The Related Factors to Marijuana Crime Rate in Denver

Hongyang Zheng

Menghan Lin

Shanyu Hou

Xinman Wu

University of Colorado at Denver

Abstract

This paper focuses on the process of building a regression model revealing the significant demographic factors associated with marijuana crime rate in Denver Area. The exploratory data, colinear checking, model selection, F-test, and diagnostic steps taken in the model building process are described in details in this paper. There are also interpretations of the coefficients based on the context of the dataset, explaining how this model works in predicting marijuana crime rate in Denver area.

Introduction

In Denver, both medical and retail marijuana are legal to people 21 or older. Since marijuana is a kind of psychoactive drug, the use of marijuana is still forbidden in many parts of the world. The physical and mental effects of marijuana may cause people to commit crimes. Our team is interested in the question that how marijuana crime rate is associated with different demographic features of neighborhoods in Denver.

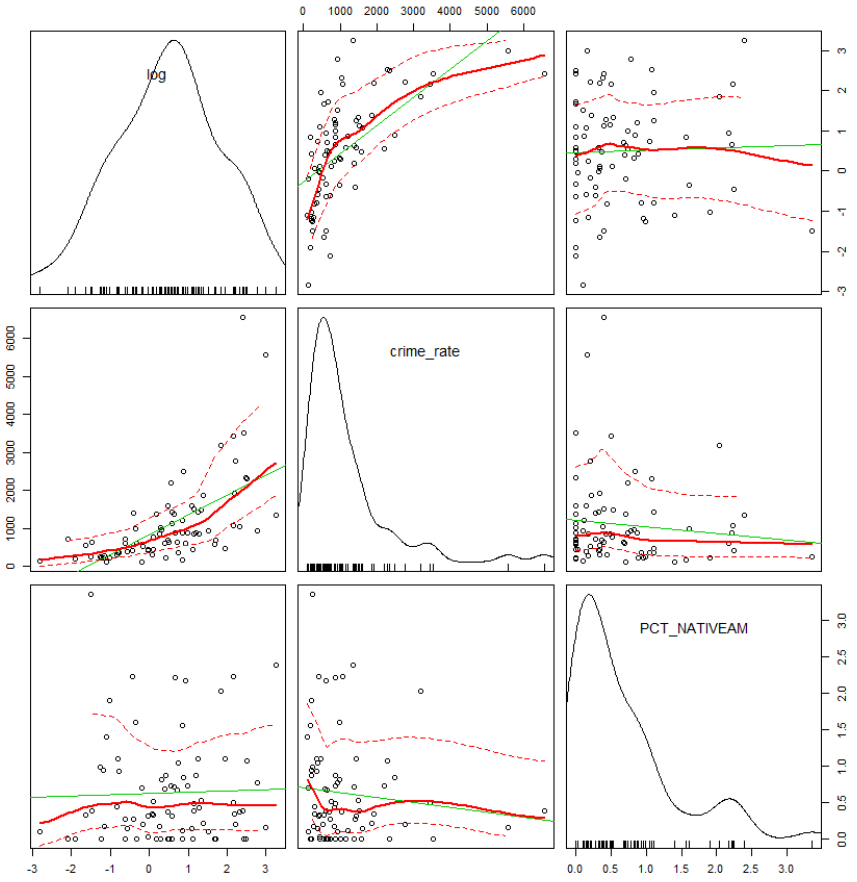
We picked variables from three datasets: *Crime Marijuana, Crime* and *American Community Survey_2010-2014* from the Denver Open Data Catalog. These data are gathered by Denver Police Department and US Census Bureau to measure the crime rate and demographic features. We combined the picked variables into a new dataset. The variables in the new dataset can be divided into two types : One is related to crimes, including marijuana crime rate per capital, crime rate per capital and whether there are police stations in the neighborhood; The other indicates the demographic features in the neighborhood, such as percentage of black people, percentage of male, and percentage of people have a college degree. The more detailed descriptions of the variables are in the appendix.

Methods

Summary data

Since the preferable model uses log(crime_rate) as response, we just show the scatterplots, boxplot and density plots regressing log(crime_rate) on other variables.

We transformed crime_rate in logarithm since the plot (log(m_crime_rate) vs crime_rate) indicates there is a logarithm relationship between them. Except transforming m_crime_rate to log(m_crime_rate), we didn’t do other transformations. Even if some variables have parabolic line (PCT_renter), it would not make sense if we transferred into a squared percentage because our variables are percentage. The density plot shows the variable crime rate is positively skewed because the long tail going in the positive direction.The density plot shows the variable percentage of native American is positively skewed because the long tail going in the positive direction, and it is bimodal.



The density plot shows the variable crime rate is positively skewed because the long tail going in the positive direction; The density plot shows the variable percentage of native American is positively skewed because the long tail going in the positive direction, and it is bimodal.

Collinearity

The combined data contains 26 variables in total. Since the neighborhood name cannot be evaluated as numerical or categorical predictor, we removed it at first. To detect the possible collinear relationship between variables, we built a linear model regressing m_crime_rate on all other variables as predictors. The R^2 is 0.63 while there is only one significant predictor according to the p-values, indicating a potential collinearity problem.

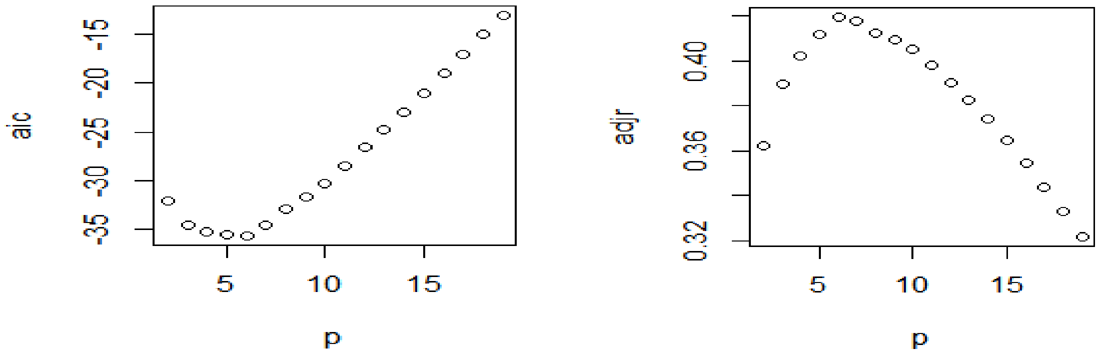
After checking the pairwise correlations, we found that PCT_female and PCT_male, PCT_non_fam and PCT_married have perfect collinearity. Thereafter, we amputated PCT_female and PCT_married as needed.

We rebuilt a linear model with the former response and the rest of predictors. When computing the variance inflation(VIF) for the new model, we found several large VIFs, where the VIF of PCT_hispanic is the largest. In the meantime, the PCT_hispanic also has large correlations with other predictors, such as PCT_White, PCT_LeCollege, and PCT_other fam. When checking the variance decomposition proprotions, PCT_hispanic, PCT_white, and PCT_black have very large variance decomposition proportions for the second largest index. Thus PCT_hispanic is amputated. Repeating the process of building new model, we found that the problems of VIF are significantly improved. However, PCT_Le19 and PCT_age2065 still have extremely large VIFs. The variance decomposition proportions of PCT_Le19, PCT_age2065 are also pretty large at the 4th large index. When exploring the pairwise correlations between predictors, PCT_Le19 has relatively larger correlations than other predictors compared to PCT_age2065.

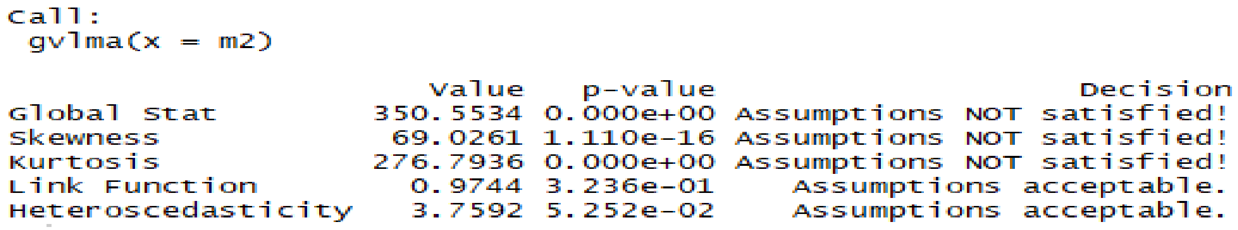
We repeated the former process to detect more possible collinearity. The VIF of the new dataset and the variance decomposition proportions suggest that there is still possible collinearity problem among PCT_age65Plus, PCT_LeCollege and PCT_white. Then, we amputated PCT_white, PCT_age65plus and PCT_LeCollege. Next, we got a dataset with very small VIF for each predictor, pairwise correlations and variance decomposition proportions. Since we were interested in whether PCT_white and PCT_non fam have association with m_crime_rate, we added those two predictors back. Although the VIFs increased after this step, both the VIF and variance decomposition proportions are still in the acceptable range. Therefore, we finally kept 19 variables in the final dataset.

Model Selection

After excluding collinearity, we used the remaining 19 variables to do model selection. Regressing marijuana_crime_rate on the rest of variables, we found that the model (model1) containing pct_male, pct_native_American, pct_Asian, pct_graduate, pct_renter, crime_rate and an intercept is a preferable model, according to the graph of Akaike information criterion(AIC) and adjusted-R^2. In the graph of AIC, the lowest point is ranked as number 6, indicating that including the above 5 variables is a better choice. In addition, the 5th point (number 6) in the right graph has the maximum adjusted-R^2 value, suggesting that model1 has a better model fitness.

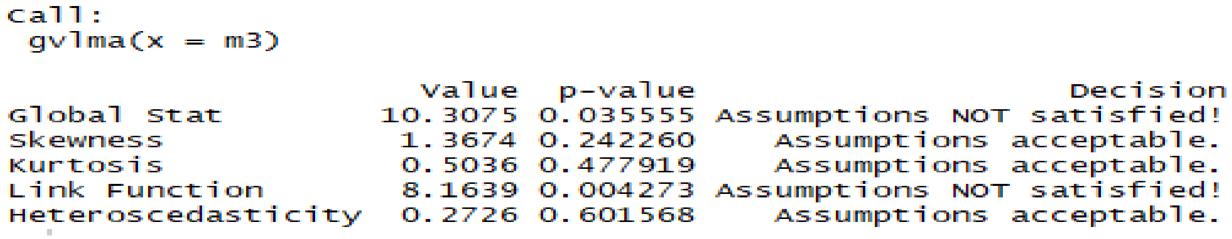


Next step we processed diagnostics for model1. We firstly did a general checking through gvlma test. The results show that model1 has proper structure, because it passes link function. However, although the error has constant variance, it is not a normal distribution(skewness and kurtosis don't hold).



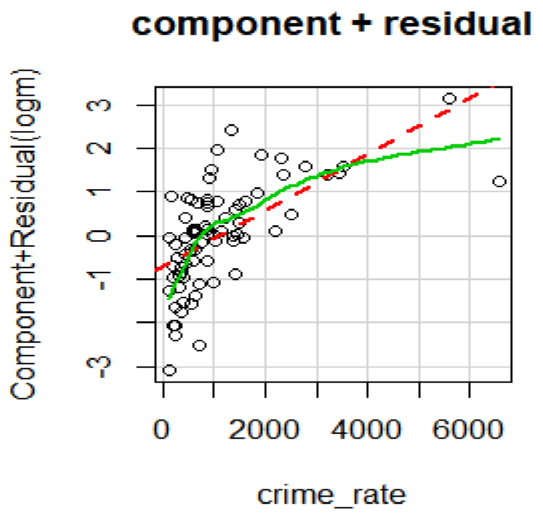
```
Call:
 gvlma(x = m2)

                        Value   p-value                   Decision
Global Stat          350.5534 0.000e+00 Assumptions NOT satisfied!
Skewness              69.0261 1.110e-16 Assumptions NOT satisfied!
Kurtosis             276.7936 0.000e+00 Assumptions NOT satisfied!
Link Function          0.9744 3.236e-01    Assumptions acceptable.
Heteroscedasticity     3.7592 5.252e-02    Assumptions acceptable.
```

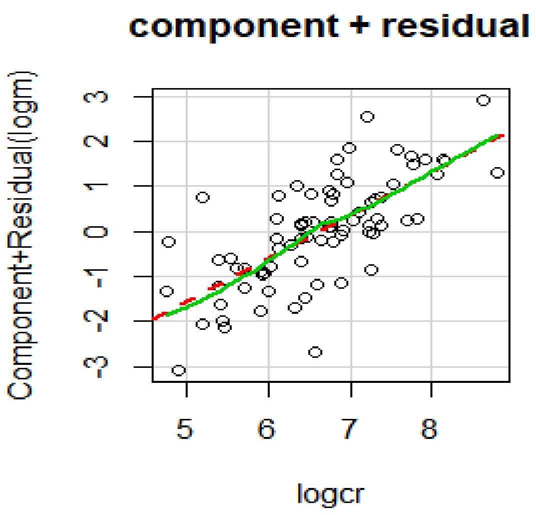
To fix our problem, we tried to transform our response. Instead of using marijuana_crime _rate, we used log(marijuana_crime_rate) as our new response, and did the model selection again. The new model (model2) regresses logm on pct_male, pct_other_family and crime_rate with an intercept. Next, we repeated the general checking to see whether a further transformation is needed. The results show that the problem with error is corrected, but structure assumptions(link function) don't hold.



```
Call:
 gvlma(x = m3)

                      Value  p-value                   Decision
Global Stat         10.3075 0.035555 Assumptions NOT satisfied!
Skewness             1.3674 0.242260    Assumptions acceptable.
Kurtosis             0.5036 0.477919    Assumptions acceptable.
Link Function        8.1639 0.004273 Assumptions NOT satisfied!
Heteroscedasticity   0.2726 0.601568    Assumptions acceptable.
```

To find the structure problem, we repeated step one--data exploration and found that logm and crime_rate has a clear log relationship. Having transferred crime_rate to log(crime_rate) in model2, all assumptions are satisfied. We noticed that there is a considerable improvement in structure of crime_rate after this change. By comparing component plus residual (cr) plots drawn before transformation(left) and after transformation(right), it is clear that the green line is much closer to the red line for the right graph, indicating that the structure problem has been fixed.





In the next step, we replaced crime_rate with log(crime_rate) and ran the model selection using the updated 19 variables. The best model suggested by AIC, Bayesain information Criterion(BIC), Mallows's C_p statistic (CP) and adjusted-R^2 plot is slightly different: model containing 7, 2, 17, and 8 variables is preferable respectively according to each graph and all of them pass the gvlma test. Since the general assumptions are satisfied, we firstly did F-test for these models in order to exclude unpreferable models.

Hypothesis Test

Those four acceptable models are as following: 2 variables model named lmodn2, 7 variables model named lmodn7, 8 variables model named lmodn8 and 17 variables model named lmodn17 (more details are in Appendix).

At first, we compared lmodn2 with other three models since lmodn2 is the simplest model.

F-test	(lmodn2,lmodn7)	(lmodn2,lmodn8)	(lmodn2,lmodn17)	(lmodn7,lmodn8)
p-value	0.02714	0.03294	0.2849	0.2849

From F-test, the p-value comparing lmodn2 with lmodm7 is 0.02714, the p-value comparing lmodn2 with lmodn8 is 0.03294 and the p-value comparing lmodn2 with modn17 is 0.3268. As a result, we got that lmodn7 and lmodn8 are better than lmodn2 but there is no evidence to conclude that lmodn17 is better.

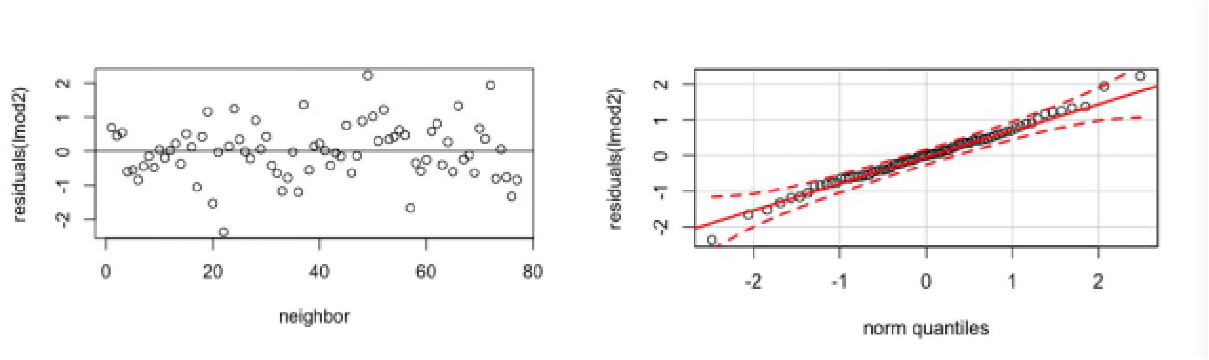
Second, to be more confidence, we compared lmodn7 and lmodn8 and wanted to decide which model is suitable. The lmodn7 is more suitable than lmodn8 in that the p-value is 0.2849. As are result, we chosen lmodn7 as we final model. Moreover, we noticed that the coefficient for pct_male is not statistically significant in lmodn7. Since the hypothesis test shows that removing pct_male doesn’t influence our model fitness much, we finally choose a simpler model (model3 with 6 variables) as our ideal model, which regresses log(m_crime_rate) on pct_white, pct_Asian, pct_other_family, median_earnings, pct_fb and log(crime_rate).

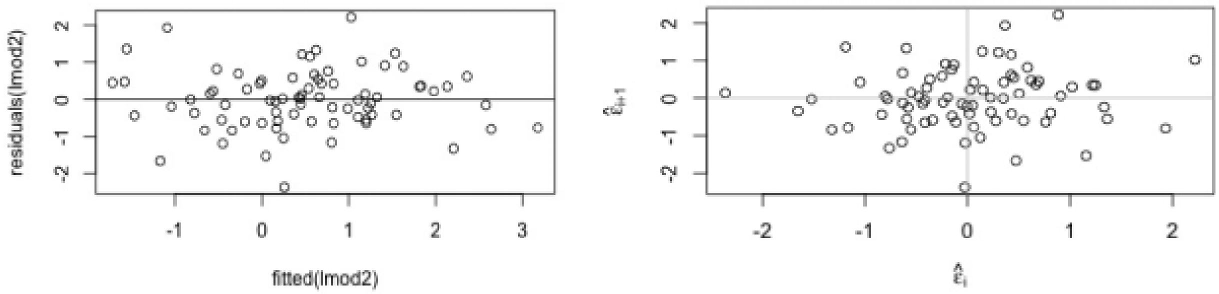
Diagnostics

We do general assumptions checking using gvlma during the above model selection process. Next, we do diagnostics in error assumptions, structural assumptions and influential observations in detail. For error assumptions, there is no test in gvlma to show whether the error terms are independent to each other, so we think it is an essential step to double check these assumptions and find potential problems.

--Error Assumptions

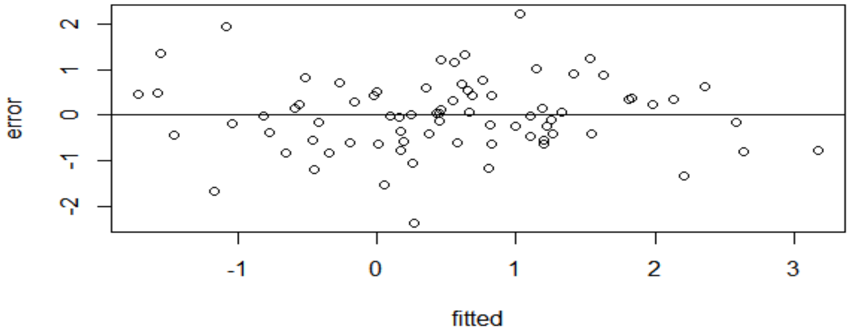
Model3 satisfies all those error assumptions well. The scatter plot of residuals against fitted value shows random pattern and constant thickness when going along horizontal axis, which suggests the constant variance of the residuals. The residuals seem to have some cyclical pattern, when plotted against a vector which labels all the neighborhoods from 1 to 78. However, the plot of successive pairs of residuals doesn't show obvious correlated relationship. The p-value of Durbin-Watson test is about 0.12, which convinced us the residuals of this model are not serial correlated. When checking the normality of the residuals, we found that the residuals fits the normal distribution pretty well in the qq-plot, with all the points fall in the confidence band.

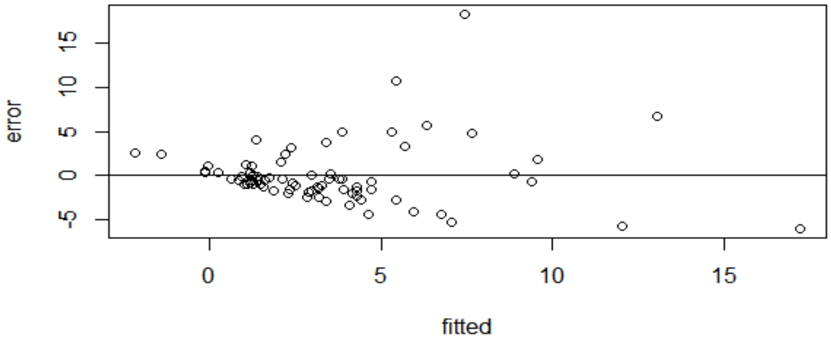




--Structural Assumptions

We used two methods to check the structural assumptions. Firstly, we drew fitted-value vs. residuals plot. A random scatter of points around 0 in the below graph suggests that the linearity assumption for model3 holds. Therefore, we conclude that the general structure of model3 is proper. Compared with model3, although model1 also passed link function test, the fitted-value vs. residuals graph(the second one) for model1 is not as good as model3's, since the scatter of points in the second graph are not as random as the points in the first one.





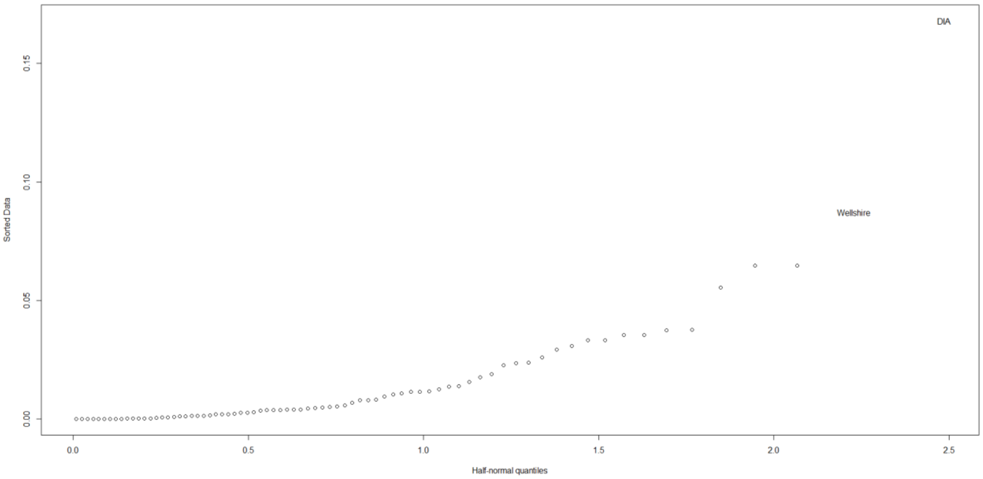
In the next step, we used cr plot as a tool to verify that there is no nonlinear relationship between each variable included in model3 and response, as cr plot can show the impact of a regressor on the fitted values. The cr plots for 3rd diagnostics in the appendix clearly show that the scatter plots for the five regressors are all linear. Especially, the green line and the red line are very close, indicating that it is not necessary to do further transformation. In a nutshell, the structure assumptions hold for model3 and the structure of each regressor in model3 is also proper.

--Influential Observation

```
No Studentized residuals with Bonferonni p < 0.05
```

```
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
10 -3.217963          0.0019678      0.15152
```

Use Bonferonni p-value to test whether the model has potential outliers. Since we have a large Bonferonni p-value (0.15>0.05), it implies the model has does not have outliers.



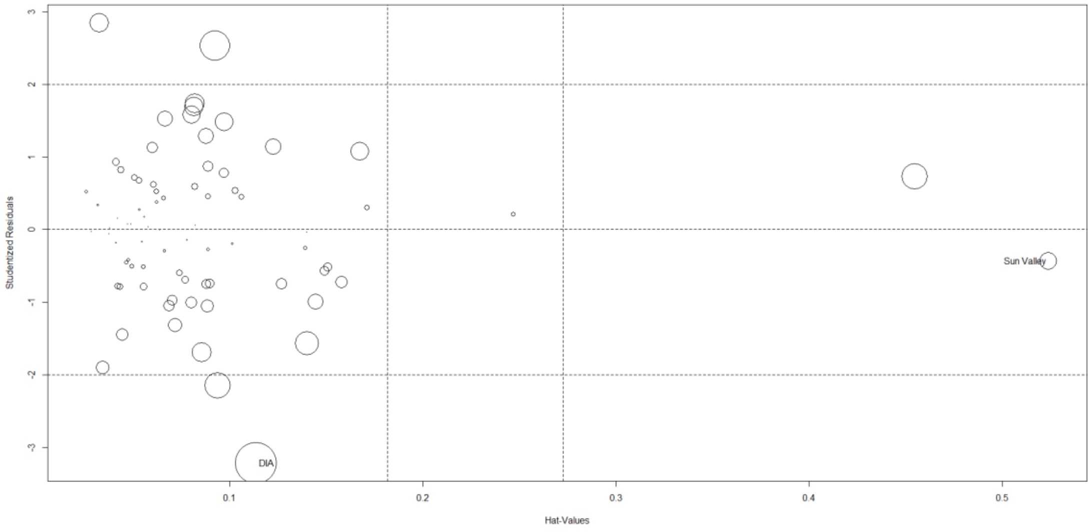
By using half-norm plot of the Cook statistics, we figured out two potential outliers: Wellshire and DIA. Then I am going to compare coefficients with/without potential outliers. After delecting neighborhood ‘DIA’, we believed there was a substantial change in coefficients. the largest change was regressor PCT_FB, with roughly 0.55 multiplicative change.

```
#DIA

                 Est. 1    SE 1  Est. 2    SE 2
(Intercept)     -7.0113  1.2470 -7.7098  1.3074
PCT_WHITE        0.0204  0.0105  0.0277  0.0109
```

```
PCT_ASIAN       -0.0844  0.0316 -0.0863  0.0336

PCT_Otherfamily  0.0419  0.0153  0.0419  0.0163

median          -0.0285  0.0104 -0.0318  0.0110

PCT_FB           0.0192  0.0171  0.0349  0.0174

log2             1.0224  0.1078  1.0463  0.1145
```

After delecting neighborhood ‘Indian Creek’, we believed there was no substantial change in coefficients. The largest change was regressor PCT_Otherfamily, with roughly 0.9 multiplicative change.

```
#Wellshire

                 Est. 1    SE 1  Est. 2    SE 2

(Intercept)     -8.1992  1.3194 -7.7098  1.3074

PCT_WHITE        0.0290  0.0108  0.0277  0.0109

PCT_ASIAN       -0.0829  0.0332 -0.0863  0.0336

PCT_Otherfamily  0.0441  0.0161  0.0419  0.0163

median          -0.0323  0.0108 -0.0318  0.0110

PCT_FB           0.0361  0.0172  0.0349  0.0174

log2             1.0982  0.1168  1.0463  0.1145
```



By using influence plot, we figured out two potential outliers: Sun Valley and DIA. Then I tried to compare the coefficient difference after dropping potential outlier.

After delecting neighborhood 'Sum Valley', we believed there was no substantial change in coefficients.

```
                 Est. 1   SE 1  Est. 2   SE 2
(Intercept)     -7.8729  1.3687 -7.7098  1.3074
PCT_WHITE        0.0291  0.0115  0.0277  0.0109
PCT_ASIAN       -0.0810  0.0360 -0.0863  0.0336
PCT_Otherfamily  0.0480  0.0218  0.0419  0.0163
median          -0.0320  0.0111 -0.0318  0.0110
PCT_FB           0.0340  0.0177  0.0349  0.0174
log2             1.0453  0.1152  1.0463  0.1145
```

Above all, we thought observation 'DIA' is an outliner, since after dropping it, there is a substantial change in our coefficients.

Results

Variable	Estimate	Estimated standard error	Test statistic	p-value
Percentage of White	0.028	0.011	2.528	0.014
Percentage of Asian	-0.086	0.034	-2.568	0.012
Percentage of Other Family	0.042	0.016	2.567	0.012
Median earning (per thousand)	-0.031	0.019	-2.896	0.005
Percentage of Foreign Born	0.035	0.017	2	0.049
Log(crime rate)	1.046	0.115	9.138	0.000

Regressing logarithm of marijuana crime rate on all the other regressors, we concluded that if two places are identical except one has 1% more percentage of white people, we expect that the place with more white people is associated with a multiplicative effect of 1.028 on the mean of marijuana crime rate; If two places are identical except one has 1% more percentage of Asian, we expect that the place with more Asian is associated with a multiplicative effect of 0.917 on the mean of marijuana crime rate; If two places are identical except one has 1% more other family household that a family only has mother or father, we expect that the place with more other family household is associated with a multiplicative effect of 1.0283 on the mean of marijuana crime rate; If two places are identical except one has 1000 dollar more median earning per year, we expect that the place with higher median wage is associated with a multiplicative effect of 0.96 on the mean of marijuana crime rate; If two places are identical except one has 1%

more foreign born population, we expect that the place with more foreign born people is associated with a multiplicative effect of 1.0283 on the mean of marijuana crime rate; If two places are identical except one has a change in crime rate from x to 2*x is associated with a multiplicative effect of 2.064 on the mean of marijuana crime rate.

Conclusion

To sum up, this project aims to analyze the related factors which would impact the marijuana crime rate. After exploring our data, checking collinearity, making model selections and doing diagnostics, we concluded that the predictors such as pct_white, pct_Asian, pct_other_family, median_earnings, pct_fb and log(crime_rate) significantly influence the log(marijuana crime rate) in each neighborhood in Denver Area. Unfortunately, we cannot make a causal conclusion because we got an observational data which only shows an association between marijuana crime rate and the predictors. As the data we got only cover the Denver area, it is unavailable for us to extend the result to a larger population, since different areas have different definitions of marijuana crime under different circumstances.

Moreover, from this project, we learned that data analytics is more difficult than we thought and a comprehensive and profound knowledge about our lectures is needed. For instance, in model selection, we first transformed our response marijuana crime rate into log(marijuana crime rate), but we forgot to transform some predictors at the same time, which caused some problems. If we had explored the new dataset at the beginning, we could have found log(crime rate) a better predictor for our model and it would have saved us a considerable amount of time.

Furthermore, we believe that there are some other methods to improve our study to some extent. For instance, our group only focused on normal distributed data and fitted it into linear model. However, there are many useful models other than linear and other possible distributions such as Poission. Therefore, we could improve our study and the way of handling data as well as collecting data when additional professional knowledge about data analyse is covered.

In addition, since we only analyzed the related factors of marijuana crime rate in Denver Area, we could explore the related factors of marijuana crime rate in other areas in Colorado or other States. We could compare the significant factors in the Denver area to which in other places. We conjecture that it is possible to find similarities and difference in the future research. Figuring out the cause of the similarities and difference is another valuable and interesting topic.

Reference

Denver Open Data Category. (2017). *Crime*. Retrieved from https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime.

Denver Open Data Category. (2017). *Crime Marijuana*. Retrieved from https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime-marijuana.

Denver Open Data Category. (2017). *American Community Survey Blk Grp (2010-2014)*. https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-american-community-survey-blk-grp-2010-2014.