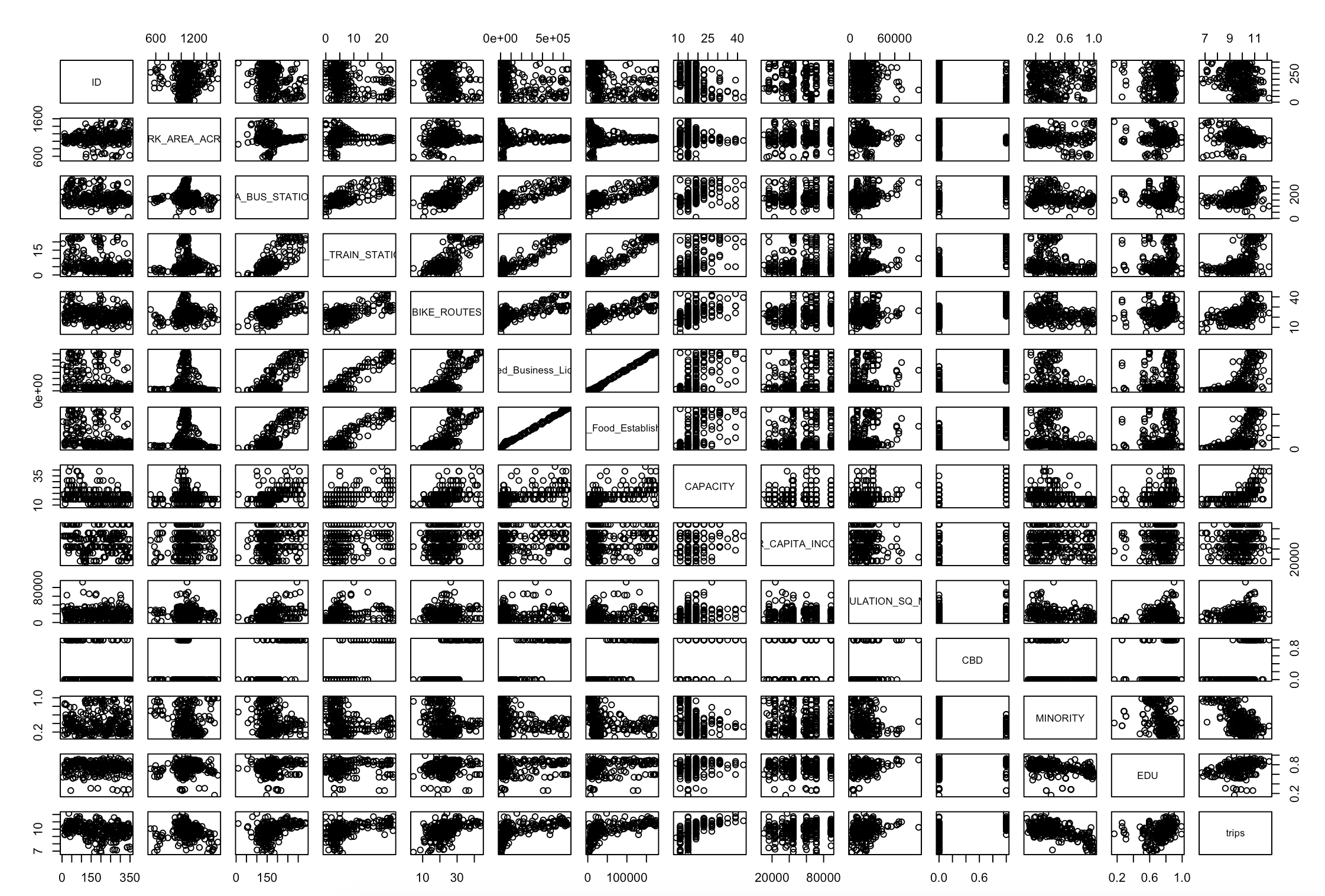
Our model analysis will separate into two parts. First part is to analyze city features variables and build the first model. Second part is to analyze crime variables and build the second model. Finally, we combine the two models together to get final model. The correlations between city features and crime are not strong, so the coefficient of our final model will not be affected.

# First Model - City Features

## 2.1 Preliminary analysis:

2.1.1 Omitted unrelated predictors and correlation analysis

We diagnose data by generating correlation matrix and scatter plots. First, we observe that some predictors are to be removed since they are not correlated with demands (trips) as well as not cause of dependent variables.



PIC(3.1)

According to the scatterplot (3.1), predictors to be omitted are:

ID, PARK\_AREA\_ACRES, PER\_CAPITA\_INCOME

2.1.2 Correlation analysis

According to the correlation matrix, the multicollinearity exists.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CTA\_TRAIN\_STATIONS | Limited\_Business License | Retail\_Food\_Establishment |
| Limited\_Business\_License | 0.9400 | 1 | 0.997 |
| CTA\_BUS\_STATIONS | 0.76 | 0.79 | 0.80 |
| Retail\_Food\_Establishment | 0.9396 | 0.997 | 1 |
| CBD | 0.85 | 0.882 | 0.882 |

If we were to use these correlated predictors in the model, we can end up with wrong conclusions from the model, e.g., concluding a variable is not significant.

We found that the correlation between business licenses and food establishments is so high that we average these two predictors as a composite predictor named avgbf.

## 2.2 Regression Model:

2.2.1 Predictor selection

According to the correlation matrix and the scatterplot, there is a strong multicollinearity problem. So we use Lasso to select the predictors. As the output in Pic(3.2.3), after considering the correlation, we decide follow predictors into our first model.

CTA\_TRAIN\_STATIONS, CTA\_TRAIN\_STATIONS, BIKE\_ROUTES, POPULATION\_SQ\_MILE, CBD, CAPACITY, MINORITY, EDU , avgbf

Pic(3.2.1)

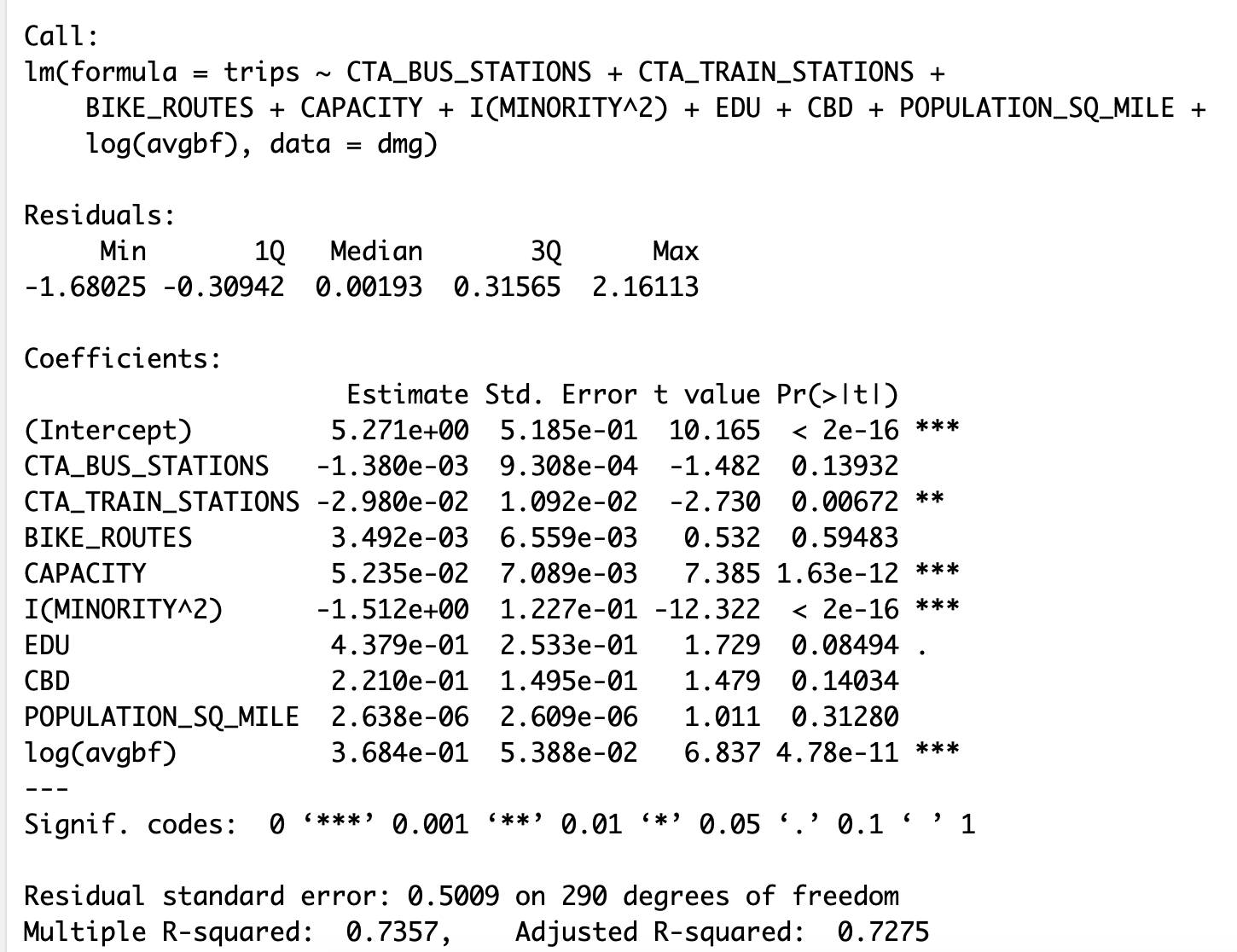
2.2.2 Decide independent variable transformation

While the predictors should be transformed may be, the bold words are our final choice:

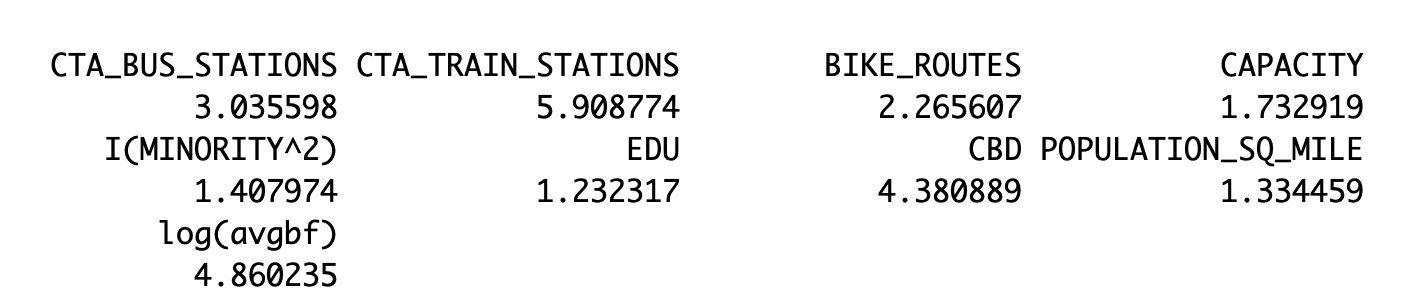
|  |  |  |
| --- | --- | --- |
| Predictors | Transform assumption1 | Transform assumption2 |
| BIKE\_ROUTES | N/A | square |
| CTA\_TRAIN\_STATIONS |
| CTA\_BUS\_STATIONS |
| CAPACITY |
| POPULATION\_SQ\_MILE |
| MINORITY | N/A | **Square** |
| EDU | N/A | N/A |
| Avgbf | **log** | Square root |
| CBD | Category | |

According to the summary of our linear model, the p-value of BIKE\_ROUTES, EDU, and CBD are 0.285, 0.078, 0.443 respectively, indicating those three predictors are not significant. After Drop those three predictors, we get our first model(with out crimes) below, with an adjusted R2 of 72.75%:

With a numerical summary:



After calculating the VIF, there is no significant multicollinearity in this model.

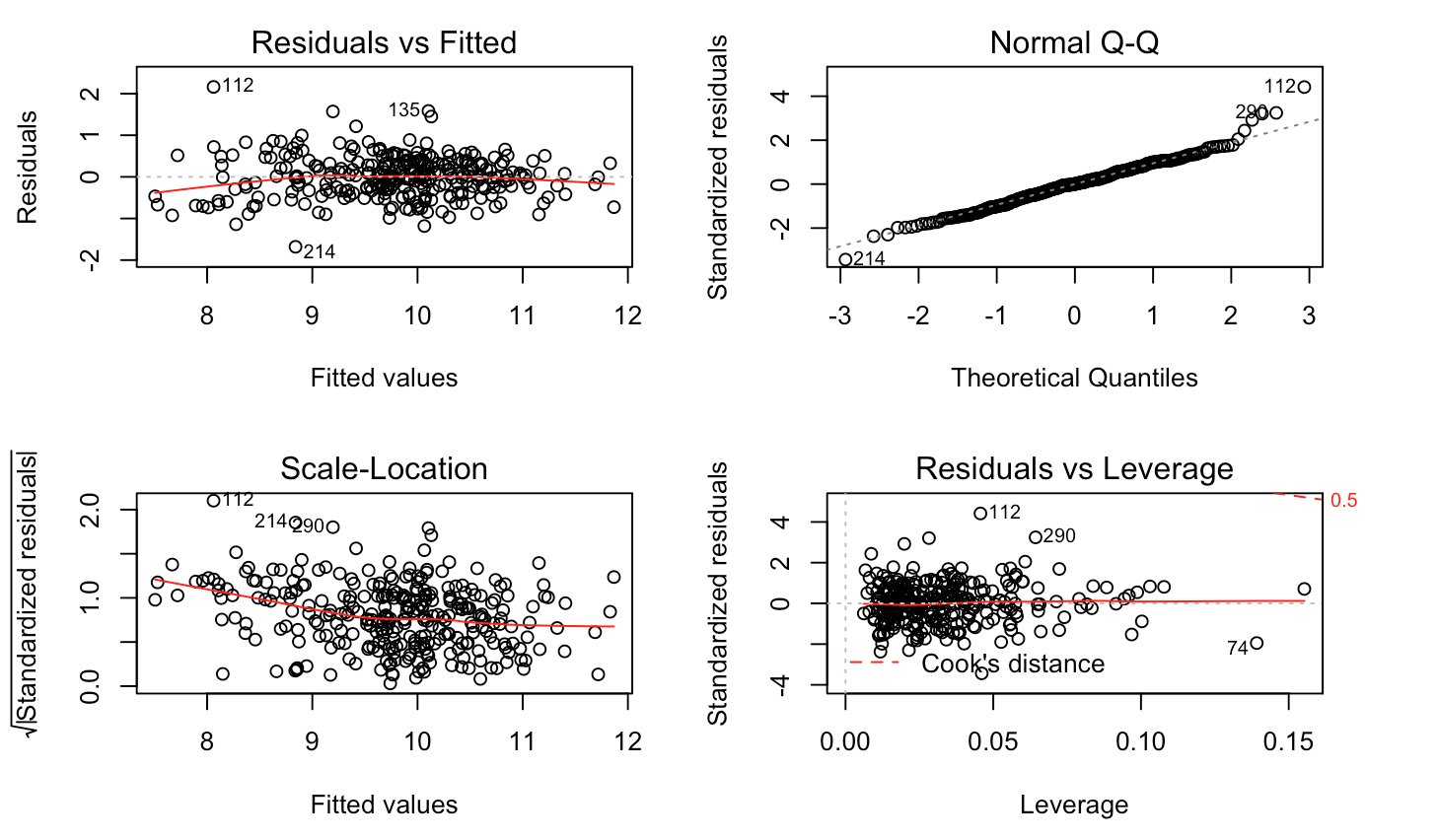


## 2.3 Model validation

In this section, we validate the linear assumptions of the regression model after we apply the square transformation. Pic(3.3) shows the diagnostic plots of the model. From left to right, up and down, the plot are for linearity, constant variance, uncorrelated errors, and error normality.

The residual plot does not indicate a particular pattern, indicating the model after partly square transformation is linear. Meanwhile, the residuals are like snowstorm, showing a pretty good match. The Q-Q plot shows the normality of errors, and there is no outlier with a Cook’s Distance more than 0.5.

According to present diagnostic, this model obeys the assumption and has no outliers.

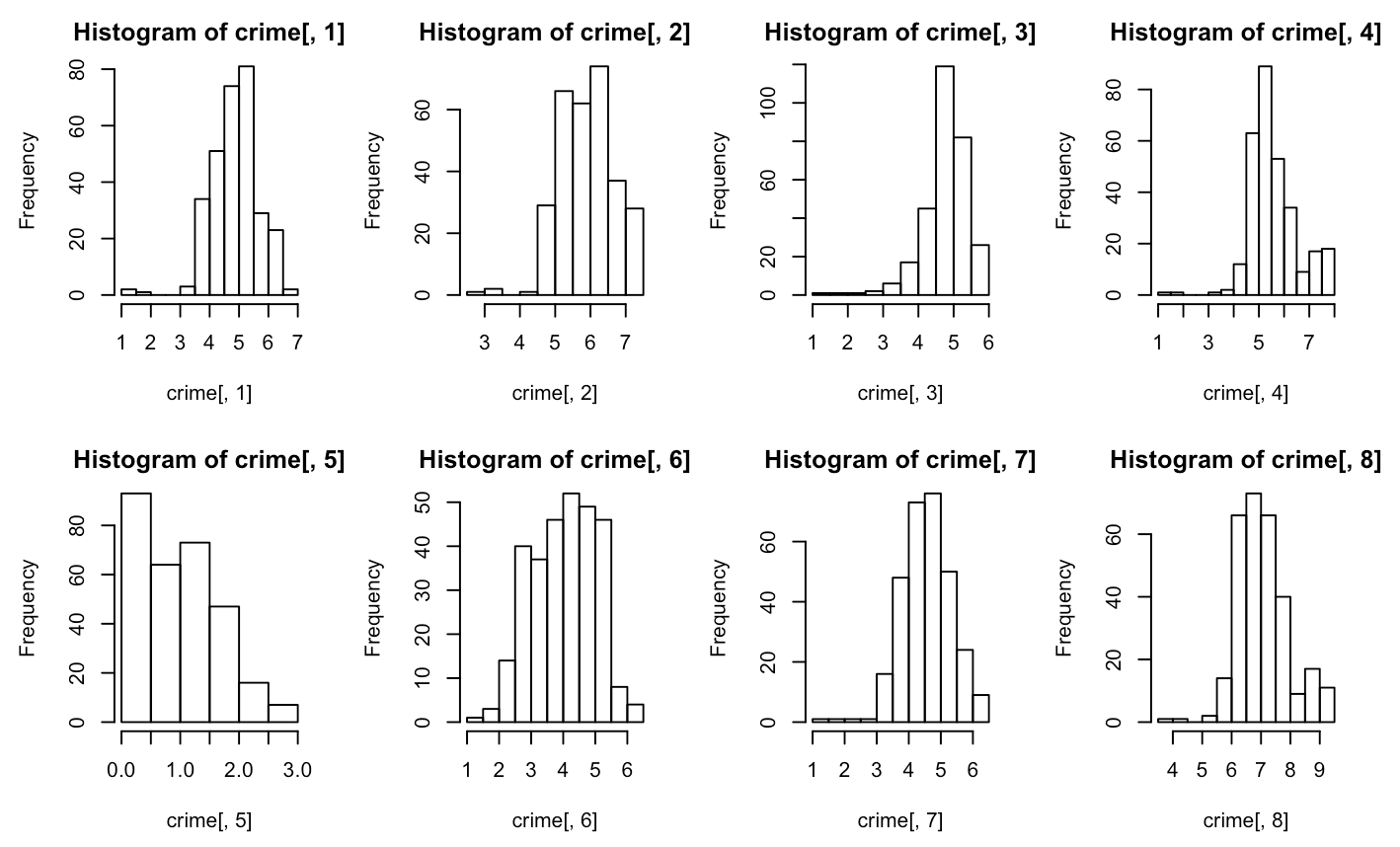


# Final Model – Crime Added

## 3.1 Preliminary analysis:

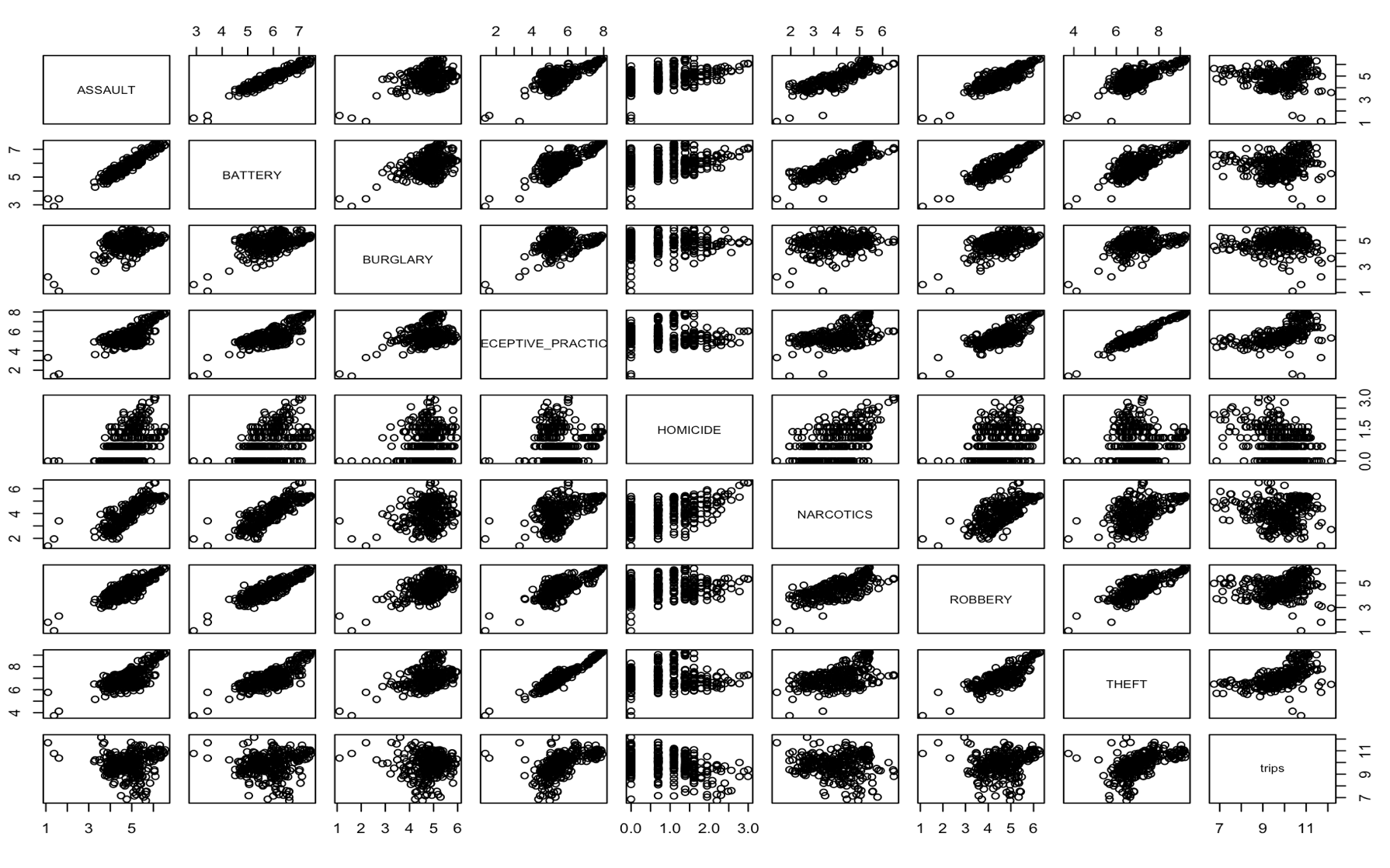
3.1.1 Omitted unrelated predictors

First, we view the histogram of the predictors to check if it is normally distributed.



Second, we perform preliminary analysis by looking at the correlations between each crime variable and the bike demand ('trips'). The crime variables with moderate correlations are selected.

Picture below is the scatterplot of the crime predictor with trips. Every crime predictor seems to correlate to trips, so we cannot delete any of the crime predictors only depending on scatterplot.



3.1.2 Correlation analysis

When consider the correlation relationships, we take all the eight crimes into consideration, meanwhile keep the demographic predictors selected in Section two in the model. Because there maybe correlation between crime predictors and demographic predictors.

The correlation matrix is too big to be list, so I list the high correlated predictors, excepting the correlation within the demographic predictors, which I has listed in Section 2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | CTA\_BUS\_STATIONS | CTA\_TRAIN\_STATIONS | ASSAULT | BATTERY | DECEPTIVE\_PRACTICE | THEFT | avgbf  (CBD) |
| ASSAULT | 0.7423 |  | 1 | 0.95 |  |  |  |
| BATTERY | 0.7480 |  | 0.95 | 1 |  |  |  |
| DECEPTIVE\_PRACTICE | 0.8422 | 0.7374 |  | 0.7726 | 1 |  | 0.7606 |
| THEFT | 0.8286 | 0.7122 |  | 0.7724 | 0.9496 | 1 | 0.7454 |
| NARCOTICS |  |  | 0.825 | 0.8387 |  |  |  |
| ROBBERY |  |  | 0.8524 | 0.8877 | 0.7789 | 0.81 |  |

If we were to use these correlated predictors in the model, we can end up with wrong conclusions from the model, e.g., concluding a variable is not significant.

According to the VIF analysis, there is multicollinearity between theft, battery, deceptive, and assault, making battery insignificant. Meanwhile, there is also multicollinearity between crimes and demographic predictors, making the THEFT has the opposite effect on the trips to the BURGLARY.

The multicollinearity comes from the behavior of these crimes, we will explain all the multicollinearity in Section 5.

## 3.2 Regression Model:

3.2.1 Predictors exploration and transformation

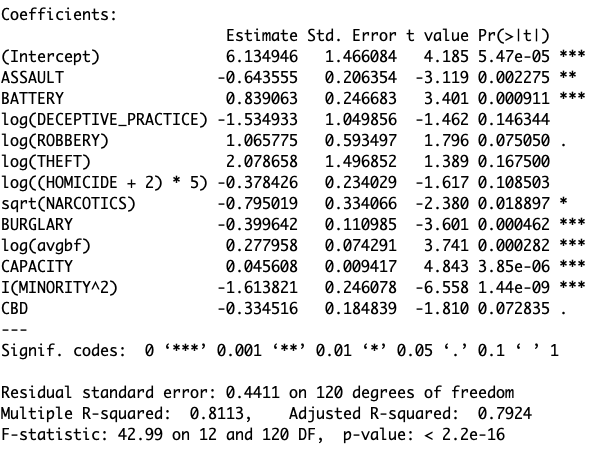
According to the scatter plot and the histogram graphic, we decide a transformation below, the bold words are our final transformation choice:

|  |  |  |
| --- | --- | --- |
| Predictors | Transform assumption1 | Transform assumption2 |
| ASSAULT | **N/A** | square |
| BATTERY |
| BURGLARY |
| THEFT | N/A | **Log(x+a)** |
| ROBBERY | N/A | **Log(x+a)** |
| DECEPTIVE\_PRACTICE | N/A | **Log(x+a)** |
| HOMICIDE | **Log[(x+a)\*b]** | Square root |
| NARCOTICS | N/A | **Square root** |

3.2.2 Predictor Selection

As the eight crimes are the predictors, we tried three predictor selection algorithms, the stepwise, the Ridge, and the Lasso. These three predictors have the MSE from test sets 0.2378, 0.2723, 0.0.2718 respectively. We choose the stepwise selection as our final model, because it does best in this case. The output has been added in the appendix.

Here is the final model after selection by stepwise:

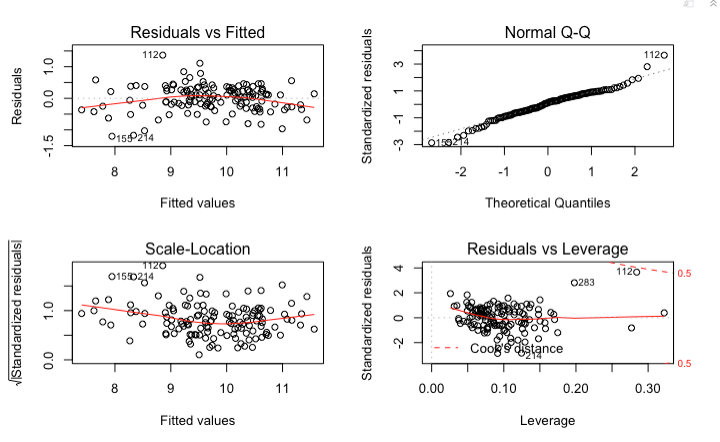


Finally we get the model:

## 3.3 Model validation

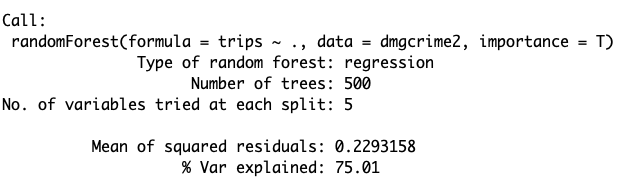
In this section, we validate the linear assumptions of the regression model after we apply the square transformation. shows the diagnostic plots of the model. From left to right, up and down, the plot are for linearity, constant variance, uncorrelated errors, and error normality.

The residual plot does not indicate a particular pattern, indicating the model after partly square transformation is linear. Meanwhile, the residuals are like snowstorm, showing a pretty good match. The Q-Q plot shows the normality of errors, and there is no outlier with a Cook’s Distance more than 0.5. For model building, no transfer needed anymore.



## 3.4 Model Improvement

We tried the random forest this time, and it only has R square of 75.01%. In this case, the linear regression does a better job.



# Interpretation

## 4.1 Predictor interpretation

***avgbf:*** It is the predictor composited by average the business licenses and food establishment. We composite these two predictors because when we do the correlation analysis, we found they have a high correlation with each other. We composite because:

1. There is no numerical predictor that can represent the CBD levels;
2. The scale and the number of food establishment and business license are close to each other, so we can average it correctly without standardized and weighted.
3. These two predictors are all related to CBD position.

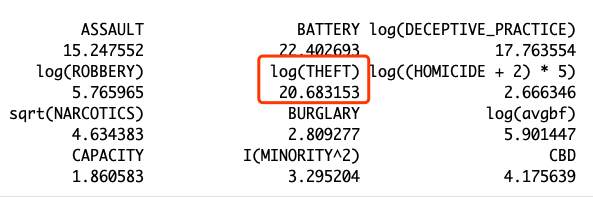
***CBD :*** It is different from avgbf because it is a dumpy predictor, where 1 is CBD, and 0 is not. We keep this predictor because (1) we would like to know if there will be any interaction between CBD and other numerical predictors, and (2) the stepwise indicates it is worthwhile to add it into the model.

***BURGLARY:*** A crime activity of unlawful entry into a building for the purpose of committing an offence, this kind of activity threat people to come to certain place, hence decrease the demand of the public bike.

1. There is multicollinearity between THEFT with any other predictors.
2. Is there any interaction effect on this model?

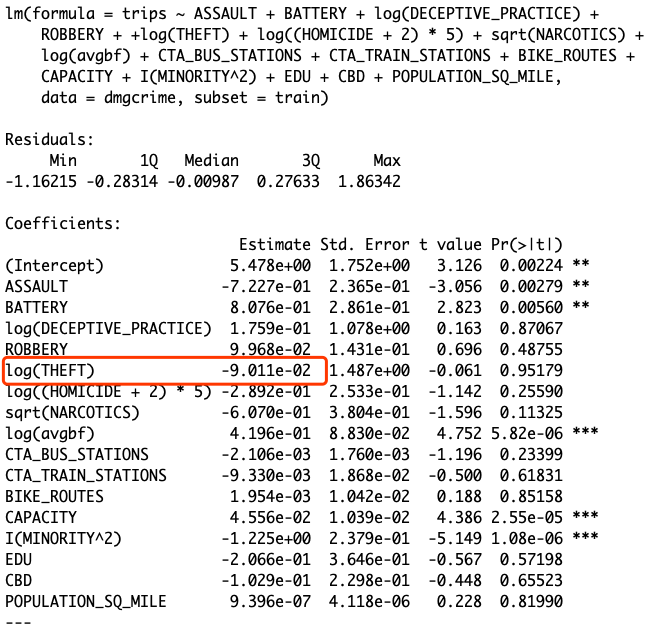
To find the answer, we do the following steps:

1. Check the VIF to see if there is multicollinearity, Here is the output:



We find that there is a severe multicollinearity between THEFT and other variables. Then we begin to omit predictors that have a high correlation with THEFT to find the answer.

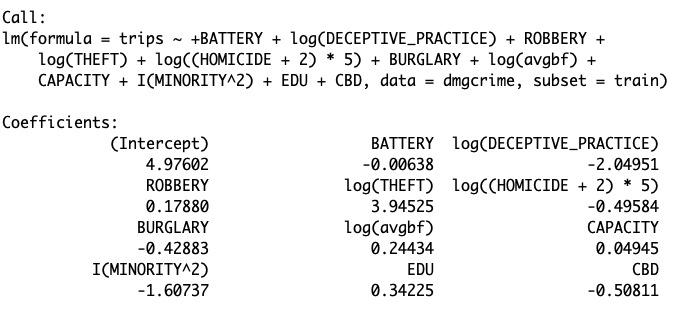
1. We find that if we delete the BURGLARY from the model, the coefficient of THEFT become negative. Now we know that is the multicollinearity cause the abnormal coefficient of THEFT.



1. Is there any interaction effect? In order to answer this question, we tried the interaction plot in R. There is no significant evidence that there is an interaction effect, so we do not add the plot here, but the plot and code can be checked in appendix.

***ASSAULT& &*** ***BATTERY***: These two predictors are all the indicates of happening of crimes. The crimes will decrease the demand of the trips because of the safety. However, the coefficient of BATTERY seems opposite. From the VIF, we also notice that the VIF of ASSAULT and BATTERY are both more than 10.

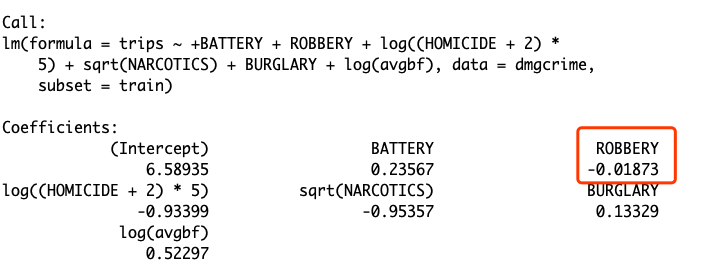
We deleted the ASSAULT and NARCOTICS from the model, the coefficient of BATTERY become negative. So we know that multicollinearity has a misleading again. But the model within ASSAULT seems to explain more, so in the final model we keep the ASSAULT.



***NARCOTICS:*** This is a crime activity of drugs addiction, we square root it to make it into a normal distributed data, and the trace of the plot seems to be square root. This crime activity always related to other crime activities. The correlation of NARCOTICS with other crime predictors are all above 50%. NARCOTICS decrease the demand of the trips because the drugs addiction always threaten people to come to that place, or people prefer to drive a car if they have to go to that place.

***HOMOCIDE:*** A crim activity of skill themselves. We transfer it in the way of 5\*log(x+2) because we find that the distribution of HOMOCIDE is obviously left skewed, and some values of HOMOCIDE are 0, while the number is not big enough after log by 10. HOMICDE decrease the demand because it is amplified by social media, making people prefer to stay away from that place.

***ROBBERY:*** A crime activity of taking property unlawfully from a person or place by force or threat of force. To our surprise, it also increase the demand of the trips. Once again, we deleted the high correlated predictors, finally find the abnormal coefficient also comes from multicollinearity.



***MINORITY***: The further information is needed to analyze these two predictors. Certain groups of people may less likely to bike.

***CAPACITY***: A higher capacity may attract more customers, thought the capacity influence the trips only in a slight level.

***DECEPTIVE\_PRACTICE:*** A crime activity of giving an appearance or impression different from the true one; misleading. We log it because of the abnormal distribution and the trace of the plot. This crimes decrease the demand of the trips in a higher effect than we expected.

## 4.2 Correlation interpretation

***CTA\_TRAIN\_STATION, CTA\_BUS\_STATION:*** In normal life, the train stations always near to the bus stations for transportation convenient.

***Retail\_Food\_Establishment, CTA\_TRAIN\_STATION, LIMITED LICENSE:*** The big volume of pedestrian produced by retail food establishment contributes to the station setting. Meanwhile, the retailers always related to business license.

***CBD, Retail\_Food\_Establishment, CTA\_TRAIN\_STATION, LIMITED LICENSE:*** The central of the business district always have more train stations and retailers than other places. Business licenses also different in active business places.

***ASSAULT, BATTWERY:*** As defined, assault involves a threat, but not bodily harm, while battery implies harm; however, the harm usually comes along with threat, making a high correlation between assault and battery.(0.95 correlation between battery and assault.

***NARCOTICS,ASSAULT,BATTERY:*** The correlation between Narcotics and Assault, Battery are both bigger than 0.85. Narcotics can be a part of reasons of Assault and Battery. Combine the analysis in correlation of Assault and Battery, Narcotics has the high probability to be the “w” cause both Assault and Battery, making a high correlation between them.

***THEFT,ASSAULT,ROBBERY, DECEPTIVE:*** relation between Theft and Assault, Robbery, Deceptive are all bigger than 0.8. Assumption 1: Income is the causality of these four crimes, however, cor(bike$PER\_CAPITA\_INCOME, bike$THEFT) is only 0.12. Assumption 2: Assault, Robbery, and Deceptive always come after the Theft.

# Conclusion

From our model, ASSAULT, BURGLARY, THEFT, CAPACITY, and MINORITY can explain 75% percent of the demands, however, the further information is needed to improve the demand like the MINORITY. What kind of groups people less likely to use the bike? Why they do not like the bike? What can we improve?

There are some places we have to improve:

1. There still be 20% of data cannot be explained by our model.
2. Multicollinearity cause a big problem to interpretation.