Public Infrastructure and Robbery Rates in Denver

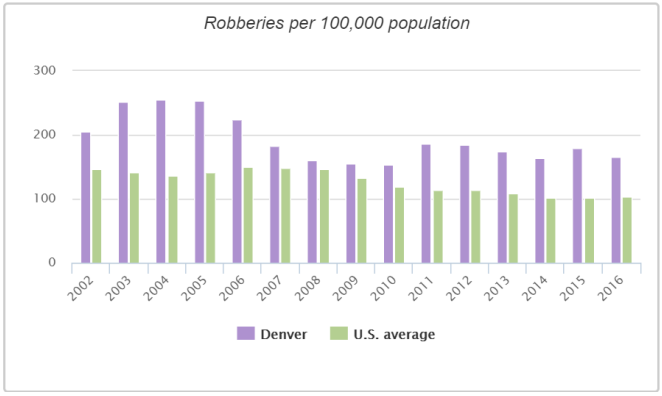
Cara You, Yuting Song, Yang Cheng, Tim Wetzel

Math 4387
Dr. French
12/4/2018

Introduction

Our group has decided to investigate the relationship between the presence of various types of infrastructure in an area and the robbery rates for those neighborhoods. For our study, we are concerned with a variety of common buildings, including police and fire stations, schools of all levels, libraries, and parks. With a wide range of infrastructure regressors, our research will draw interest from local governments and urban planners. A model that can determine how strongly linked these institutions are with robbery allows city officials to better understand the risk for robbery in neighborhoods exhibiting certain characteristics. In turn, this could improve the safety of individual neighborhoods and the city of Denver as a whole.

Denver robberies are descending three years in a row. Comparing monthly: Denver robberies are down this month (134) compared with last (94), and up this year over the same time period last year. Comparing yearly: Denver has had 1065 robberies reported so far this year, an average of 3.3 per day. In 2017, Denver averaged 99.2 per month, and the year before that (2016) Denver averaged 94.7 per month (https://crime.denverpost.com/crime/robbery/#charts). Denver has a higher crime rate compared to other cities in Colorado, and a higher robbery rate compared to other cities the United States.



From the graph above, Denver has a higher robberies rate compared to U. S. average. (http://www.city-data.com/crime/crime-Denver-Colorado.html)

In 2016, a study examined the influence of transportation infrastructure on criminal behavior. The findings suggest that transportation infrastructure can provide potential deterrent for criminal activity (Chippo and Sherri, 2016). Another study in 2015 shows the impact of green stormwater infrastructure installation on surrounding health and safety. The results showed consistent and statistically significant reductions in narcotics possession (18%–27% less) within 16th-mile. Narcotics manufacture and burglaries were also significantly reduced at multiple scales (Kondo and Michelle, 2015). All the above research gives us reason to think there may be some relationships between infrastructure and crime.

The data used for this analysis comes from open data catalogs managed by the city and county of Denver. It was created in partnership with Open Colorado. This is an observational study in which data was observed based on National Incident Based Reporting System (NIBRS). The data is collected by the Denver Police Department which strives to make the data as accurate as possible.

Methods and Results

COLLINEARITY

After selecting a subset of the available variables to begin analysis, multicollinearity checks were necessary to prevent serious problems in the model fit and interpretation of the results. To check for collinearity amongst potential regressors for the model, variance inflation scores were observed and compared. These scores were obtained using the `vif` function.

High variance inflation or correlation issues were used as justification to remove neighborhood area, kindergartens, technical colleges, colleges offering associate's degrees,

colleges offering graduate degrees, and total individual schools. This should not come as a surprise. Clearly there is a trend with the type of variable being omitted: education. Collinearity problems between two variables arise when there is a very close linear relationship between them, and there are often schools grouped together in small areas. It appears that not all these education-related regressors are necessary to adequately explain the robbery rate.

VARIABLE SELECTION

After checking the collinearity, there are 20 variables remaining. We will clear any unnecessary regressors to reduce the noise and achieve more precise estimates. Akaike's Information Criterion (AIC) is an information-based criteria for variable selection. M is the model, L(M) is the log likelihood of the model using the MLE estimates of the parameter, and is the number of regression coefficients in model M. We favor models with smaller AIC. In our model p equals 2 to 20. The model with lowest AIC is when p=11, which indicates we have 10 variables, including: PoliceStations, FireStations, BikeRacks, ParkRestrooms, PublicLibraries, HeadStartPrograms, RecreationCenters, PrekindergartenSchools, ElementarySchools..5th., and Colleges..Bachelors. We also tried to figure out the simplest model using residuals which give us the exact same result as we get from AIC criteria.

CHECKING MODEL STRUCTURE

The residual and component-plus-residual plots of fire stations indicate that there is a nonlinear relationship between fitted value robbery density and the regressor for fire stations. The marginal model plots also tell us that our model does not fit the data very well. The added variable plots look normal. Therefore, we did a transformation of the square root of the regressor fire station.

VARIABLE SELECTION

After checking the structure of the model, we transform fire stations and try to figure out the simplest model again. The variables for fire stations and recreation centers are now excluded. There are only 8 variable left in our model. After the refinement of the model, all the graphs look approximately in linear patterns, and the model fits the data well. There is no need for any further transformation.

CHECKING ERROR ASSUMPTIONS & CHECKING INFLUENTIAL OBSERVATIONS

Since the estimation and inference for a regressor model depends on the error assumption, we should check this assumption using regression diagnostics. First, we examine the constant variance of the residuals by plotting residuals against fitted values and there is not a clear issue with the nonconstant problem since the thickness of the points is roughly constant as we move from left to right along the x-axis. (Appendix). The p-value from the Shapiro-Wilk test is 0.00835, which means we reject the hypothesis that the residuals come from a normal distribution. Finally, we can examine correlated error by computing the Durbin-Watson statistic. The p-value is 0.44, which means we fail to reject the null hypothesis of uncorrelated errors.

Next, we can check for influential observations. We identify the influential observations as they can have a substantial impact on model fit and conclusion. These observations are neighborhoods 2 and 16. We can remove them to see if the fit changes. The coefficient of `HeadStartProgram` changed by about 50% after removing the number 16. The coefficient of `CollegesBachelors` changed by about 40% after removing the number 2. When we remove them together, we can see almost all the coefficients except for `BikeRacks` and `ParkRestrooms` change substantially. After excluding two observations, the coefficients of `PrekindergardenSchools`, `ElementarySchools`, and `CollegesBachelors`

change from significant to insignificant. Therefore, it changes the interpretation of the conclusion. We decide to exclude them, and go back to the procedure.

ADD NEW VARIABLE IN TO THE MODEL

Our model doesn't fit well when checking the model structure because there are too many zeroes present in our variable. We decided to add a few more variables from demographic dataset and try to find better relationship. Investigating collinearity for this new set of variables returns similar results as in the previous section.

VARIABLE SELECTION

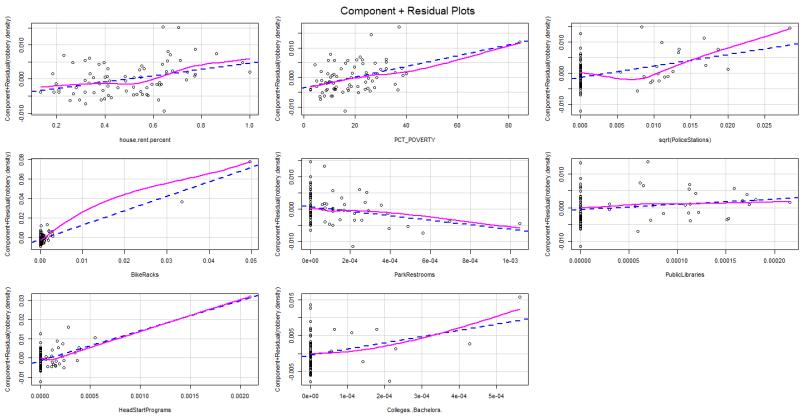
After the collinearity check, we get a model with variables "house.rent.percent", "bachelor.rate", "PCT_POVERTY", "MED_HH_INCOME", "PoliceStations", "BikeRacks", "ParkRestrooms", "PublicLibraries", "HeadStartPrograms", "PrekindergartenSchools", "ElementarySchools..5th.", "Colleges..Bachelors". We use stepwise selection with AIC criterion to simplify the model this time. It turns out variables bachelor.rate and MED_HH_INCOME are excluded. We notice that the two new variables house.rent.percent and PCT_POVERTY have incredibly small coefficients compared with other regressors. This is caused by different scales.

CHECKING MODEL STRUCTURE AGAIN

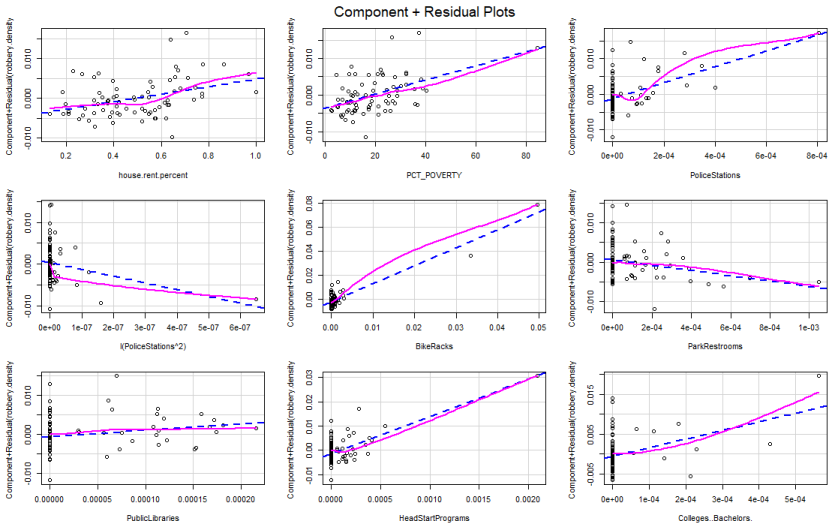
Next, we plot all the plots again (please find the full version of the plots in our code part). The data- and model-driven curves for the bike rack variable are not similar. The marginal relationship between the response and the house rent percent and percent poverty variables are not very deficient, because the curves don't fit the data very well. It's difficult to detect clearly defined non-linear patterns in any of the plots, but there are clearly some influential observations in the police station, bike racks, park restrooms, head start program, and college bachelor plots.

The loess and least-squares fits don't match very well in police station. The residual plot and the component-plus-residual plot of fire stations indicate that there is a nonlinear relationship between robbery density and fire stations, and also between robbery density and bike racks. The marginal model plots also tell us that our model does not fit the data very well. The added variable plots look okay. Therefore, we will transform our fire station variable using square root and square and compare the component-plus-residual plot of each model:

sqrt(PoliceStations):

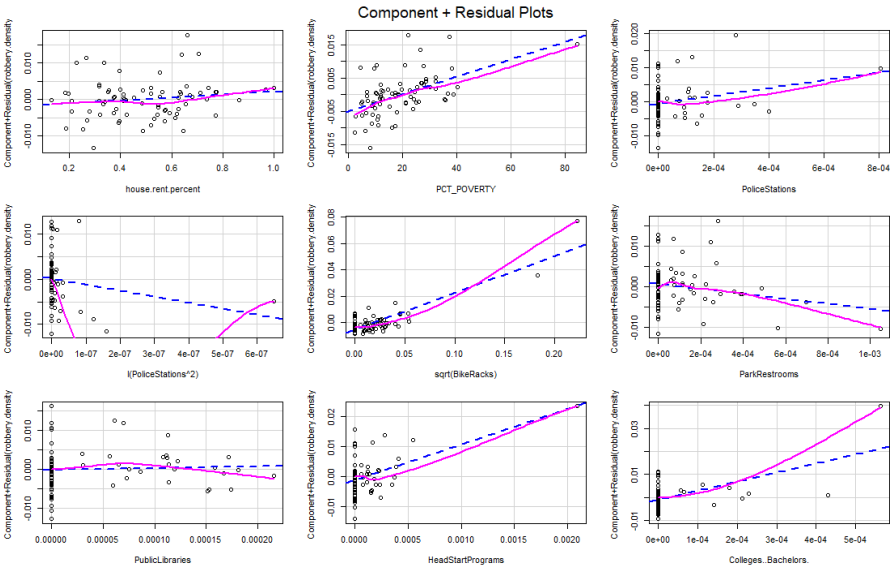


I(PoliceStations^2) :



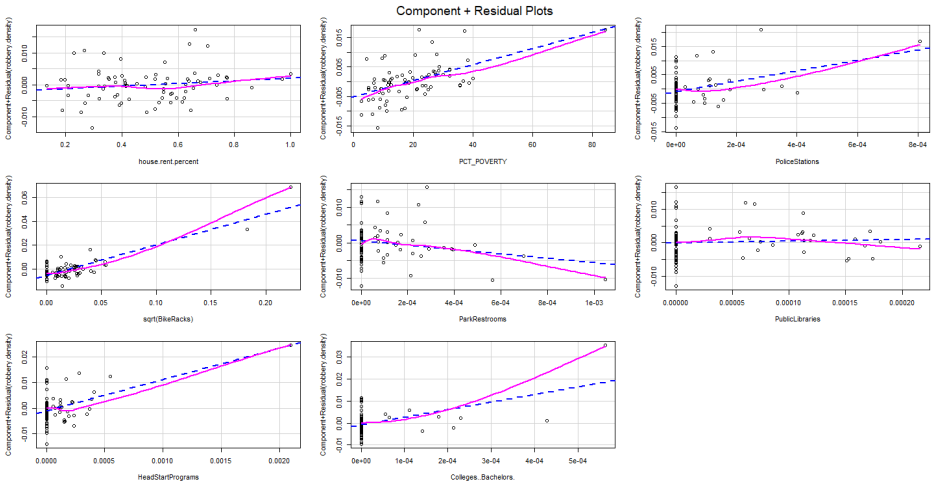
Comparing sqrt(PoliceStations) to I(PoliceStations^2), I(PoliceStations^2) fits the data better, and the loess and least-square fits match better, which contains the squared police station variable. Next we checked if a further transformation on bike racks based on squared police station was necessary or not:

I(PoliceStations^2) + sqrt(BikeRacks):



Comparing I(PoliceStations^2) + sqrt(BikeRacks) to I(PoliceStations^2), I(PoliceStations^2) still fits the data better, and the loess and least-square fits match better. We also tried to transform only the bike racks variable by taking its square root without any transformation on fire station:

sqrt(BikeRacks):



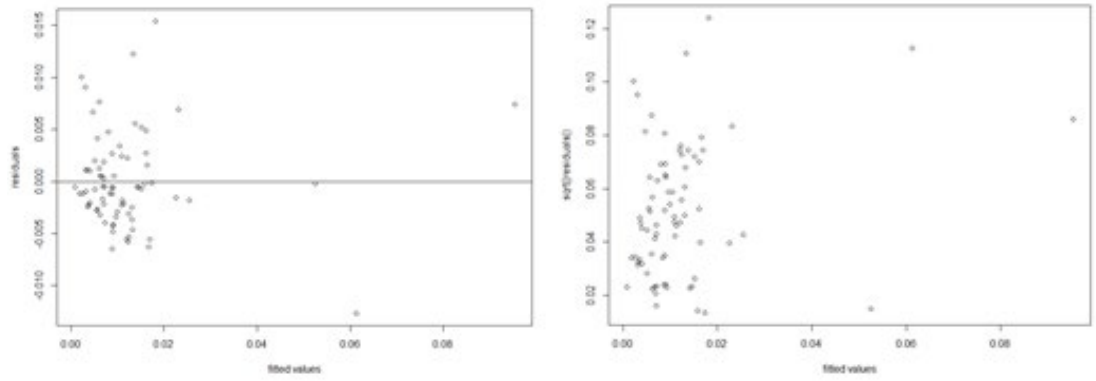
Comparing sqrt(BikeRacks) to I(PoliceStations^2), I(PoliceStations^2) still fits the data better, and the loess and least-square fits match better. Therefore, our final model only transforms police station by taking the square of police station.

Our model now features robbery.density as response with regressors house.rent.percent, PCT_POVERTY, PoliceStations, I(PoliceStations^2), BikeRacks, ParkRestrooms, PublicLibraries, HeadStartPrograms, and Colleges..Bachelors. We then plotted all the graphs again to double check our transformed model.

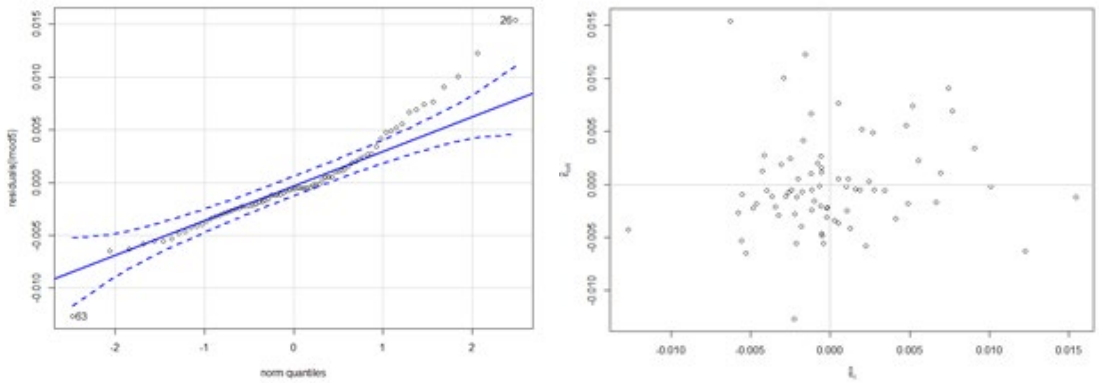
There isn't an obvious systematic pattern from the plots. The model does not fit police station and bike racks very well due to the influential observations in the upper tail. The data- and model-driven curves for the bike rack variable is more similar after transformation. The curves also fit the data better. It's difficult to detect clearly defined non-linear patterns in any of the plots. There are still some influential observations in the police station, bike racks, park restrooms, head start program, and college bachelor plots. After the refinement of the model, all the graphs look approximately in linear patterns, and the model fits the data better, and the loess and least-square fits match better. No need for any further transformation.

CHECKING ERROR ASSUMPTION

First, we examine the constant variance.



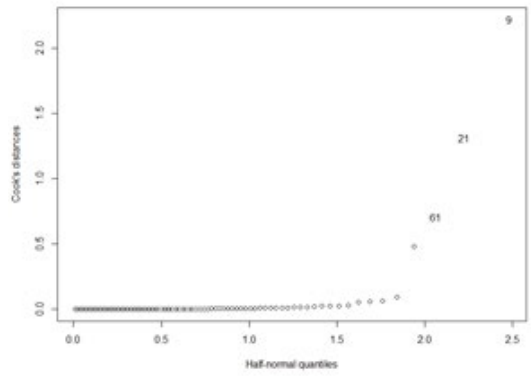
From the plots, there is not a clear problem with nonconstant variance of the errors because the thickness of the points is roughly constant as we move from left to route. Second, we use q-q plot and Shapiro-Wilk test to assess the normality of the residuals.



The P-value from the test is 0.00182. Both the graph and the result of the test indicate that the residuals are not normally distributed. Finally, we will check the independence of the residuals by plotting successive pairs of the residuals and Durbin-Watson test. There is no clear slope in the graph, and the P-value form DW test is 0.1303, which shows that we fail to reject the uncorrelated errors.

CHECKING INFLUENTIAL OBSERVATIONS

Then, we can check for influential observations by using half-norm plot.



From the graph, neighborhood number 9 seems influtional, so, we can remove it to check if our result change substantially. From the output (see Appendix), the magnitude and the significant of our coefficients, and the R square only change a little, which will not affect the conclusion. Therefore, we decide to keep it.

HYPOTHESIS TEST

During the model structure section, we experimented with various transformations to attempt to get better fit from certain variables. An F-test was used to decide whether the reduced model was preferable to the original. Our hypotheses for the test were as follows:

Null Hypothesis: The value of the squared PoliceStations coefficients is not 0 when all other regressors are not 0.

Alternative Hypothesis: The value of the squared PoliceStations coefficient is 0 when all other regressors are not 0.

To perform the test computationally, the `anova` function from the base package of R was used. Using this method and the two models described above, the produced p-value was approximately 0.8048. Given this result, there was ample evidence to reject the null hypothesis that the squared PoliceStations coefficient is nonzero when all other regressors in the model are also nonzero.

INTERPRETATION

Now that we have arrived at a final model, we can observe the coefficients for the regressor variables and provide an interpretation of our findings. The following estimated values are obtained from the summary of the final regression model.

Coefficient	***Estimated Value***	***p-value***
Intercept	-0.0002228	0.899387
Percentage of rental properties	0.0095252	0.015524
Percentage of residents in poverty	0.0001918	0.013695
Police Stations	0.0228991	0.000107
Police Stations (Squared)	-0.0012705	0.804794
Bike racks	1.4724195	< 2e-16
Park restrooms	-6.7440609	0.055904
Public libraries	16.3829863	0.083385
HeadStart programs	15.2903870	1.95e-05
Colleges with bachelor's degrees	18.1490836	0.041231

There are a number of regressors with statistically significant effect. Specifically, we found bike racks, police stations, and HeadStart programs to have a substantial positive link to robbery rates in Denver neighborhoods. Other regressors with lesser positive correlations to robbery rate include rental properties rates, poverty percentages, public libraries, and colleges offering bachelor's degrees.

Of all the regressors remaining in the final model, only one was associated with lower rates of robbery: park restrooms. The p-value for this variable is high-- almost 0.2. The decision to keep park restrooms in the model was based off the conducted hypothesis test. According to the test, our model would be worse off with the exclusion of the variable. This is a difficult measurement

to interpret in the context of our problem, as there is no clear reason why park restrooms should decrease robberies.

Conclusions

Our study hoped to find the effect of the density of infrastructure on the robbery rate for Denver neighborhoods. Since there are many omitted variables that are not included in our model which are likely to be correlated with the predictors and the response, we cannot make a causal conclusion. Things like the average parental education level or the density of single mothers in a neighborhood, the lower the education level of the parents, the more likely for a person to commit to the crime. Also, the low-educated parents will tend to send their children to HeadStart programs due to the fact that HeadStart aims to disadvantaged children, which will lead the area to have more programs. That would overestimate the effect of HeadStart programs on crime rate. In this case, probably, the true causal effect of the head start programs on robbery crime rate is negative. Although we already exclude an variable about education level, the percentage of people with bachelor's degrees, the percentage of people with higher education level may not be a good index we need. We may need percentage of people with lower than high school education level as our regressor. More tests are needed. In terms of colleges with bachelor's degrees, the density of colleges may be correlated with racial diversity. And intercultural conflict may lead to higher robbery rate. And that would also overestimate the effect of density of colleges with bachelor degree on crime rate. Also it is possible college students are more vulnerable to robbery. Because there is high correlation between college bachelor and bike rack (0.710), it may be that more college means more students' bike could be stolen. However we don't have data about percentage of students in college in specific neighborhood. More surveys are needed. Without any doubt, percentage of poverty and percentage of rented house (The number of rented household decided

by total houses hold) did influence robbery rate since floating population may have less responsibility to the neighborhood. And also they have less connection with the neighborhood which makes them vulnerable. Also neighborhood with more poverty people might have less security which makes them vulnerable. And there are problems in terms of reverse causality, it is likely that higher crime rates lead to higher density of the police stations since more police forces are needed.

Also, our study only uses the data from Denver. Since Denver has its own unique characteristics, it may not be representative of the United States; the relationship between the density of infrastructure and the robbery crime rate may be different in other states. Therefore, the result cannot be applied to the areas beside Denver.

Policy

We need more data to figure out whether the high correlation between police stations and robbery rates is because more police stations makes it easier for easier to report robberies. If it is not the case that more police stations lead to more robbery but it is the deficiency of the police station that causes the problem, perhaps we can try to improve the police station efficiency, such as to increase the police force in the neighborhoods where these institutions are present.

The presence of bike racks is also highly correlated with robbery. While we do not have sufficient evidence to causally link this with robbery, perhaps we can improve bike rack construction quality and design. We could also educate bike owners about how to sufficiently lock their bike, such as to choose a fixed immovable bike rack and to use a U-shape lock. We could even implement a mobile application to aid bike owners in tracking their bikes (a strategy which has been successfully implemented in Vancouver, Washington). Poverty percent is highly correlated with robbery. Both HeadStart programs and colleges offering bachelor's degrees have

high correlation with robbery density, which is also probably due to their correlation with poverty percent. Therefore, we could also improve the living condition of the community in order to get poorer people alternative choice to make a living rather than committing robberies, such as: offering more jobs, or improving the coverage of social welfare.

Finally, house renting rate is highly correlated with robbery, which is probably because rented house are more vulnerable to robberies due to their vacancy and mobility. Policy makers should make permanent housing more affordable in an effort of reducing the number of families renting their houses. Rented properties are also associated with a higher rate of robbery, establishing a more permanent community in which more high quality infrastructures are offered may combat this.

Improvement & Extension

From this study, we've learned ways to deal with raw data. Initially, we had 32 variables and we used several methods to inspect collinearity, select variables, check if our model fit the data well, exclude the extreme observations, and repeat these steps many times to get our final model. The model structure checking is difficult since we have a lot of zero values in our data, so there are many points concentrated at the lower-left corner and the lines do not actually fit the data very well. We have to do a lot of transformations and go back through the variable selection to get what we need to include in our final model. For improvement, we think we can collect more information about the neighborhood characteristics and add more controls in the model and lower the impact of omitted variable bias to draw a more causal statement. In terms of the policy, we think we can do further research to find out the types of the robbery and the location where it happened and to find out the main target of the crime.

Reference

Chippo, Sherri Bennett. “A Study of the Relationship Between Transportation Infrastructure and Criminal Behavior.” (2016).

Kondo, Michelle C., et al. “The impact of green stormwater infrastructure installation on surrounding health and safety.” American journal of public health 105.3 (2015): e114-e121.