Impact of Violent Crime Incidence on Communities using Regression Analysis

Gabrielle Taylor
Linear Models | Dr. Alan Huebner
University of Notre Dame

May 4, 2022

1. Introduction

Violent crime has a uniquely powerful role in defining neighborhoods. Since 1995, the rates of violent crime in the United States have declined significantly, but social disparities persist.[2] The exposure to violent crime has found to damage the health and development[4] of not only individual victims, their families, but entire communities by effecting nearby residents.[3] The variation in levels of violent crime is linked to complex characteristics of neighborhoods, including segregation, inequality, immigration, land use, social control, residential instability, social capital (i.e. poverty), and social trust, as well as the characteristics of their nearby neighborhoods.[2] So, although overall violent crime rates have declined substantially, the distribution of violent crime remains about the same due to communities that were initially the most violent generally remaining the most violent.[2]

Within neighborhoods, research has indicated that violent crime occurs in a small number of "hot spots" known as "micro places" (i.e. either street intersections or two blocks faces' on both sides of a street between two intersections).[5] Both informal social controls, such as collective efficacy, and formal social controls, such as the presence of law enforcement, could prevent hot spots.6 In many communities of color, concentrated disadvantage, crime, and imprisonment appear to interact in a continually destabilizing feedback loop. In disadvantaged segregated neighborhoods residents may also be more likely to be detached from social institutions and disregard the law.[7] Evidence suggests that community policing can improve communities' relationships with law enforcement and contribute to strategies such as hot-spot policing that seem to reduce violent crime.[8] Although there has been evidence to indicate that neighborhood characteristics contribute to social disparities, none of the major national sources of crime data provide comprehensive information at the neighborhood level.[2] Due to the absence of annual national neighborhood-level data, researchers are unable to compare trends across and within communities in relations to violent crime.

As for some of the other influential factors towards violent crime: residential instability appears to be interrelated; a Los-Angeles study in the mid-1990s shown the effects of violent crime on instability to be twice as strong as the instability on crime, while other research suggests that the effects of instability on violence lead to reducing community efficacy.[9] Vacancies and evictions can also lead to violent crime by destabilizing communities and creating venues for crime.[2] Some empirical research, however, suggests that mixed-use areas, which combine commercial and residential properties, have lower rates of crime than do commercial-only areas, perhaps by reducing crimes of opportunity.[10] In addition, some studies have shown that violent crime in a neighborhood can lead to crime victims most likely committing crimes, where findings in the U.S. Department of Justice report suggest they're higher levels of violent crime victims in young adults/ teenagers and the elderly compared to others.[11]

By examining these factors and possibly more, identifying the root causes of violent crime along with how it can point to solution-base results in the reduction of its incidence and impact. A regression analysis was performed on violent crimes and community data to obtain potential predictive factors in identifying violent crimes incidences in communities. This report will go through by section: the study data in how the model

was designed and data collection source, the methodologies used to obtain our finalized model, the output results provided from the model, and lastly a discussion of the results along with futuristic recommendations.

2. Study Data

The data source for the study data was retrieve from the University of California Irvine (UCL) Machine Learning Repository, which contains a collection of databases, domain theories, and data generators used for empirical analysis·1 Their data collection was based on communities in the U.S. combined with socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR.

In the study data, 125 variables were found to have any plausible connection to crime along with potential dependent variables, 4 were non-predictive (i.e. community name, state, country code, fold), and 18 that could predict the considered 'Index Crimes' by the FBI (e.g. murders, rapes, robberies, assaults, larcenies, auto-thefts, arsons) along with 2,215 observations that had 221 non-responses, which is explained in further detail in the data source's documentation. Note that most of the related missing data was on the LEMAS data, where communities not found in both the census and crime datasets were omitted. Thus, the FBI notes that uses this data to evaluate communities is over-simplistic with many relevant factors not being included (e.g. communities with large numbers of visitors are viewed with higher per capitia crime, measured by residents, than communities with fewer visitors with all other factors remaining constant).

The variables explored in the analysis were the following: (i) percentage of African-Americans in the population, denoted as `racePctBlack`, (ii) percentage of Hispanic heritage in the population, denoted as `racePctHisp`, (iii) percentage of age group 12 to 29 in the population, denoted as `agePct12t29`, (iv) percentage of age group 65 and up in the population, denoted as `agePct65up`, (v) percentage of total number of divorces, denoted as `totalPctDiv`, (vi) percentage of individuals in dense housing, denoted as `pctPersDenseHouse`, (i.e. more than 1 individual per room), (vii) median household income, denoted as `medIncome`, (in U.S. dollar, $), (viii) percentage of median gross rent in housing income, denoted as `medRentPctHouseInc`, (ix) percentage of individuals under the poverty level, denoted as `pctPopUnderPov`, (x) percentage of vacant houses boarded up, denoted as `pctVacantBoarded`, (xi) percentage of individuals born in the same state they're currently living, denoted as `pctBornSameState`, (xii) percentage of individual's born in a foreign country (outside of U.S.), denoted as `pctForeignBorn`, (xiii) population density (individuals per square mile), denoted as `popDensity`, and (xiv) land area (in square miles) were used as the predictors, denoted as `landArea`; whereas (xv) violent crimes per population of 100,000 individuals, denoted as `violentCrimesPerPop`, was used as the response. In the U.S., the following type of crimes considered violent are murder, rape, robbery, and assault. Although there were more variables that could be considered, these were selected based on previous research studies, as mentioned in the Introduction Section. In the final model, 9 predictor variables along with a sample-size of 1,994 observations were analyzed.

3. Methodolgy

Before any data analysis were performed, the regression model underwent explanatory data analysis that involved some preprocessing procedures, such as the renaming of variables, changing the data types to the appropriate variables, and the removal of missing values (on the desired variables used in the analysis); along with a univariate (examined each predictor's distribution) and bivariate analysis (examined each predictor against the response). As the regression model was fitted, it was checked for multicollinearity using the Variance Inflation Factor (VIF), where predictors (with a VIF greater than 10) that indicated the presents of multicollinearity were removed.

Model Selections were than performed on the regression model to identify the best fitted model in terms of predictivity. The two types of model selection performed were: backward elimination (a.k.a. p-value based selection) and Akaike Information Criteria (AIC). Both methods undergo an automated iteration process that outputs their final recommended model as being the best fitted. Previous studies had found through

simulated data that p-value based selection performs poorly to small changes in data and confirms that the AIC method is more robust.

Model diagnostics were performed shortly after to check whether any of the model assumptions were upheld; thus, checking on the validity of the model's inference(s). The diagnostics analysis consisted of checking for constant error variance using a diagnostic plot and/ or hypothesis testing: Breusch-Pagan (BP) Test; normal errors using a Q-Q plot, histogram, and or hypothesis testing: Shapiro-Walk Test; and independent/ uncorrelated errors using a lagged residual plot and/ or hypothesis testing: Durbin-Watson (DW) Test. Most researchers prefer to rely on the results for the diagnostic/ Q-Q plot(s), histogram(s), and then hypothesis testing in that respected order. This is due to the plots being the less controversial compared to the others. Research has also found that hypothesis testing tends to have over sensitivity to large sample sizes.

Further investigation on the model was performed on the individual observations to check for influence (with a Cook's distance greater than the F statistic threshold) and whether any suggested their removal. Model transformation was also used to facilitate in improving the model's fit and correct violations of model assumptions (e.g. heteroscedasticity). The Box-Cox Method was used to transform the response value, where diagnostics of the model were checked once more to indicate whether the transformation worked appropriately. Lastly, reported inferences and predictions were made from the final model using 95% confidence intervals (CI) explaining the true mean and of a particular instance. In addition, an importance plot/ table was used in identifying the importance of each predictor in relations to the response using random forest modeling. In the next section, we'll discuss the results.

4. Results

In the results section, all of the data analysis performed were from Steps 1-7 with an alpha level of 0.05.

4.1. Exploratory Data Analaysis

In Step 1, all exploratory data analysis were performed on each predictor variables against the response variable, violent crimes per population, in hopes to capture key factors that determines its predictive behavior. The univariate and bivariate analyses along with their summary outputs were performed on each of the predictor variable in Figure 4.1.1-4.1.4 and Figure 4.1.5-4.1.8, respectfully. Only 4 of the predictors will be referenced in Section 4.1.

For the univariate analysis:

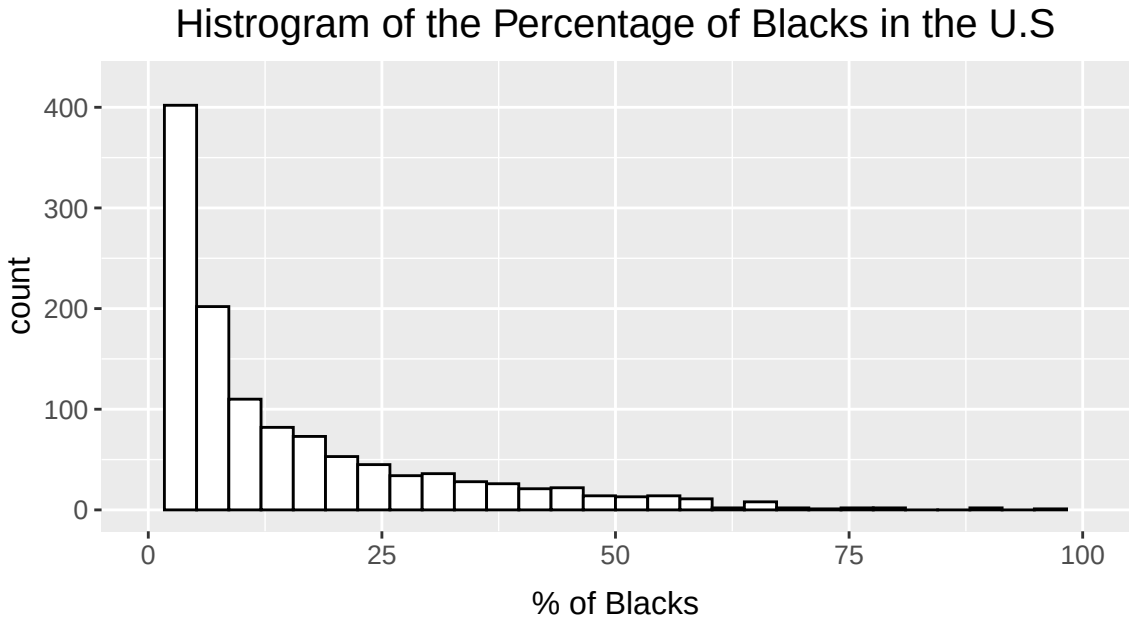


Figure 4.1.1

Figure 4.1.1 displays the distribution of African-Americans in the U.S. population as a percentage with a heavy right skewed tail shape of a standard deviation of 14.10% and a median of 3.15%.

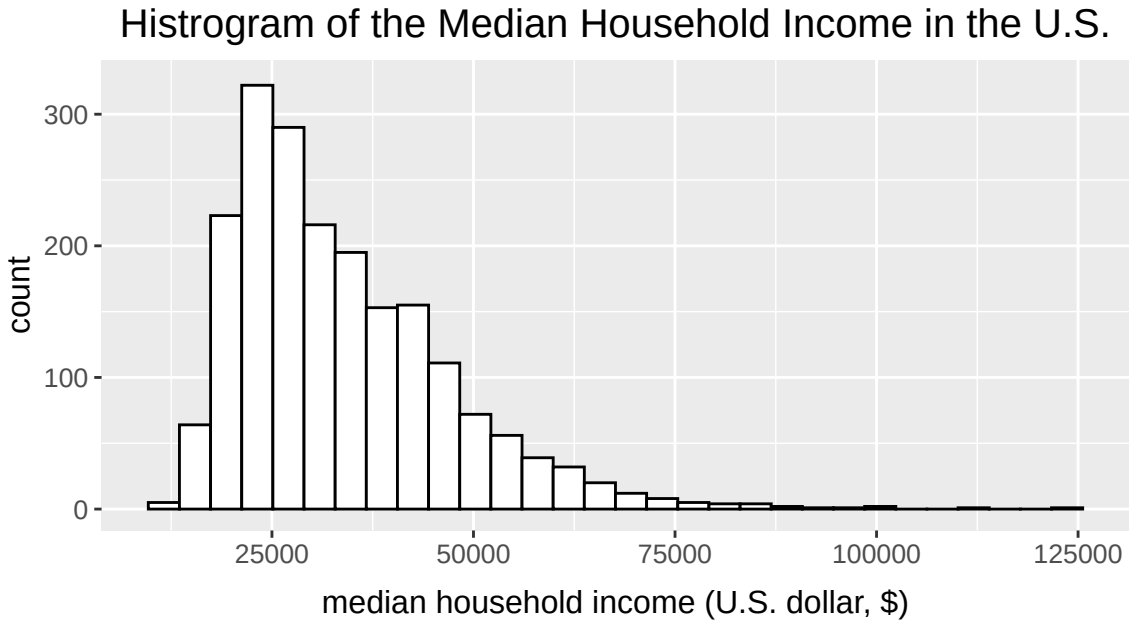


Figure 4.1.2

Figure 4.1.2 displays the distribution of the median household income in the U.S. population as dollars with a right skewed tail shape of a standard deviation of $13,391.74 and a median of $30,896.

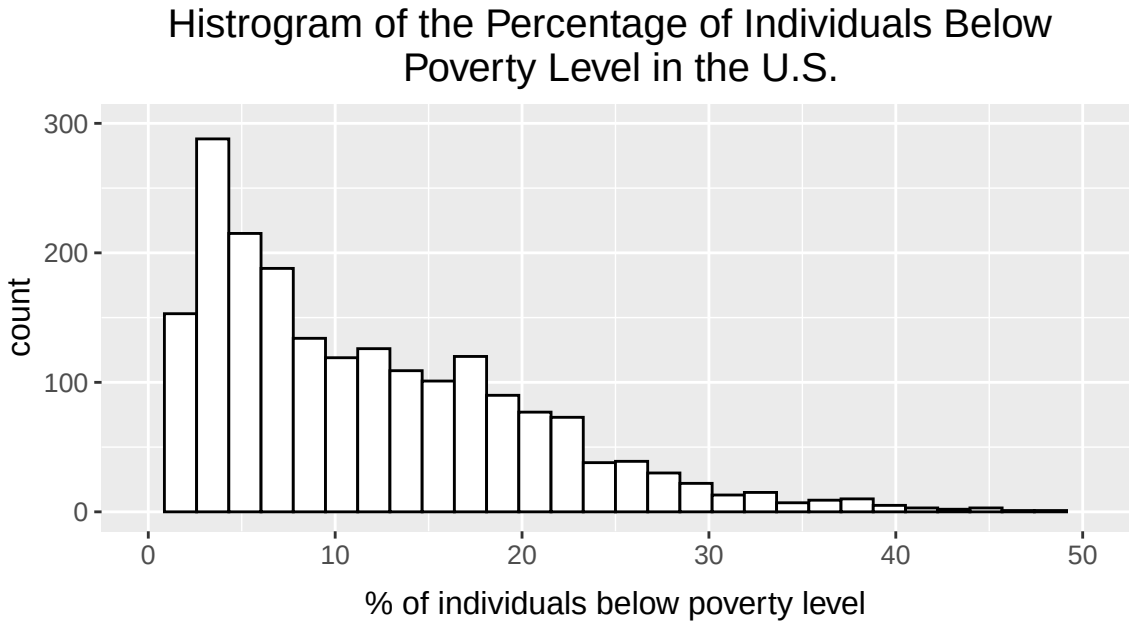


Figure 4.1.3

Figure 4.1.3 displays the distribution of individuals under the poverty level in the U.S. population as a percentage with a right skewed tail shape of a standard deviation of 8.51% and a median of 9.65%.

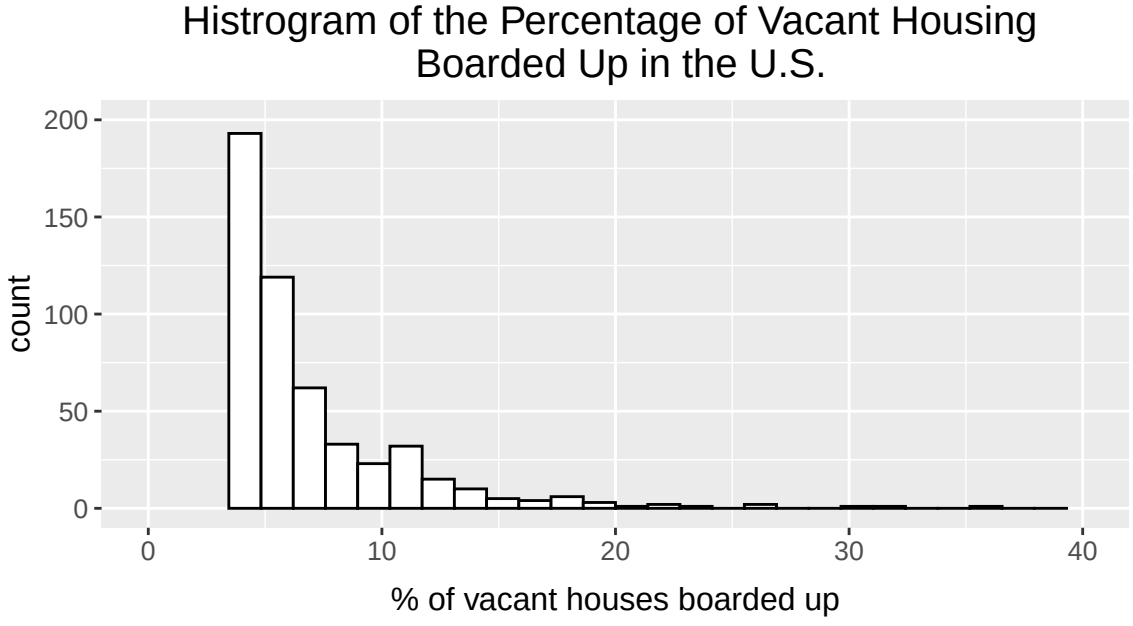


Figure 4.1.4

Figure 4.1.4 displays the distribution of the number of vacant houses boarded up in the U.S. population as a percentage with a heavy right skewed tail shape of a standard deviation of 3.45% and a median of 1.74%.

Overall, based on the high levels of skewness among each of the predictor variables it can be infer that the data may have to undergo some type of transformation to transform the model data to a more Gaussian or Normal distribution.

For the bivariate analysis:

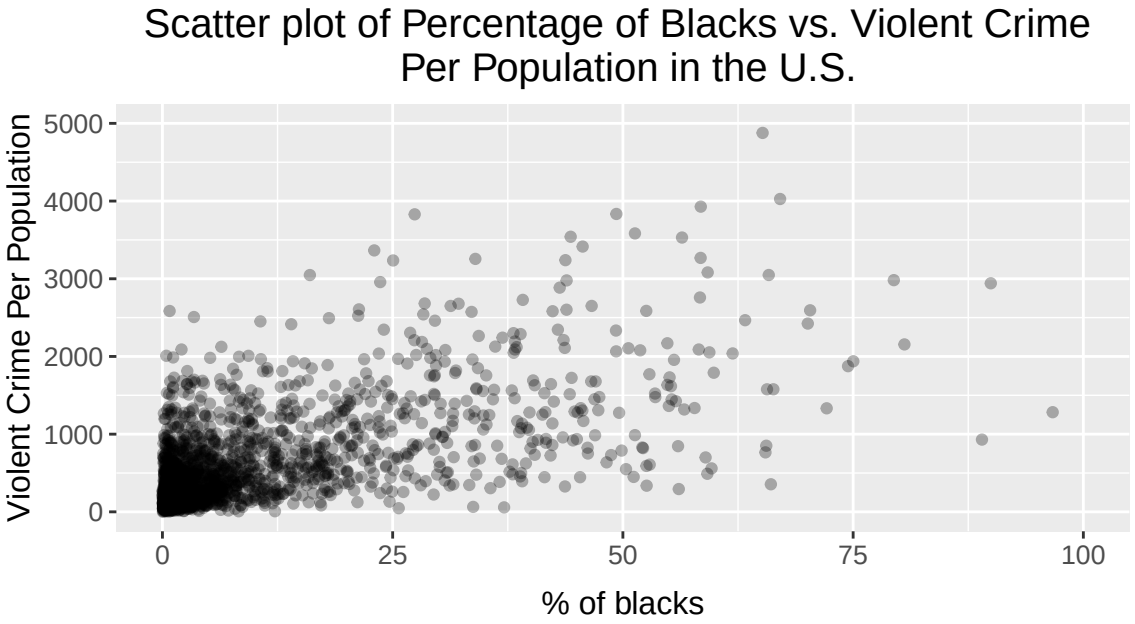


Figure 4.1.5

Figure 4.1.5 displays a scatter plot describing the relationship between the percentage of African-Americans in the U.S. population and the number of violent crimes per population. The plot suggests that the variables have a moderately positive linear trend. Indicating that the higher the percentage of African-Americans is within the population, the higher the count for violent crimes' incidence per population.

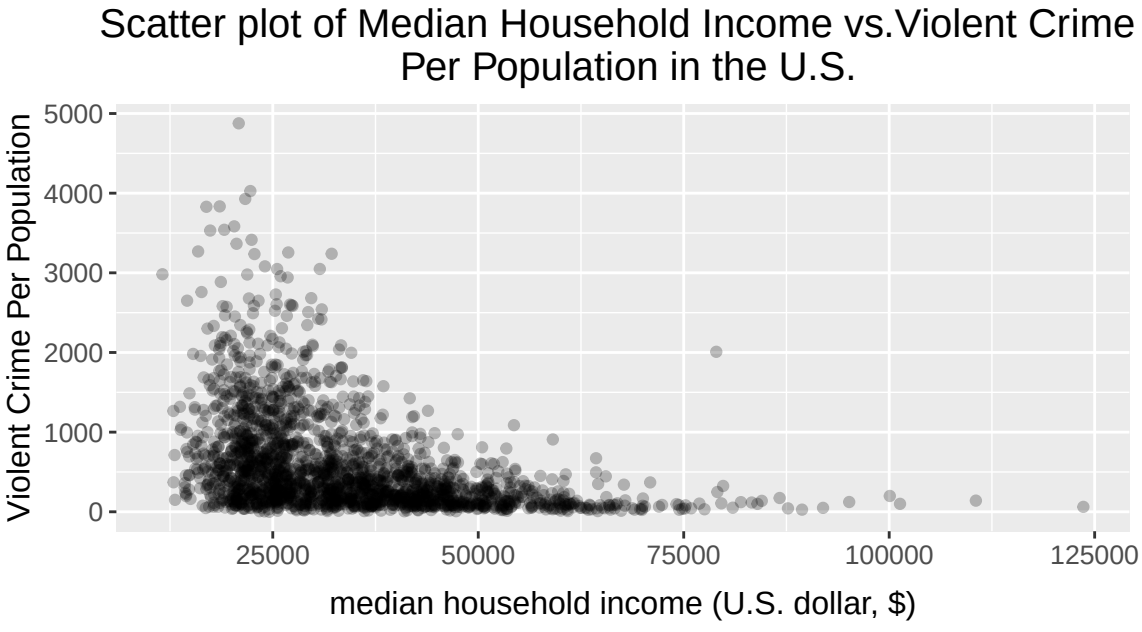


Figure 4.1.6

Figure 4.1.6 displays a scatter plot describing the relationship between the median household income in the U.S. population and the number of violent crimes per population. The plot suggests that the variables have a strong (steeper slope) negative linear trend with the expect for a few potential outliers. Indicating that the lower the median household income is within the population, the higher the count for violent crimes' incidence per population.

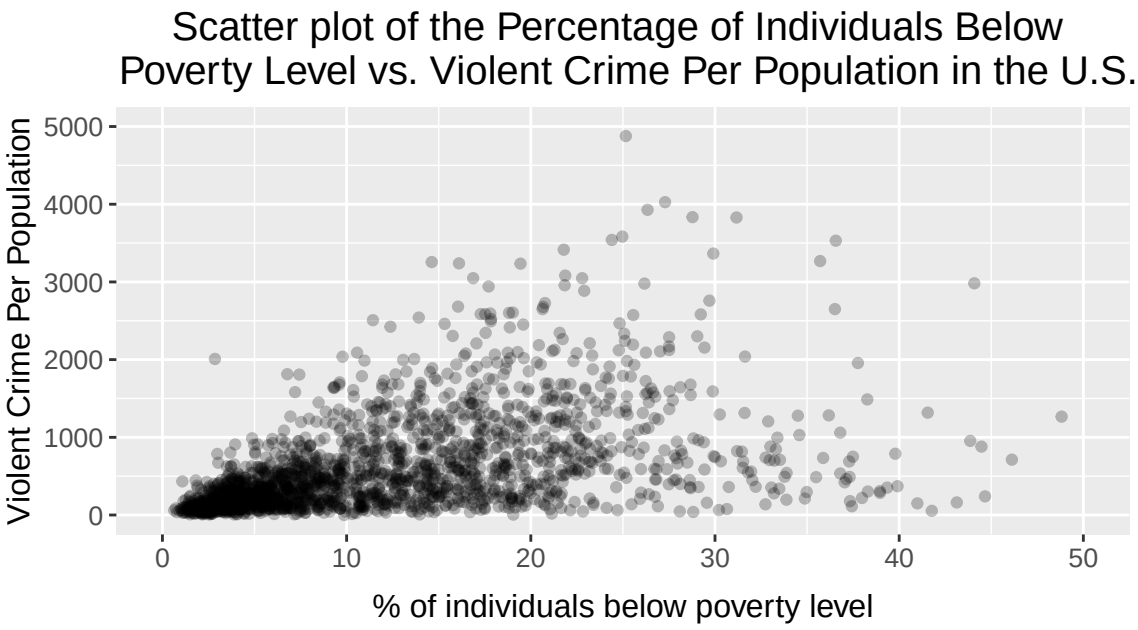


Figure 4.1.7

Figure 4.1.7 displays a scatter plot describing the relationship between the percentage of individuals under the poverty level in the U.S. population and the number of violent crimes per population. The plot suggests that the variables have a moderately (wider slope) positive linear trend. Indicating that the higher the percentage of individuals under the poverty level is within the population, the higher the count for violent crimes' incidence per population.

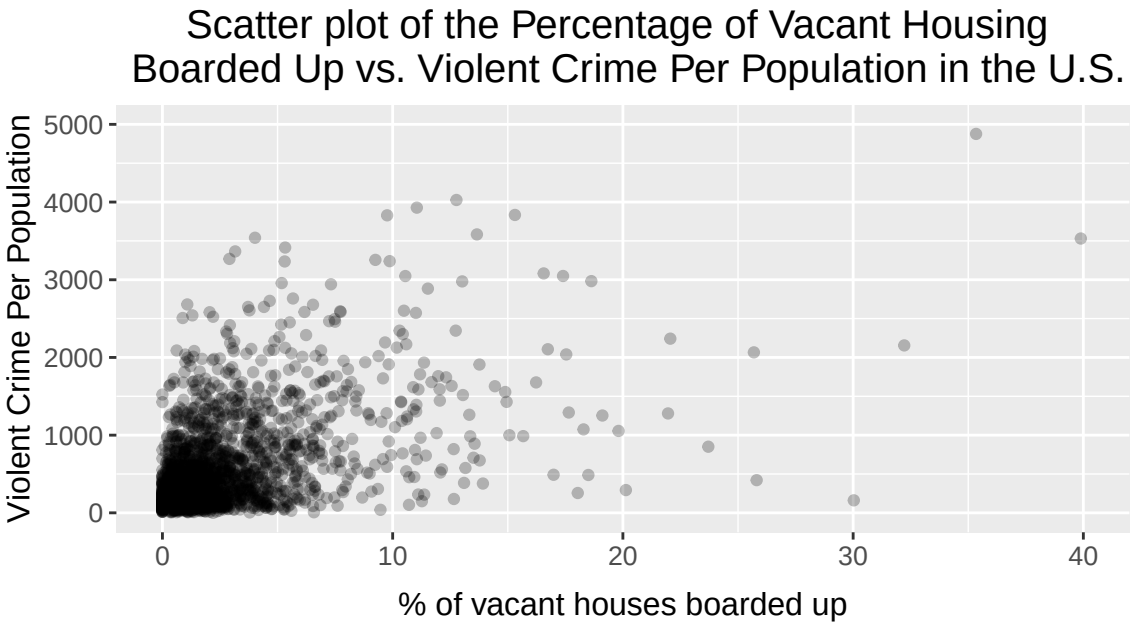


Figure 4.1.8

Figure 4.1.8 displays a scatter plot describing the relationship between the percentage of vacant housing boarded up in the U.S. population and the number of violent crimes per population. The plot suggests that the variables have a (steep slope) positive linear trend with the exception for a few potential outliers.

Indicating that the higher the percentage of vacant housing boarded up is within the population, the higher the count for violent crimes' incidence per population.

Overall, based on the bivariate analysis, we could expect for all the predictor variables to have positive association towards the response, violent crimes per population.

4.2. Fit a Linear Model

In Step 2, all the 14 predictors along with the response variable were fitted as a linear regression model, which can be denoted as follows

$$ViolentCrimesPerPop = racePctBlackX_1 + racePctHispX_2 + agePct12t29X_3 + agePct65upX_4 + totalPctDivX_5 + pctPersDenseHouseX_6 + medIncomeX_7 + medRentPctHouseIncX_8 + pctPopUnderPovX_9 + pctVacantBoardedX_{10} + pctBornSameStateX_{11} + pctForeignBornX_{12} + popDensityX_{13} + landAreaX_{14}.$$

Using the model's summary statistic output, the following variables were found to be statistically significant from Equation 1: `racePctBlack`, `agePct65up`, `totalPctDiv` , `pctPersDenseHouse`, `medRentPctHouseInc`, `pctPopUnderPov`, `pctVacantBoarded`, `pctForeignBorn`, and `landArea` along with a coefficient of determination, R-squared, of 0.5995. Thus, suggesting that 59.95% of the variation in the number of violent crimes per population can be explained by all the predictor variables in the model.

4.3. Perform Model Selection

In Step 3, two types of model selection methods (e.g. p-value based selection and AIC) were performed on the regression model in Equation 1, where both provided their best recommended fitted models. For the p-value based selection, its final recommended model removed the following variables: `agePct12t29`, `racePctHisp`, `pctBornSameState`, `medIncome`, `popDensity`, `pctPopUnderPov`, and `landArea`. As for the AIC model selection, its final recommended model removed the same following variables as the p-value based selection with the exception for `pctPopUnderPov` and `landArea`. The initial model had an AIC value of 23812.24 that reduced to 23805.02 with the removal of the 5 predictor variables. Both methods for the most part matched our initial intuitions with the exception for the removal of `pctPopUnderPov` and `landArea` in the p-value based selection. However, based on the results being similar and AIC being known as more robust than the p-value based, the AIC recommended model was chosen to conduct all further data analyses. Thus, the new regression model can be denoted as follows

$$ViolentCrimesPerPop = racePctBlackX_1 + agePct65upX_4 + totalPctDivX_5 + pctPersDenseHouseX_6 + medRentPctHouseIncX_8 + pctPopUnderPovX_9 + pctVacantBoardedX_{10} + pctForeignBornX_{12} + landAreaX_{14}.$$

Based on the model summary statistic output for Equation 2, it had found all the predictor variables to be highly statistically significant with the exception of `pctPopUnderPovX_9` only having 90% confidence and `landArea` just meeting the 95% confidence threshold for significance. This makes sense for the p-value based selection's recommendation excluding `pctPopUnderPov` and `landArea`; however, based out literature review, studies show these variables to play a factor towards the increase adverse effects in violent crimes' incidence among communities. In addition, although the, R-squared, value was barely affected with a decline of 0.5%, the adjusted coefficient of determination, Ra-squared, had increased from 0.5967 to 0.5972. Thus, suggesting the removal of those predictors improved the model's overall performance on predictability in the variation of the number of violent crimes per population.

4.4. Perform Model Diagnostics

In Step 4, three levels of diagnostics were performed on regression model in Equation 2 that check for any violations on model assumptions.

Perform the appropriate methods to check for constant error variance assumption:

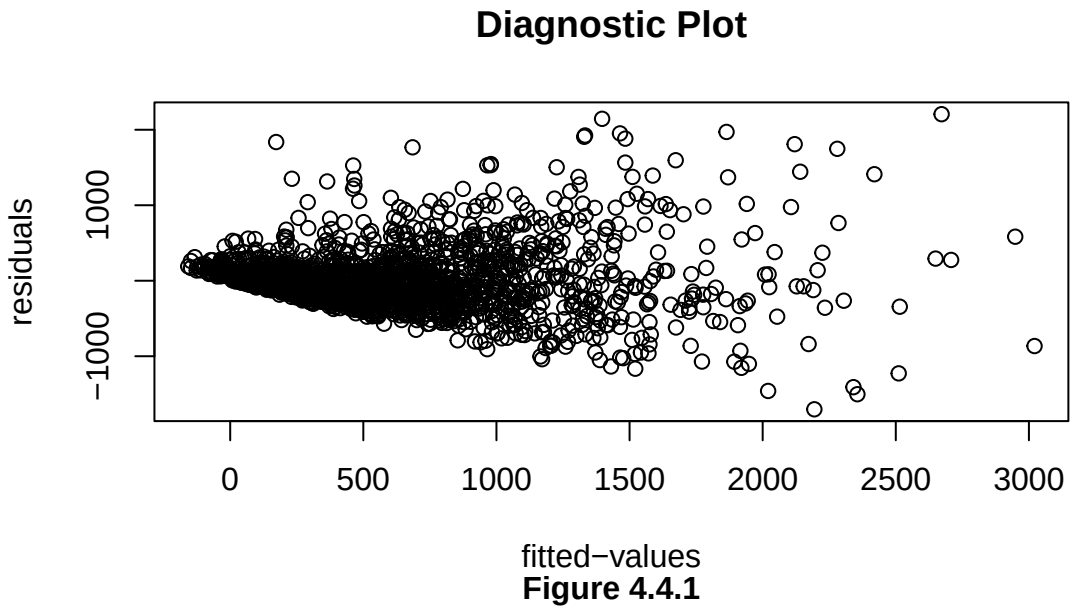
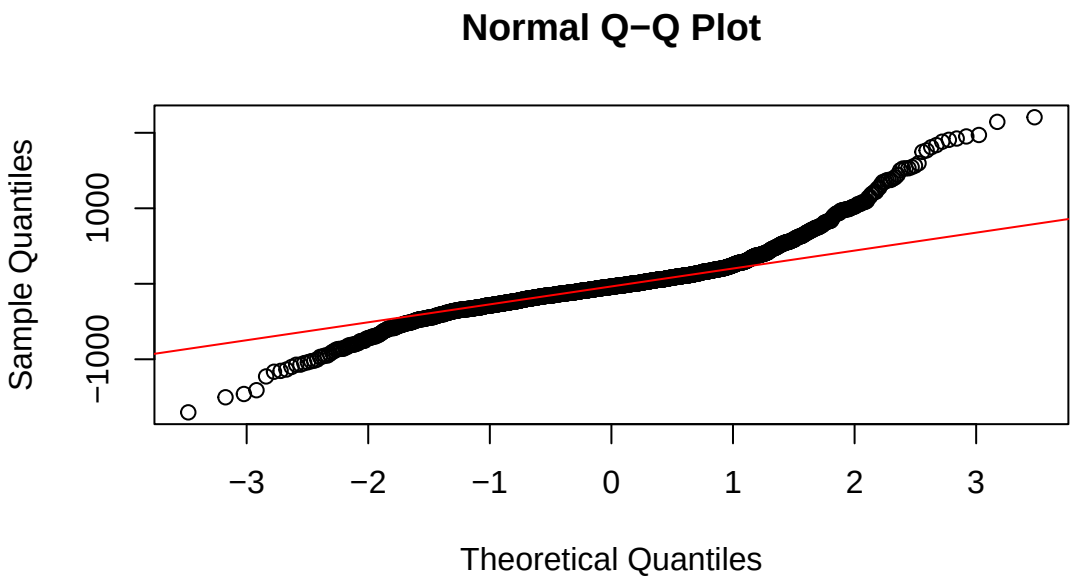


Figure 4.4.1

Based on the diagnostic plot in Figure 4.4.1, the relationship between the model's fitted-values and residuals demonstrated a negative linear pattern; thus, indicating the residuals have non-constant variance. Based on the BP Test, the test statistic demonstrates a p-value, $< 2.2e\text{-}16$, below any appropriate alpha level; thus, indicating the residuals have non-constant variance. Using both the diagnostic plot and hypothesis testing, the BP Test, the results suggests that the model assumption for constant error variances is violated.

Perform the appropriate methods to check for the assumption of normal errors:



Based on the Q-Q plot in Figure 4.4.2 (above), the relationship between the sample and theoretical quantiles displays a bend within its overall trend line; thus, indicating the residuals are not normally distributed.

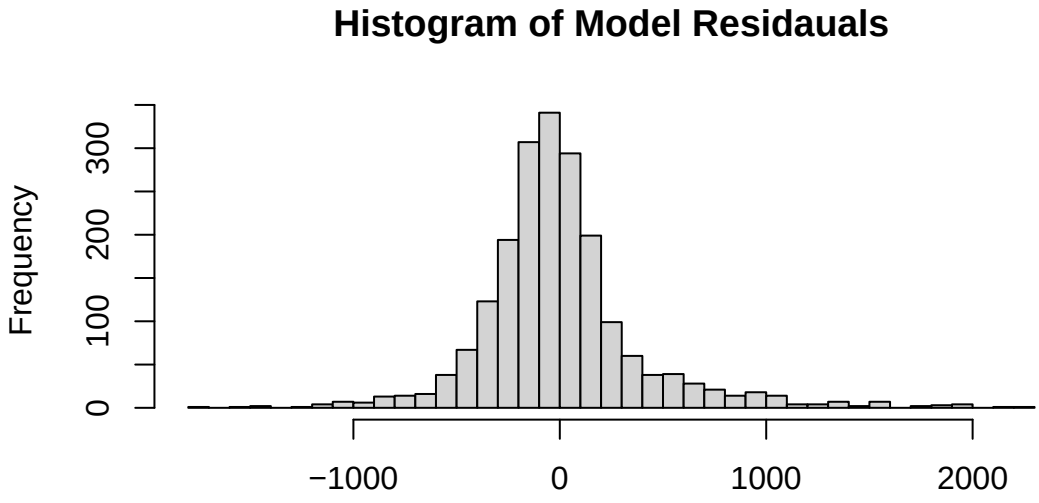


Figure 4.4.3

By looking at the distribution for the model residuals using the histogram in Figure 4.4.3, we notice the bend in the Q-Q plot is demonstrated around the residual's center of shift from a mean of zero. In addition to the Shapiro-Walks Test, the test statistic demonstrates a p-value, $< 2.2e\text{-}16$, below any appropriate alpha level; thus, indicating the residuals are non-normal. All method results suggests that the model assumption for normal errors is violated.

Perform the appropriate methods to check for independent/ uncorrelated errors:

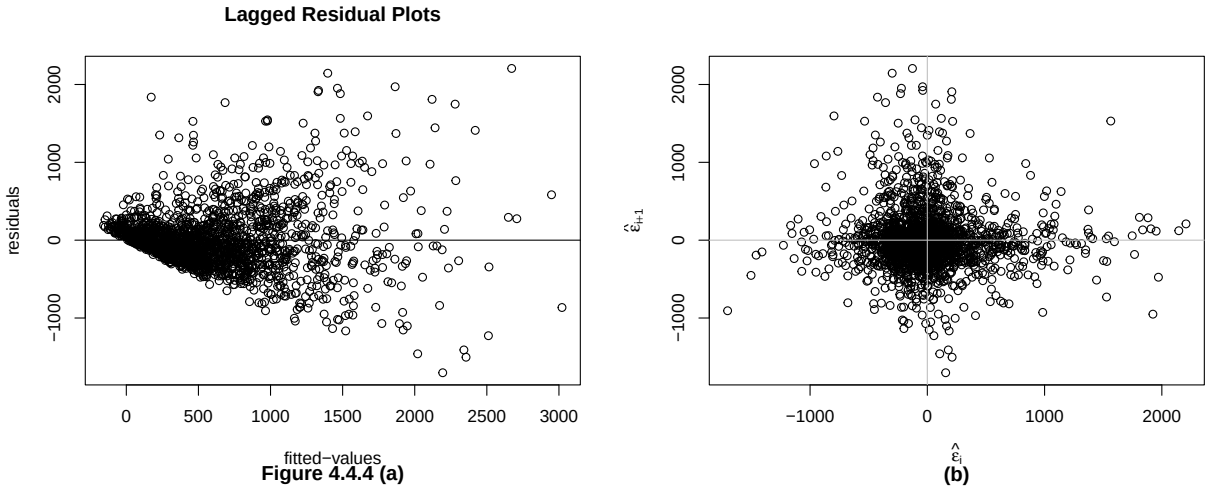


Figure 4.4.4 (a)

Based on lagged residual plot(s) in Figure 4.4.4 (a-b), the relationship between the residuals and fitted-values shown in (a) and the successive pairs of residuals shown in (b) display a heavy cluster density around the origin; thus, indicating the residuals to be dependent/ correlated. However, based on the DW Test, the test statistic demonstrates a p-value, 0.9419, above any appropriate alpha level; thus, indicating the residuals are independent/ uncorrelated. Despite the opposing results from the DW test, we conclude that the model assumption for error independence is violated.

As inferred in the exploratory analysis, the model diagnostics suggests that the model's performance is being adversely affected by potential outliers that might be influential. In determining whether that's the case, further investigation must be taken place. Also suggesting that there's a necessity to transform the skewed data to being more normally distributed, which is demonstrated in the next two sections accordingly.

4.5. Investigate Fit for Individual Observations

In Step 5, further investigations were performed on the model data to determine if any outliers, leverages, and/ or influential points are present. Based on the rule of thumb (absolute value of residuals that exceed a value of 3) and using the absolute standardized residuals, the model data had identified roughly 39 outliers. Based on the rule of thumb for leverages, 156 observations were identified with high leverage, where 12 of those were also outliers. Using the Cook's distance with a F statistic of 0.9345 (threshold of F distribution), there were evidence of disproportionately large influential observations in the regression model. Thus, despite the presents of outliers and leverage, no observations were removed from the model due to no evidence of there being high influence. However, this does not solve our model's performance and validity. So, instead we must transform our model.

4.6. Apply Transformations to the Model

In Step 6, the model data underwent a Box-Cox Transformation to correct the model assumptions and performance.

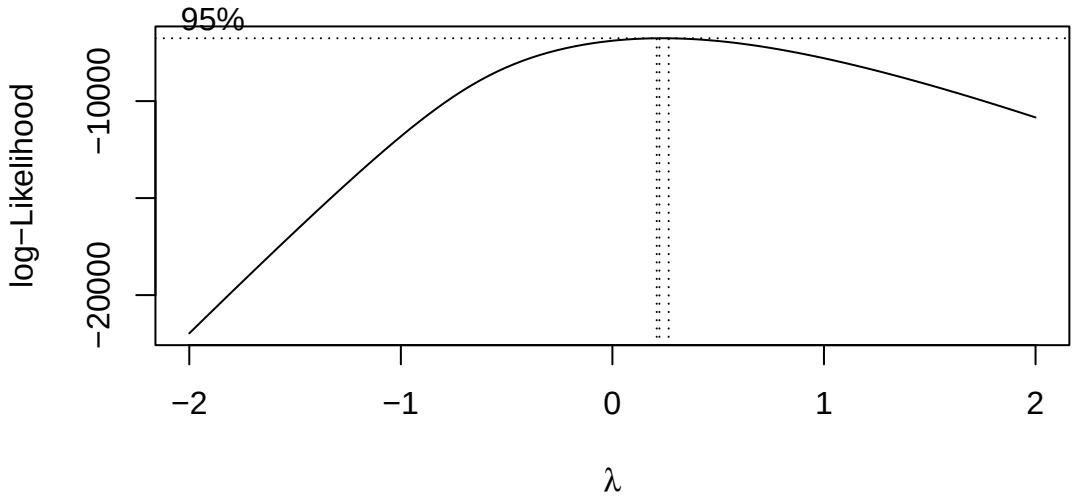


Figure 4.6.1. (above) represents the log transformation of the Box-Cox method, which shows the 95% confidence bounds and optimal value for the transformation index (i.e. lambda). The optimal lambda value was found to be about 0.2222, which was used to transform model data.

The transformed model was than re-fitted, which can be defined as follows

$ViolentCrimesPerPop^{\lambda} = racePctBlackX_1 + agePct65upX_4 + totalPctDivX_5 + pctPersDenseHouseX_6 + medRentPctHouseIncX_8 + pctPopUnderPovX_9 + pctVacantBoardedX_{10} + pctForeignBornX_{12} + landAreaX_{14}$.

Once the model data from Equation 2 was transformed, all model assumptions had to be re-performed to check the validity of the model inference in Equation 3.

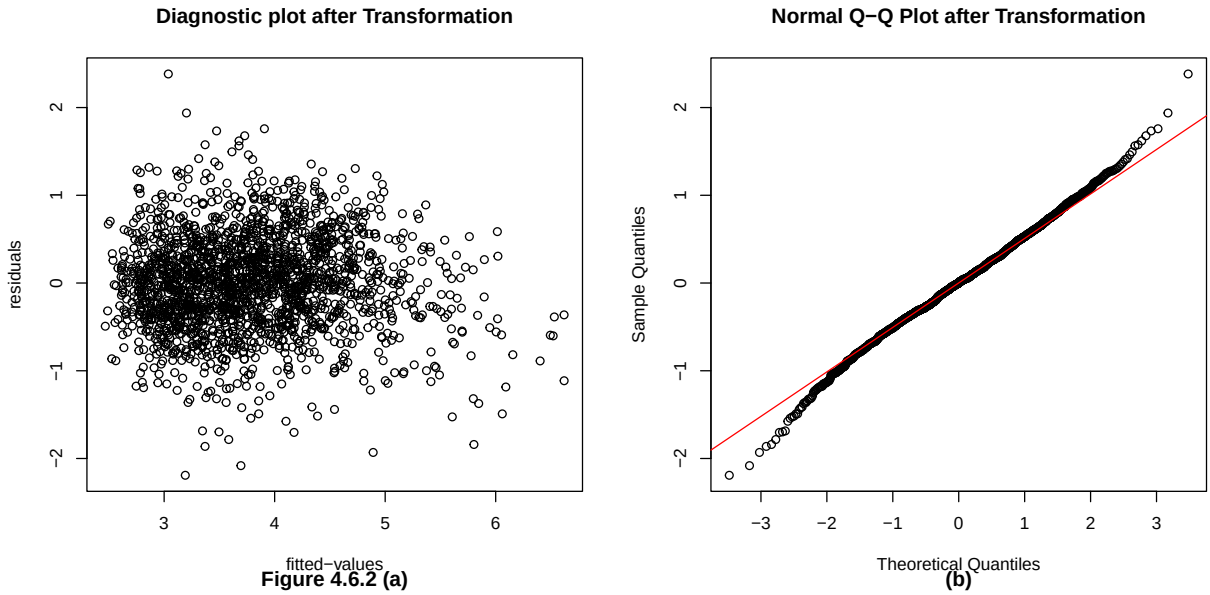
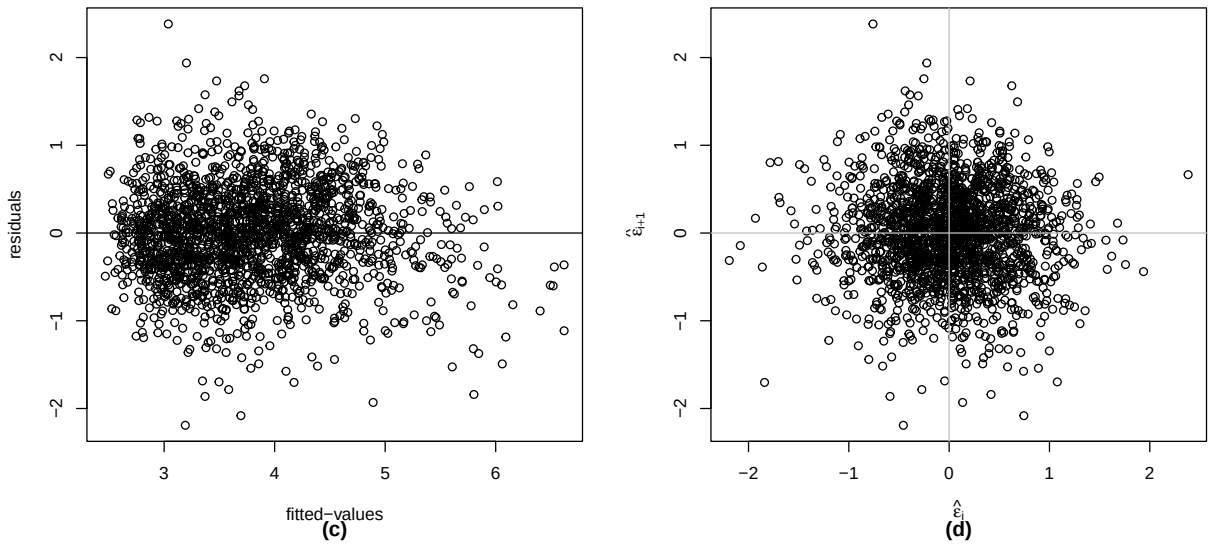


Figure 4.6.2 (a) (b)



As a result in Figure 4.6.2 (a-d), the diagnostic plot shown in (a) and the Q-Q plot shown in (b) displays both model assumptions are met for constant error variance and normal error. Although the model diagnostics had improved overall in terms of correcting some of the assumptions, the lagged residual plots shown in (c-d) had still shown evidence of some correlated residuals. Despite this, we still infer that using the transformed model in Equation 3 would be the best suited model for making predictions.

4.7. Report Inferences and Make Predictions using the Final Model

In Section 7, model inferences and predictions were made based on Equation 3, where all numerical units are in terms of the logarithmic value of the response variable. Although the model's R-squared value increased to roughly 0.6338, it must be noted it is in terms of explaining the variation of the logarithmic value of the response variable as well.

Based on the table in Figure 4.7.1, these are the reported estimates and corresponding p-values for each of the parameters in the model. Since the parameter estimates are in terms of the logarithmic value of the

Table 1: Figure 4.7.1

Parameters	Estimates	p-values
(Intercept)	1.20699849607727	5.13284672130926e-20
racePctBlack	0.019900547858859	1.54271898765532e-69
agePct65up	0.0126710251284758	4.20346665913393e-06
totalPctDiv	0.117025614738102	4.66511843362633e-116
pctPersDenseHouse	0.0264264853659176	8.79114498959445e-11
medRentPctHouseInc	0.0220455274223214	1.07899764120672e-05
pctPopUnderPov	0.00961875990055719	1.74470095407041e-05
pctVacantBoarded	0.0166823243601443	8.30288943508287e-05
pctForeignBorn	0.0133361012522799	6.82654344233981e-08
landArea	0.000222701546109803	0.0374907720521936

response, they are not easy to interpret.

Thus, when examining predictor's confidence intervals of their inverse function we are able to interpret the results back in the original units of the model. For instance, we are 95% confidence that the slope estimate for the percentage of African-Americans in the U.S. population is between 2.1888e-08 and 2.2365e-08 in the actual value of violent crimes per population. Note that the percentage of African-Americans were found to be the most important variable using an importance plot through random forest modeling.

Lastly, taken the median values of all the predictor variables from the final model we inferred the following:

For populations at the median levels of land area; and percentage of African-Americans, age groups of 65 and up, total number of divorces, individuals per dense housing, median gross rent of housing income, vacant houses boarded up, individuals born out of the U.S., and the population under the poverty level, we are 95% confident that the true mean of violent crimes per population of 100,000 individuals is between roughly 284.3 and 306.6 instances per population.

For populations at the median levels of land area; and percentage of African-Americans, age groups of 65 and up, total number of divorces, individuals per dense housing, median gross rent of housing income, vacant houses boarded up, individuals born out of the U.S., and the population under the poverty level, we are 95% confident that a particular value of violent crimes per population of 100,000 individuals is between roughly 58.8 and 966.3 instances per population.

5. Conclusion/ Recommendations

Based on the analysis results, we can infer that larger percentages of African-Americans, age groups of 65 and up, total number of divorces, individuals per dense housing, median gross rent of housing income, vacant houses boarded up, individuals born out of the U.S., and the population under the poverty level would lead to higher numbers of violent crimes incidence per population.Thus, based on the model, targeting communities among these factors must be provided with assistance in reforming the communities on a whole. Based on the research from peer literature review, our analysis for the most part coincides with the literature results. Many research suggests that in order to reduce and improve the safety and health of many of theses communities, we must conduct further investigation on what cycles crime. It would be recommended to provide broader social process and economic gains to those of communities with higher levels of violent crimes. These factors are known to be linked to violent crime and could reduce levels of disparities along with saving the community on a whole from the impact of these crimes.

References

1. Michael Redmond, 2011. "UCI Machine Learning Repository: Communities and Crime Unnormalized Data Set". [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized#
2. Huduser.gov. 2016. "Neighborhoods and Violent Crime | HUD USER". [online] Available at: https://www.huduser.gov/portal/periodicals/em/summer16/highlight2.html
3. Patrick Sharkey and Robert J. Sampson. 2015. "Violence, Cognition, and Neighborhood Inequality in America," in Social Neuroscience: Brain, Mind, and Society, Russell Schutt, Matcheri S. Keshavan, and Larry J. Seidman, eds. Cambridge: Harvard University Press.
4. E.g., David J. Hardin. 2009. "Collateral Consequences of Violence in Disadvantaged Neighborhoods," Social Forces 88:2, 757–84.
5. Anthony A. Braga, Andrew V. Papachristos, and David M. Hureau. 2009. "The Concentration and Stability of Gun Violence at Micro Places in Boston, 19802008," Journal of Quantitative Criminology 26:1, 33–53.
6. David Weisburd, Elizabeth R. Groff, and Sue-Ming Yang. 2013. "Understanding and Controlling Hot Spots of Crime: The Importance of Formal and Informal Social Controls," Prevention Science 15:1, 31–43.
7. Robert J. Sampson and Charles Loeffler. 2010. "Punishment's Place: The Local Concentration of Mass Incarceration," Daedalus 139:3, 20–31.
8. Anthony A. Braga. 2015. "Better Policing Can Improve Legitimacy and Reduce Mass Incarceration," Harvard Law Review Forum 129, 233–41.
9. Boggess and Hipp 2010. See also Hipp, Tita, and Greenbaum.
10. James M. Anderson, John M. MacDonald, Ricky Bluthenthal, and J. Scott Ashwood. 2013. "Reducing Crime by Shaping the Built Environment with Zoning: An Empirical Study of Los Angeles," University of Pennsylvania Law Review 161, 699–756.
11. Craig A. Perkins, 1997. "Bureau of Justice Statistics Special Report: Age Patterns of Victims of Serious Violent Crime," in U.S. Department of Justice.

Appendices

Appendix A - Step 1: Exploratory Data Analysis

```
# Structure - Get the data

# Load Libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(lmtest)
library(rms)
library(MASS)
library(randomForest)
library(knitr)
library(kableExtra)

# Load data
community_crime <- read.table("CommViolPredUnnormalizedData.txt", sep = ",")
glimpse(community_crime)

# Renames all the column names from the table data to the appropriate
# corresponded data info provided from UCL Machine Learning Repository
colnames(community_crime) <- c("communityName", "state", "countyCode", "communityCode",
    "fold", "population", "householdSize", "racePctBlack", "racePctWhite", "racePctAsian",
    "racePctHisp", "agePct12t21", "agePct12t29", "agePct16t24", "agePct65up", "numUrban",
    "pctUrban", "medIncome", "pctWage", "pctFarmSelf", "pctInvInc", "pctSocSec",
    "pctPubAsst", "pctRetire", "medFamInc", "perCapInc", "perCapWhite", "perCapBlack",
    "perCapIndian", "perCapAsian", "perCapOther", "perCapHisp", "numUnderPov", "pctPopUnderPov",
    "pctLess9thGrade", "pctNotHSGrad", "pctBSorMore", "pctUnemployed", "pctEmpl",
    "pctEmplManu", "pctEmplProfServ", "pctOccupManu", "pctOccupMgmtProf", "pctDivMale",
    "pctNevMarrMale", "pctDivFemale", "totalPctDiv", "persPerFam", "pctFam2Par",
    "pctKids2Par", "pctYoungKids2Par", "pctTeen2Par", "pctWkMomYoungKids", "pctWkMom",
    "numKidsBornNevMarr", "pctKidsBornNevMarr", "numImmig", "pctImmig3", "pctImmig5",
    "pctImmig8", "pctImmig10", "pctPopImmig3", "pctPopImmig5", "pctPopImmig8", "pctPopImmig10",
    "pctSpeakEngOnly", "pctNotSpeakEngWell", "pctLargHouseFam", "pctLargHouseOccup",
    "persPerHouse", "persPerOwnHouse", "persPerRentHouse", "pctPersOwn", "pctPersDenseHouse",
    "pctHouseLess3BR", "medNumBR", "houseVacant", "pctHouse", "pctOwnHouse", "pctVacantBoarded",
    "pctVacMore6M", "medYrHouseBuilt", "pctHouseNoPhone", "pctWOFullPlumb", "ownLowQ",
    "ownMedVal", "ownHiQ", "ownQrange", "rentLowQ", "rentMed", "rentHiQ", "rentQrange",
    "medRent", "medRentPctHouseInc", "medOwnCostPctInc", "medOwnCostPctIncWOMtg",
    "numInShelters", "numStreet", "pctForeignBorn", "pctBornSameState", "pctSameHouse85",
    "pctSameCity85", "pctSameState85", "lemasSwFT", "lemasSwFTPerPop", "lemasSwFTFieldOps",
    "lemasSwFTFieldPerPop", "lemasTotalReq", "lemasTotalReqPerPop", "policReqPerOffic",
    "policPerPop", "racialMatchCommPol", "pctPolicWhite", "pctPolicBlack", "pctPolicHisp",
    "pctPolicAsian", "pctPolicMinor", "officAssgnDrugUnits", "numKindsDrugsSeiz",
    "policAvewOTWk", "landArea", "popDensity", "pctUsePubTrans", "policCars", "policOperBudg",
    "lemasPctPolicOnPatr", "lemasGangUnitDeploy", "lemasPctOfficDrugUn", "policBudgPerPop",
    "murders", "murdPerPop", "rapes", "rapesPerPop", "robberies", "robbPerPop", "assaults",
    "assaultPerPop", "burglaries", "burglPerPop", "larcenies", "larcPerPop", "autoTheft",
    "autoTheftPerPop", "arsons", "arsonsPerPop", "violentCrimesPerPop", "nonViolPerPop")

# Creates vectors of the names of variables (by labeling them to their
# appropriate data type)
```

```r
names_num <- c("countyCode", "communityCode", "policOperBudg")
names_int <- c("perCapInc", "perCapWhite", "perCapBlack", "perCapIndian", "perCapAsian",
    "perCapOther", "perCapHisp", "lemasSwFT", "lemasSwFTFieldOps", "lemasTotalReq",
    "officAssgnDrugUnits", "numKindsDrugsSeiz", "policCars", "lemasGangUnitDeploy",
    "rapes", "robberies", "assaults", "burglaries", "larcenies", "autoTheft", "arsons")
names_dbl <- c("medNumBR", "ownLowQ", "ownMedVal", "ownHiQ", "ownQrange", "rentLowQ",
    "rentMed", "rentHiQ", "rentQrange", "medRent", "lemasSwFTPerPop", "lemasSwFTFieldPerPop",
    "lemasTotalReqPerPop", "policReqPerOffic", "policPerPop", "racialMatchCommPol",
    "pctPolicWhite", "pctPolicBlack", "pctPolicHisp", "pctPolicAsian", "pctPolicMinor",
    "policAvewOTWk", "lemasPctPolicOnPatr", "policBudgPerPop", "rapesPerPop", "robbPerPop",
    "assaultPerPop", "burglPerPop", "larcPerPop", "autoTheftPerPop", "arsonsPerPop",
    "violentCrimesPerPop", "nonViolPerPop")

# Change the appropriate data types
community_crime[names_num] <- lapply(community_crime[names_num], as.numeric)
community_crime[names_int] <- lapply(community_crime[names_int], as.integer)
community_crime[names_dbl] <- lapply(community_crime[names_dbl], as.double)

# Identifies all of the variable's class/ data type
sapply(community_crime, "class")

# Accounts for all the missing values for each variable in the data.frame
na_count <- sapply(community_crime, function(y) sum(length(which(is.na(y)))))
data.frame(na_count)

# Select desirable predictor variables and the response for the predictive
# model along with the removal of any NA values
community_crime_tidy <- community_crime %>%
    dplyr::select(violentCrimesPerPop, racePctBlack, racePctHisp, agePct12t29, agePct65up,
        totalPctDiv, pctPersDenseHouse, medIncome, medRentPctHouseInc, pctPopUnderPov,
        pctVacantBoarded, pctBornSameState, pctForeignBorn, popDensity, landArea) %>%
    na.omit()

# Final data.frame dimensions
dim(community_crime_tidy)

# Save data as a file write.csv(community_crime_tidy,
# 'community_crime_tidy.csv')


# Explore the data - All analysis were performed on the following 4 predictor
# variables: `racePctBlack`,`medRentPctHouseInc`, `pctPopUnderPov`, and
# `pctVacantBoarded`.

# Univariate Relationships: - Histograms for each of the following predictor
# variables:

# Distribution of % of African-Americans in the population
ggplot(community_crime_tidy, aes(x = racePctBlack)) + geom_histogram(color = "black",
    fill = "white", bins = 30) + ylim(0, 425) + xlim(0, 100) + labs(title = "Histrogram of the Percentag
    x = "% of Blacks", caption = "Figure 1.1") + theme(plot.title = element_text(size = 15,
    hjust = 0), axis.title.x = element_text(size = 12, vjust = -1), axis.title.y = element_text(size =
    vjust = 1), axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 10),
```

```
    plot.caption = element_text(color = "black", face = "bold", size = 10, hjust = 0.5,
        vjust = -1))
```

```
# Summary Output for % of African-Americans in the population
summary(community_crime_tidy$racePctBlack)

# Distribution of Median household Income
ggplot(community_crime_tidy, aes(x = medIncome)) + geom_histogram(color = "black",
    fill = "white", bins = 30) + ylim(0, 325) + labs(title = "Histrogram of the Median Household Income
    x = "median household income (U.S. dollar, $)", caption = "Figure 1.4") + theme(plot.title = elemen
    hjust = 0), axis.title.x = element_text(size = 12, vjust = -1), axis.title.y = element_text(size =
    vjust = 1), axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 10),
    plot.caption = element_text(color = "black", face = "bold", size = 10, hjust = 0.5,
        vjust = -1))
```

```
# Summary Output for the median house income
summary(community_crime_tidy$medIncome)

# Distribution of % individuals below poverty level
ggplot(community_crime_tidy, aes(x = pctPopUnderPov)) + geom_histogram(color = "black",
    fill = "white", bins = 30) + ylim(0, 300) + xlim(0, 50) + labs(title = "Histrogram of the Percentag
    x = "% of individuals below poverty level", caption = "Figure 1.?") + theme(plot.title = element_te
    hjust = 0), axis.title.x = element_text(size = 12, vjust = -1), axis.title.y = element_text(size =
    vjust = 1), axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 10),
    plot.caption = element_text(color = "black", face = "bold", size = 10, hjust = 0.5,
        vjust = -1))
```

```
# Summary Output for % of individuals below poverty level
summary(community_crime_tidy$pctPopUnderPov)

# Distribution of % of vacant housing boarded up
ggplot(community_crime_tidy, aes(x = pctVacantBoarded)) + geom_histogram(color = "black",
    fill = "white", bins = 30) + ylim(0, 200) + xlim(0, 40) + labs(title = "Histrogram of the Percentag
    x = "% of vacant houses boarded up", caption = "Figure 1.?") + theme(plot.title = element_text(size
    hjust = 0), axis.title.x = element_text(size = 12, vjust = -1), axis.title.y = element_text(size =
    vjust = 1), axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 10),
    plot.caption = element_text(color = "black", face = "bold", size = 10, hjust = 0.5,
        vjust = -1))
```

```
# Summary Output for % of vacant housing boarded up
summary(community_crime_tidy$pctVacantBoarded)

# Bivariate Relationships - Scatter plots for each predictor against
# `violentCrimesPerPop`:

# % of African-Americans in the population vs. violent crimes committed per
# population
ggplot(community_crime_tidy, aes(y = violentCrimesPerPop, x = racePctBlack)) + geom_point(alpha = 0.3) +
    ylim(0, 5000) + xlim(0, 100) + labs(title = "Scatter plot of Percentage of Blacks vs. Violent Crime
    x = "% of blacks", y = "Violent Crime Per Population", caption = "Figure 2.1") +
    theme(plot.title = element_text(size = 15, hjust = 0), axis.title.x = element_text(size = 12,
        vjust = -1), axis.title.y = element_text(size = 12, vjust = 1), axis.text.x = element_text(size
        axis.text.y = element_text(size = 10), plot.caption = element_text(color = "black",
            face = "bold", size = 10, hjust = 0.5, vjust = -1))
```

```r
# Median houeshold Income vs. violent crimes committed per population
ggplot(community_crime_tidy, aes(y = violentCrimesPerPop, x = medIncome)) + geom_point(alpha = 0.25) +
    labs(title = "Scatter plot of Median Household Income vs. Violent Crime Per \n Population in the U.S
        x = "median household income (U.S. dollar, $)", y = "Violent Crime Per Population",
        caption = "Figure 2.4") + theme(plot.title = element_text(size = 15, hjust = 0),
    axis.title.x = element_text(size = 12, vjust = -1), axis.title.y = element_text(size = 12,
        vjust = 1), axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 10),
    plot.caption = element_text(color = "black", face = "bold", size = 10, hjust = 0.5,
        vjust = -1))
```

```r
# % of Individuals Below Poverty Level vs. violent crimes committed per
# population
ggplot(community_crime_tidy, aes(y = violentCrimesPerPop, x = pctPopUnderPov)) +
    geom_point(alpha = 0.25) + ylim(0, 5000) + xlim(0, 50) + labs(title = "Scatter plot of the Percentag
    x = "% of individuals below poverty level", y = "Violent Crime Per Population",
    caption = "Figure 2.?") + theme(plot.title = element_text(size = 15, hjust = 0),
    axis.title.x = element_text(size = 12, vjust = -1), axis.title.y = element_text(size = 12,
        vjust = 1), axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 10),
    plot.caption = element_text(color = "black", face = "bold", size = 10, hjust = 0.5,
        vjust = -1))
```

```r
# % of Vacant Housing Boarded Up vs. violent crimes committed per population
ggplot(community_crime_tidy, aes(y = violentCrimesPerPop, x = pctVacantBoarded)) +
    geom_point(alpha = 0.25) + ylim(0, 5000) + xlim(0, 40) + labs(title = "Scatter plot of the Percentag
    x = "% of vacant houses boarded up", y = "Violent Crime Per Population", caption = "Figure 2.?") +
    theme(plot.title = element_text(size = 15, hjust = 0), axis.title.x = element_text(size = 12,
        vjust = -1), axis.title.y = element_text(size = 12, vjust = 1), axis.text.x = element_text(size
        axis.text.y = element_text(size = 10), plot.caption = element_text(color = "black",
            face = "bold", size = 10, hjust = 0.5, vjust = -1))
```

Appendix B - Step 2: Fit a Linear Model

```
# Fit a Linear Model
step2_model <- lm(violentCrimesPerPop ~ racePctBlack + racePctHisp + agePct12t29 +
    agePct65up + totalPctDiv + pctPersDenseHouse + medIncome + medRentPctHouseInc +
    pctPopUnderPov + pctVacantBoarded + pctBornSameState + pctForeignBorn + popDensity +
    landArea, data = community_crime_tidy)
# Summary Output for model
summary(step2_model)

# Check for Multicollinearity:
i <- 0  # initializing
for (i in seq_along(vif(step2_model))) {
    if (vif(step2_model)[i] > 10) {
        print(paste("Multicollinearity is present in the following: ", (vif(step2_model)[i] >
            10), ".", sep = ""))
    } else {
        print(paste("Multicollinearty isn't present."))
    }
}
```

Appendix C - Step 3: Perform Model Selection

```
# Fit a Linear Model as an OLS (used for backward selection)
step2_model_ols <- ols(violentCrimesPerPop ~ racePctBlack + racePctHisp + agePct12t29 +
    agePct65up + totalPctDiv + pctPersDenseHouse + medIncome + medRentPctHouseInc +
    pctPopUnderPov + pctVacantBoarded + pctBornSameState + pctForeignBorn + popDensity +
    landArea, data = community_crime_tidy)

# Perform Backward (p-value based) Model Selection
fastbw(step2_model_ols, rule = "p", sls = 0.05)

# Perform Akaike Information Criteria model selection
stepAIC(step2_model)

# Model Used for Data Analysis
step3_model <- lm((violentCrimesPerPop + 1) ~ racePctBlack + agePct65up + totalPctDiv +
    pctPersDenseHouse + medRentPctHouseInc + pctPopUnderPov + pctVacantBoarded +
    pctForeignBorn + landArea, data = community_crime_tidy)
# Summary Output for the model
summary(step3_model)
```

Appendix D - Step 4: Perform Model Diagnostics

```
# Check assumption of constant error variance using diagnostic plot
plot(step3_model$fitted.values, step3_model$residuals, main = "Diagnostic Plot",
    xlab = "fitted-values", ylab = "residuals", sub = substitute(paste(bold("Figure 4.4.1"))),
    cex.sub = 0.9)
```

```
# using hypothesis testing: Breusch-Pagan (BP) Test
bptest(step3_model)

# Check assumption of normal errors using QQ plot
qqnorm(step3_model$residuals)  # using QQplot
qqline(step3_model$residuals, col = "red")  # add qqline to plot
```

```
# using model's histogram
hist(step3_model$residuals, breaks = 30, main = "Histogram of Model Residauals",
    xlab = "Residuals", sub = substitute(paste(bold("Figure 4.4.3"))), col.sub = "black",
    cex.sub = 0.9)
```

```
# using hypothesis testing: Shapiro-Walk Test
shapiro.test(step3_model$residuals)

# Check assumption of independent/ uncorrelated errors
n <- dim(community_crime_tidy)[1]  # total sampling-size
par(mfrow = c(1, 2))
# using lagged residual plot
plot(main = "Lagged Residual Plots", residuals(step3_model) ~ fitted.values(step3_model),
    na.omit(community_crime_tidy), ylab = "residuals", xlab = "fitted-values", sub = substitute(paste(b
    col.sub = "black", cex.sub = 0.9)
abline(h = 0)
plot(tail(residuals(step3_model), n - 1) ~ head(residuals(step3_model), n - 1), xlab = expression(hat(e
    ylab = expression(hat(epsilon)[i + 1]), sub = substitute(paste(bold("(b)"))),
    col.sub = "black", cex.sub = 0.9)
abline(h = 0, v = 0, col = grey(0.75))
```

```
# using hypothesis testing: Durbin-Watson Test
dwtest(step3_model)
```

Appendix E - Step 5: Investigate Fit for Individual Observations

```
# Check for outliers: Calculates the absolute standardized residuals as a
# summary output.
summary(abs(rstandard(step3_model)))
# Identifies the observations that are outliers
if (which(abs(rstandard(step3_model)) > 3)) {
    print(paste("Based on the rule of thumb, there are ", table(abs(rstandard(step3_model)) >
        3)["TRUE"], " outliers.", sep = ""))
} else {
    print("Based on the rule of thumb, there are no outliers present.")
}

# Check leverages: Computes Average Leverage
mean(hatvalues(step3_model))
# High-Leverage points: Identifies high leverage points
which(2 * mean(hatvalues(step3_model)) <= hatvalues(step3_model))
# Counts observations with high leverage
table(2 * mean(hatvalues(step3_model)) <= hatvalues(step3_model))["TRUE"]

# Counts the relations of high-leverages against outliers
table(2 * mean(hatvalues(step3_model)) <= hatvalues(step3_model), abs(rstandard(step3_model)) >
    3, dnn = c("high-leverage", "outlier"))


# Check for influential points: Find F distr. threshold
n <- dim(model.matrix(step3_model))[1]  # total number of sample-size
p3 <- dim(model.matrix(step3_model))[2]  # total number of parameters for step3_model
num.df <- p3  # degrees of freedom for the numerator
den.df <- n - p3  # degreses of freedom for the denominator
(F.thresh <- qf(0.5, p3, n - p3))  #computes F distribution threshold
# Compute Cook's distances
cooks_dist <- cooks.distance(step3_model)
# Prints Cook's distance summary
summary(cooks_dist)
# Identifies the observations that are influential
if (cooks_dist > F.thresh) {
    print(paste("Based on the rule of thumb, theses are the following disproportionately large influenti
        which(cooks_dist > F.thresh), ".", sep = "")
} else {
    print("Based on the rule of thumb, there are no disproportionately large influential observations in
}
```

Appendix F - Step 6: Apply Transformations to the Model

```
# Apply the box-cox method to the model
bc <- boxcox(step3_model, plotit = T)
```

```
# Optimal value of lambda
(lambda <- bc$x[which.max(bc$y)])

# Fitted new model using transformation recommended by the box-cox method
step6_model <- lm((violentCrimesPerPop + 1)^lambda ~ racePctBlack + agePct65up +
    totalPctDiv + pctPersDenseHouse + medRentPctHouseInc + pctPopUnderPov + pctVacantBoarded +
    pctForeignBorn + landArea, data = community_crime_tidy)
# Summary Output for transformed model
summary(step6_model)

# Re-check Model Diagnostics for new model:
par(mfrow = c(2, 2))
# Diagnostic plot - checks for constant variance assumption
plot(step6_model$fitted.values, step6_model$residuals, main = "Diagnostic plot after Transformation",
    xlab = "fitted-values", ylab = "residuals", sub = substitute(paste(bold("Figure 4.6.2 (a)"))),
    col.sub = "black", cex.sub = 0.9)
# Q-Q plot - checks for normal error assumption
qqnorm(step6_model$residuals, main = "Normal Q-Q Plot after Transformation", sub = substitute(paste(bol
    col.sub = "black", cex.sub = 0.9)
# lagged residual plot - checks for independent/ uncorrelated error assumption
plot(main = "Lagged Residual Plots after Transformation", residuals(step6_model) ~
    fitted.values(step6_model), na.omit(community_crime_tidy), ylab = "residuals",
    xlab = "fitted-values", sub = substitute(paste(bold("(c)"))), col.sub = "black",
    cex.sub = 0.9)
abline(h = 0)
plot(tail(residuals(step6_model), n - 1) ~ head(residuals(step6_model), n - 1), xlab = expression(hat(e
    ylab = expression(hat(epsilon)[i + 1]), sub = substitute(paste(bold("(d)"))),
    col.sub = "black", cex.sub = 0.9)
abline(h = 0, v = 0, col = grey(0.75))
```

Appendix G - Step 7: Report Inferences and Make Predictions using the Final Model

```
# Creates vector of predictor names in order of the summary output
predictor_names <- c("(Intercept)", "racePctBlack", "agePct65up", "totalPctDiv",
    "pctPersDenseHouse", "medRentPctHouseInc", "pctPopUnderPov", "pctVacantBoarded",
    "pctForeignBorn", "landArea")
# Create data.frame table of parameter's estimates and pvalues
table_info <- as.data.frame(matrix(c(predictor_names, step6_model$coefficients, summary(step6_model)$coe
    4]), ncol = 3, dimnames = list(c(), c("Parameter(s)", "Estimate(s)", "p-value(s)"))))
# Displays figure into a table formatt
table_info %>%
    kable(caption = "Figure 4.7.1", digits = 4) %>%
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F)

# Reports R-Squared value
summary(step6_model)$r.squared

# Data.frame containing only the variables used in the final model
community_crime_tidy_modeled <- community_crime_tidy %>%
    dplyr::select(violentCrimesPerPop, racePctBlack, agePct65up, totalPctDiv, pctPersDenseHouse,
        medRentPctHouseInc, pctPopUnderPov, pctVacantBoarded, pctForeignBorn, landArea)

# Determines the most importance of the variables from the original model
# (before transformation) Creates a random forest model
regressor_orig <- randomForest(violentCrimesPerPop ~ ., community_crime_tidy_modeled,
    importance = TRUE)
# produces importance plot
varImpPlot(regressor_orig)
```

```
# produces importance table
importance(regressor_orig)

# Determines the most importance of the variables from the final model (after
# transformation) Creates a random forest model
regressor_final <- randomForest(violentCrimesPerPop^lambda ~ ., community_crime_tidy_modeled,
    importance = TRUE)
# produces importance plot
varImpPlot(regressor_final)
```

```
# produces importance table
importance(regressor_final)

# alpha level
alpha <- 0.05
# number of parameters for step6_model
p6 <- length(step6_model$coefficients)
# slope estimate
x_bar <- step6_model$coefficients["racePctBlack"]
# standard error of slope estimate
se <- summary(step6_model)$coefficients[, 2]["racePctBlack"]
# Z score
z <- qt(p = alpha/2, df = n - p6, lower.tail = F)

# Compute and Report 95% CI for the slope of that the most important variable
```

```r
important_ci <- x_bar + c(-1, 1) * z * se/sqrt(n)
# Inverse transformation values converted back to original units of the
# response variable
important_ci^(1/lambda)

# Define new_df
new_df <- data_frame(racePctBlack = median(community_crime_tidy$racePctBlack), agePct65up = median(commu
    totalPctDiv = median(community_crime_tidy$totalPctDiv), pctPersDenseHouse = median(community_crime_t
    medRentPctHouseInc = median(community_crime_tidy$medRentPctHouseInc), pctVacantBoarded = median(comm
    pctForeignBorn = median(community_crime_tidy$pctForeignBorn), pctPopUnderPov = median(community_cri
    landArea = median(community_crime_tidy$landArea))

# 95% CI for a prediction of the true mean
(confidence_ci <- predict(step6_model, newdata = new_df, interval = "confidence",
    level = 0.95))
# Inverse transformation values converted back to original units of the
# response variable
c(confidence_ci[2]^(1/lambda), confidence_ci[3]^(1/lambda))

# 95% CI for a prediction of a particular observation performs 95% CI for
# individual values
(prediction_ci <- predict(step6_model, newdata = new_df, interval = "prediction",
    level = 0.95))
# Inverse transformation values converted back to original units of the
# response variable
c(prediction_ci[2]^(1/lambda), prediction_ci[3]^(1/lambda))
```