UNIVERSITY OF NOTRE DAME | DATA SCIENCE

Linear Model Analysis of Violent Crimes in USA Communities

Linear Models Final Project Report
Tao Tang

Index

1.Introduction

Understanding where violent crime happens can be a key to understanding why it happens. Environmental, Social and population characteristics may be important predictors of the level of violent crime in a population. Determining which are most influential on the level of violent crime will provide valuable input to neighborhood design, urban development, and policing practices. In this project, we did choose four variables we feel the most important to violent crimes, and then used linear modeling concepts and knowledges to build regression model and report on the model results and diagnostics.

2.Dataset Description

This project utilized the USA Communities and Crimes Dataset, sourced from the UCI Dataset Repository. The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units. The crime attributes (N=18) that could be predicted are the 8 crimes considered 'Index Crimes' by the FBI)(Murders, Rape, Robbery, ...), per capita (actually per 100,000 population) versions of each, and Per Capita Violent Crimes and Per Capita Nonviolent Crimes). The per capita crimes variables were calculated using population values included in the 1995 FBI data (which differ from the 1990 Census values). The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault.

In this project, we used ViolentCrimesPerPop as response variable, and the following variables as predictors:

PctKidsBornNeverMar: percentage of kids born to never married

PctPersDenseHous: percent of persons in dense housing (more than 1 person per room)

racePctWhite: percentage of population that is Caucasian

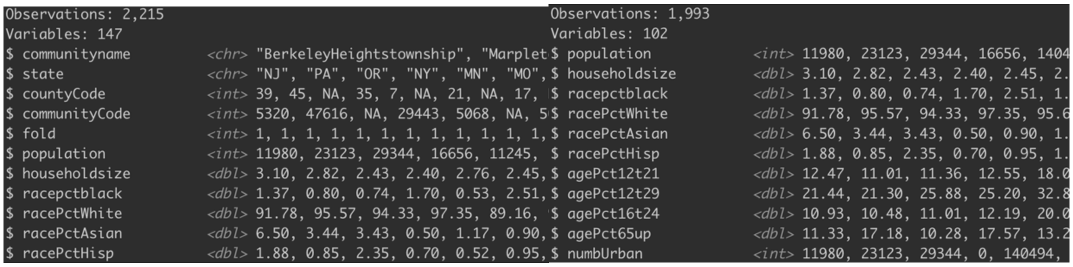
racepctblack: percentage of population that is African American

3.Data Cleaning

It can be seen that there are some variables in the dataset which will not contribute to the accurate prediction of violent crime levels. These include Community Identifies: State, County, Community and Community name which can be removed from the feature dataset.

Analyzing the dataset for missing values or null data in the dataset showed that while there are no 'na' values in the dataset, missing values appear in the form of "?". An analysis of the columns with missing values shows there are 25 columns with missing values which describe mostly community identifiers and policing information. Most of these have a large proportion of the data missing (> 50%) which renders these variables unusable. Therefore, those columns with large proportions of missing data are removed from the dataset. There are also missing values for the response variable "ViolentCrimesPerPop". The samples with missing data for the response variable "ViolentCrimesPerPop" are removed from the dataset.

In Figure 3.3, before the cleaning process, the data contains 2215 observations and 147 variables. After the cleaning, the data contains 1998 observations and 102 variables.



```
Observations: 2,215
Variables: 147
$ communityname          <chr> "BerkeleyHeightstownship", "Marplet
$ state                  <chr> "NJ", "PA", "OR", "NY", "MN", "MO",
$ countyCode             <int> 39, 45, NA, 35, 7, NA, 21, NA, 17,
$ communityCode          <int> 5320, 47616, NA, 29443, 5068, NA, 5
$ fold                   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ population             <int> 11980, 23123, 29344, 16656, 11245,
$ householdsize          <dbl> 3.10, 2.82, 2.43, 2.40, 2.76, 2.45,
$ racepctblack           <dbl> 1.37, 0.80, 0.74, 1.70, 0.53, 2.51,
$ racePctWhite           <dbl> 91.78, 95.57, 94.33, 97.35, 89.16,
$ racePctAsian           <dbl> 6.50, 3.44, 3.43, 0.50, 1.17, 0.90,
$ racePctHisp            <dbl> 1.88, 0.85, 2.35, 0.70, 0.52, 0.95,
```

```
Observations: 1,993
Variables: 102
$ population             <int> 11980, 23123, 29344, 16656, 1404
$ householdsize          <dbl> 3.10, 2.82, 2.43, 2.40, 2.45, 2.
$ racepctblack           <dbl> 1.37, 0.80, 0.74, 1.70, 2.51, 1.
$ racePctWhite           <dbl> 91.78, 95.57, 94.33, 97.35, 95.6
$ racePctAsian           <dbl> 6.50, 3.44, 3.43, 0.50, 0.90, 1.
$ racePctHisp            <dbl> 1.88, 0.85, 2.35, 0.70, 0.95, 1.
$ agePct12t21            <dbl> 12.47, 11.01, 11.36, 12.55, 18.0
$ agePct12t29            <dbl> 21.44, 21.30, 25.88, 25.20, 32.8
$ agePct16t24            <dbl> 10.93, 10.48, 11.01, 12.19, 20.0
$ agePct65up             <dbl> 11.33, 17.18, 10.28, 17.57, 13.2
$ numbUrban              <int> 11980, 23123, 29344, 0, 140494,
```

Figure 3.3 Data glimpse before and after cleaning

4.Exploratory Data Analysis (EDA)

Exploring the predictor variable ViolentCrimesPerPop. In Figure 4.1, showed that it has a continuous distribution which is left skewed. The value range is from 0 to 4877.

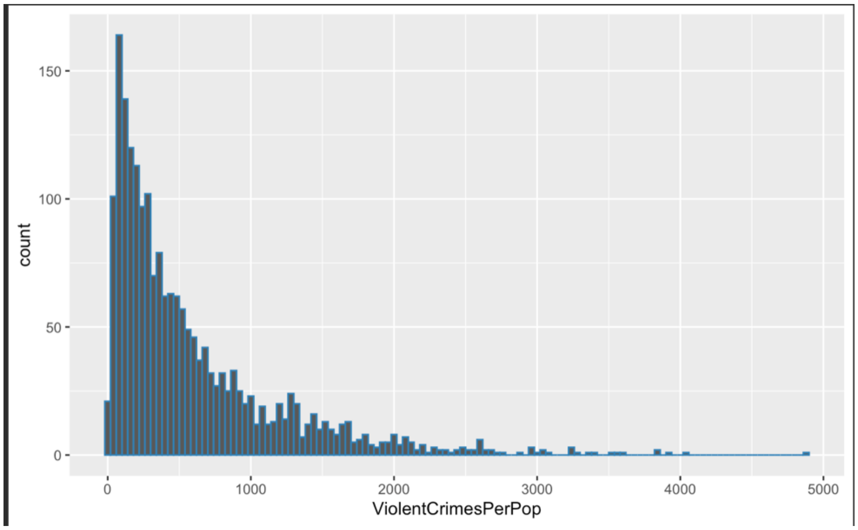


Figure 4.1 Distribution of ViolentCrimesPerPop

Next, we chose four (4) predictors we feel are the most important violent crimes and produced scatter plots of the response variable vs. these predictors. In Figure 4.2, we can see that the percentage of kids born to never married is positively correlated to ViolentCrimesPerPop. The percent of persons in dense housing (more than 1 person per room) is positively correlated to ViolentCrimesPerPop. The percentage of population that is Caucasian is negatively correlated to ViolentCrimesPerPop. The percentage of population that is african american is positively correlated to ViolentCrimesPerPop.

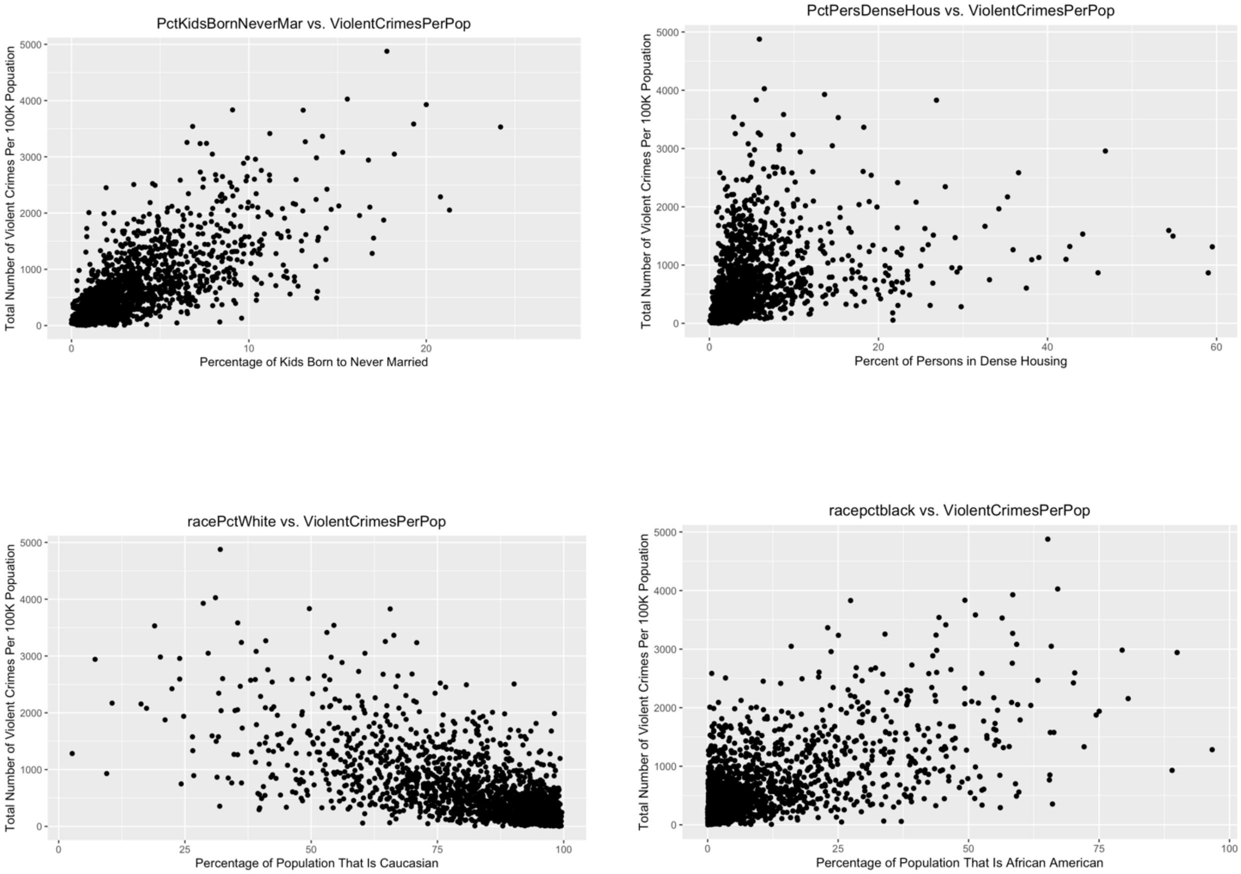
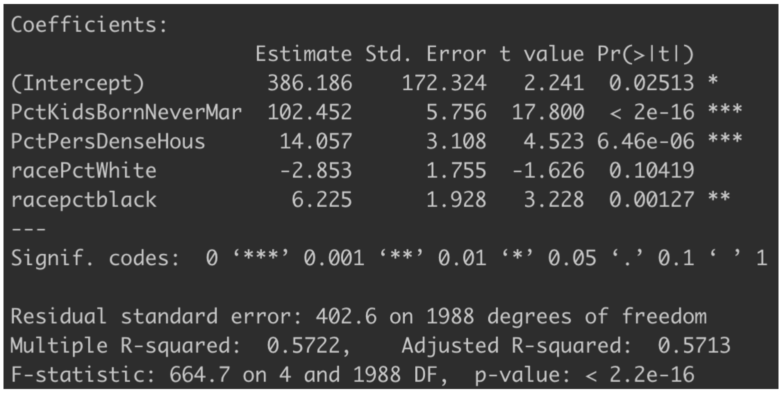


Figure 4.2 response variable vs. the four predictors

5. Fit a Linear Model

Here we fit a linear model with ViolentCrimesPerPop as the response and the four variables chosen in previously as the predictors. In Figure 5.1, the model suggested a positive association between the ViolentCrimesPerPop and PctKidsBornNeverMar, PctPersDenseHous,racepctblack; and a negative association between the ViolentCrimesPerPop and racePctWhite. Based on the output, the ViolentCrimesPerPop (total number of violent crimes per 100K popuation) without considering any variables is 387.429, on average. For every one-unit increase in racePctWhite, we expect the ViolentCrimesPerPop to descrease 2.861, on average. Approximately 57% of the variation in ViolentCrimesPerPop is explained by the predictors. In addition, the p-value of F-statistics is small enough to tell us that at least one of the predictors is significant important.



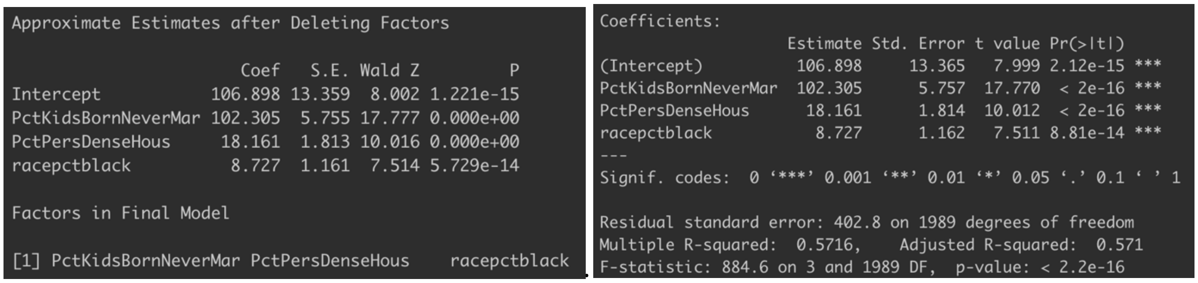
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          386.186    172.324   2.241  0.02513 *  
PctKidsBornNeverMar  102.452      5.756  17.800  < 2e-16 ***
PctPersDenseHous      14.057      3.108   4.523 6.46e-06 ***
racePctWhite          -2.853      1.755  -1.626  0.10419    
racepctblack           6.225      1.928   3.228  0.00127 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 402.6 on 1988 degrees of freedom
Multiple R-squared:  0.5722,	Adjusted R-squared:  0.5713 
F-statistic: 664.7 on 4 and 1988 DF,  p-value: < 2.2e-16
```

Figure 5.1

6. Perform Model Selection

In this step, we performed model selection via automated backward selection and Akaike Information Criterion methods. The selection is based on the four predictors we chose previously.

First, we used automated backward selection. In Figure 6.1, the output suggested three predictors PctKidsBornNeverMar, PctPersDenseHous, and racepctblack in final model, with lower adjusted R-squared than the model built previously.

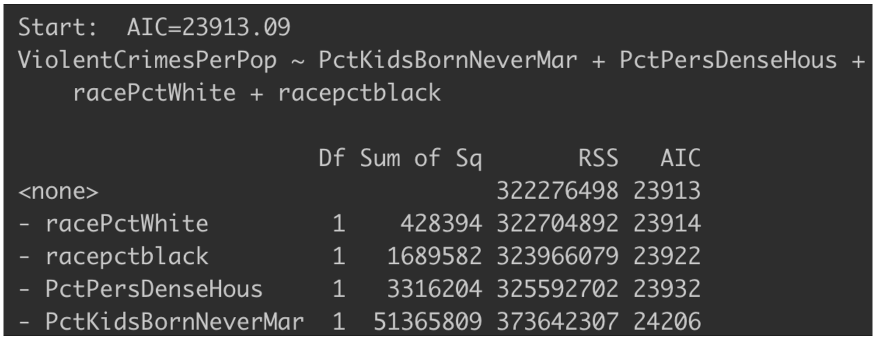


```
Approximate Estimates after Deleting Factors

                     Coef   S.E. Wald Z         P
Intercept         106.898 13.359  8.002 1.221e-15
PctKidsBornNeverMar 102.305  5.755 17.777 0.000e+00
PctPersDenseHous   18.161  1.813 10.016 0.000e+00
racepctblack        8.727  1.161  7.514 5.729e-14

Factors in Final Model

[1] PctKidsBornNeverMar PctPersDenseHous    racepctblack
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          106.898     13.365   7.999 2.12e-15 ***
PctKidsBornNeverMar  102.305      5.757  17.770  < 2e-16 ***
PctPersDenseHous      18.161      1.814  10.012  < 2e-16 ***
racepctblack           8.727      1.162   7.511 8.81e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 402.8 on 1989 degrees of freedom
Multiple R-squared:  0.5716,	Adjusted R-squared:  0.571 
F-statistic: 884.6 on 3 and 1989 DF,  p-value: < 2.2e-16
```

Figure 6.1 Output of automated backward selection

Next, we applied Akaike Information Criterion on the model we built previously. In Figure 6.2, the AIC suggested that same four predictors of the previous model.



```
Start:  AIC=23913.09
ViolentCrimesPerPop ~ PctKidsBornNeverMar + PctPersDenseHous + 
    racePctWhite + racepctblack

                      Df Sum of Sq       RSS   AIC
<none>                             322276498 23913
- racePctWhite         1    428394 322704892 23914
- racepctblack         1   1689582 323966079 23922
- PctPersDenseHous     1   3316204 325592702 23932
- PctKidsBornNeverMar  1  51365809 373642307 24206
```

Figure6.2 Output of automated backward selection

In conclusion, we simply decided to choose the model with the four predictors.

7. Apply Diagnostics to the Model

The estimation of and inference from the regression model depend on several assumptions. These assumptions should be checked using regression diagnostics before using the model in earnest. In this step, we check the independence, constant variance and normality of the errors. The errors are not observable, but we can examine the residuals. It is not possible to check the assumption of constant variance just by examining the residuals alone — some will be large and some will be small, but this proves nothing. We need to check whether the variance in the residuals is related to some other quantity. Here we are going to plot the fitted values and residuals. If all is well, you should see constant symmetrical variation (known as homoscedasticity) in the vertical direction. Nonconstant variance is also called heteroscedasticity. In Figure 7.1, there is a pattern on this plot, and thus the model assumption of constant error variance is not upheld.

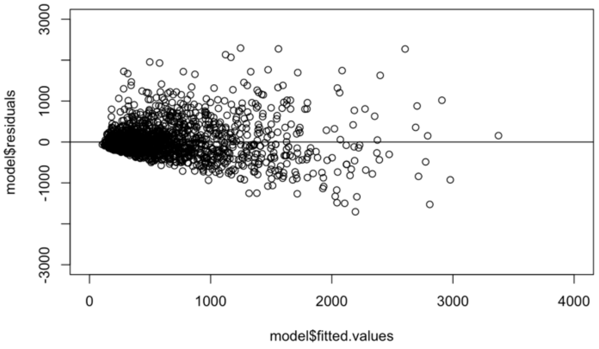


Figure 7.1 Fitted.value vs. Residuals

The tests and confidence intervals we used are based on the assumption of normal errors. The residuals can be assessed for normality using a *Q–Q* plot. In Figure 7.2, non-normality is found based on the plot. Thus the assumption of normal errors is not upheld.

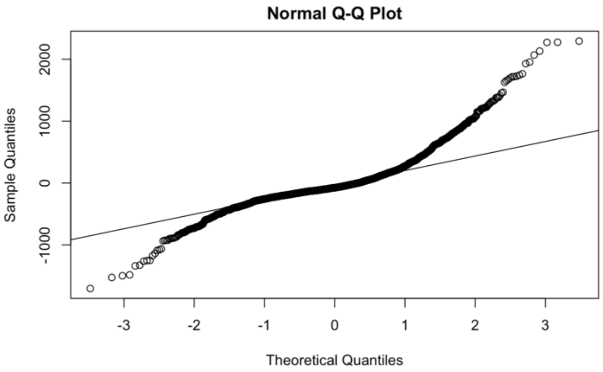


Figure 7.2 Q-Q plot

In Figure 7.3, the plot looks random. Therefore, based on the plot I would assume that the errors are uncorrelated, indicating the model assumption is upheld.

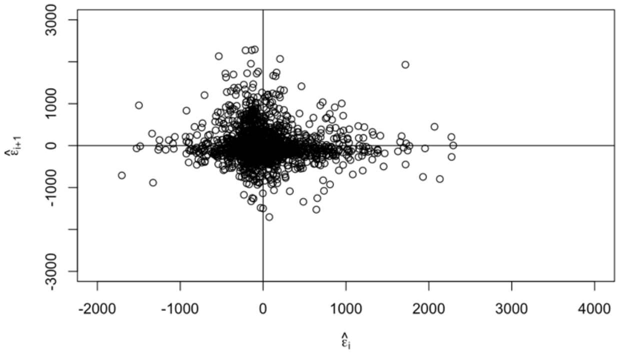


Figure 7.3 lagged residual plot

8. Investigate Fit for Individual Observations

We may find that some observations do not fit the model well. Such points are called outliers. Other observations change the fit of the model in a substantive manner. These are called influential observations. It is possible for a point to have both these characteristics. In this step, we used standardized residuals to detect outliers, and Cook's Distance to check influential points. The output indicates that there are 45 points that having an absolute value bigger than 3, so those observation are considered outliers by our rule of thumb threshold of three. However, none of them exceed that F threshold of 0.87. Therefore, we could say that there are no influential observations, and we do not need to remove those outliers.

9.Apply Transformations to the Model

Transformations of the response and/or predictors to improve the fit and correct violations of model assumptions were not met in step 4 of applying diagnostic to the model. The Box–Cox method is a popular way to determine a transformation on the response. It is designed for strictly positive responses and chooses the transformation to find the best fit to the data. Before applying the transformation, let's determine whether there is a quadratic relationship between the model's residuals and each predictor. In Figure 9.1, those graphics indicated that the four predictors do not have quadratic relationship with residuals, so we do not need to use polynomial regression in this case.

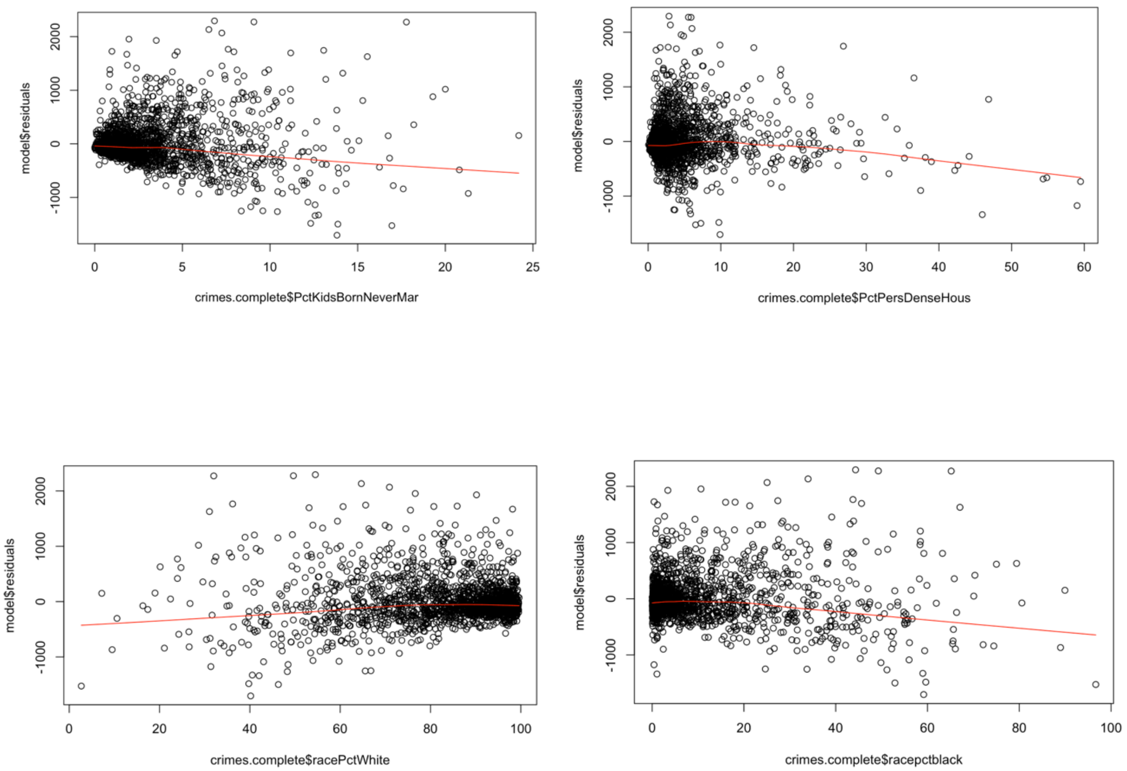


Figure 9.1 residuals vs the four predictors

In Figure 9.2, the Box-Cox analysis on the response variable ViolentCrimesPerPop of this model recommends λ = 0.263.

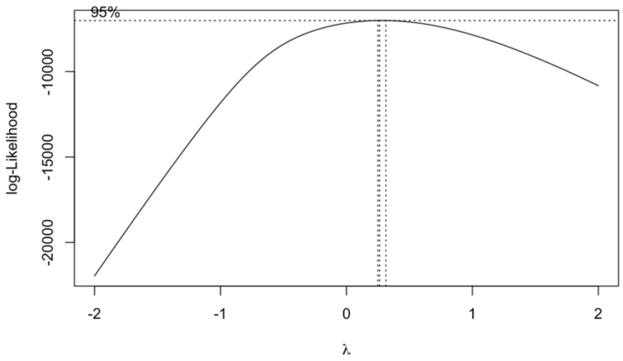


Figure 9.2

In the end, we fit a new model with the new transformed response variable (ViolentCrimesPerPop)^ λ. In Figure 9.3, those graphics showed the new model corrected violations of model assumptions were not met in step 4 after applying diagnostic to the initial model.

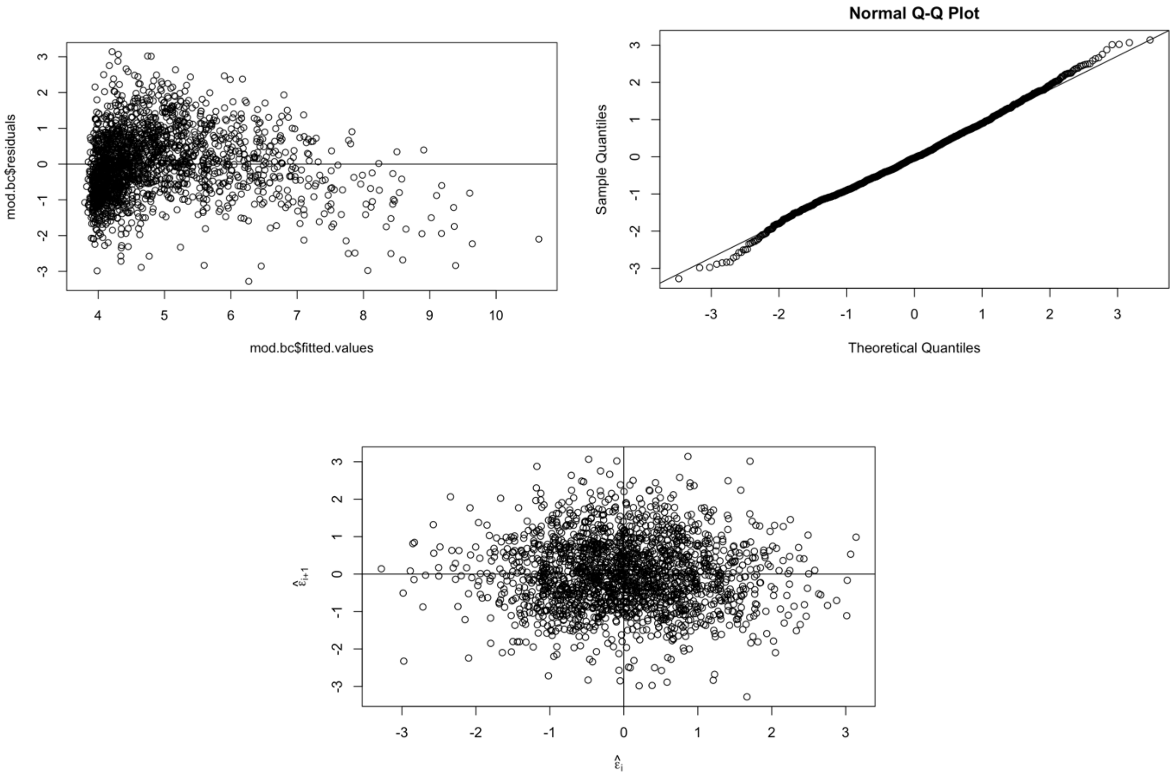
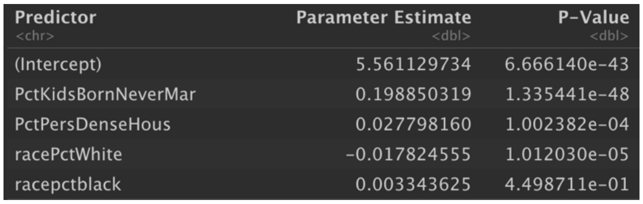


Figure 9.3 Diagnostics to the Final Model

10. Report Inferences and Make Predictions

In Figure 10.1, the parameter estimates and p-values for the final model are reported.



Predictor <chr>	Parameter Estimate <dbl>	P-Value <dbl>
(Intercept)	5.561129734	6.666140e-43
PctKidsBornNeverMar	0.198850319	1.335441e-48
PctPersDenseHous	0.027798160	1.002382e-04
racePctWhite	-0.017824555	1.012030e-05
racepctblack	0.003343625	4.498711e-01

Figure 10.1

Based on the final output, we are 95% confident the slope of PctKidsBornNeverMar lies in the interval (0.17295, 0.2247503). In addition, we are 95% confident the true mean ViolentCrimesPerPop at the median of each predictors lies in the interval (4.407291, 4.50599). Moreover, we are 95% confident the ViolentCrimesPerPop for the particular observation that at the median values of each predictor, which PctKidsBornNeverMar is 2.08, PctPersDenseHous is 2.47, racePctWhite is 89.64, and racepctblack is 3.14, lies in the interval (2.644148, 6.269132). As we can tell that the prediction interval is wider than confidence interval.

11.Conclusion

In this project, we only selected four predictors for our model. The P-Value of the variable racepctblack (percentage of population that is African American) is 0.45 which is not statistically significant and indicates that we cannot reject the null hypothesis which allows us to conclude that changes in the predictor racepctblack are not associated with changes in the response ViolentCrimesPerPop. There is some similarity between these features, which indicates actions to reduce correlated features should have been taken in the data preparation phase. However, the R-squared of the model is approximately 53%, which means 53% of variation in the response are explained by the four predictors. By using R-squared (with a scale of 0 to 1) as a measure of predictive performance, the model can be considered to have a moderately good performance on this dataset.