Predicting U.S Crime Rate Applying Multivariable Linear Regression

Jie Liu(jl5788), Yu Si (ys3507), Jiaqi Chen (jc5681), Keming Zhan(kz2383), Yiru Gong (yg2832)

Abstract

The increasing crime rate of the United States in the 21st century has triggered significant attention. In this project, we created a model to predict the crime rate based on multiple socioeconomic variables. We conducted initial visualization of the crime rate dataset with various plots, followed by model selection, model interaction, and cross-validation. With these methods, we selected an optimal model to predict the crime rate in counties. Subsequently, we used model diagnostics to determine whether the selected model needed further transformation before excluding existing outliers.

Introduction

The crime rate in the United States has been declining since 1990. This pattern continued until 2015 when crime rates started to gradually climb and increased significantly in 2020[1]. Between 2019 and 2020, the murder rate rose 30% in the United States based on the CDC database[2]. As the crime rate keeps growing, more people became concerned about this issue and require an accurate crime rate anticipation model. The fundamental goal of this project is to gain insights to anticipate the crime rate in different cities and states based on socioeconomic and demographic information. Among various factors that contribute to crime occurrence, the education and infrastructure level of the region potentially account for relevant crime rates[3]. Additionally, income level also exerts a potential influence on social stability [4].

Methods

Data import and clean We import the original dataset—-County Demographic Information and create a new variable CRM_1000(the crime rate per 1,000 population) as a target variable, and we transform the type of the variable region to factor in R.

Data exploratory analysis The aim of this part is to understand the distribution of each variable and the correlation of two variables and eliminate the effect of multicollinearity to satisfy the assumption of independence. In order to visually show this information, we constructed a summary table for all variables, a boxplot of the target variable, histograms of independent variables, and a correlation plot.

Model selection Aiming at selecting the most appropriate model, we applied forward, backward and stepwise selection for comparison. Stepwise selection is finally utilized based on the effectiveness of AIC value adjustment.

Model interaction Interactions occur when variables act together to impact the output of the process. Use an interaction plot to show how the relationship between one categorical factor and a continuous response depends on the value of the second categorical factor. On an interaction plot, parallel lines indicate that there is no interaction effect while different slopes suggest that interaction might be present. The more nonparallel the lines are, the greater the strength of the interaction. Based on our analysis, we formed the interaction between the categorical variable and continuous variable, in which region is the categorical variable. After forming an interaction and according to the model, we could use adjusted R^2 to compare models with different interaction terms and use ANOVA to confirm the large model is superior to the original model without interaction.

Cross validation We divided data into 5-fold and built models using 4/5 of data while testing models using the remaining repetitively. RMSE(root-mean-square-deviation) and R square of the finally established models were evaluated to select the optimal. The model with lower RMSE indicates a smaller deviation and a higher R square suggests better fitness of the model.

Model diagnostics In order to maximize the likelihood and stabilize the variance, we decided to use box-cox transformation to transform non-normal dependent variables into a normal shape. After framing the box-cox plot, we found an effective power transformation to achieve normal residuals. We built a new model with the transformed value, graphed QQ plots with the original model and the new model, and ultimately compared them to determine the optimal model. Then the outliers in the model are identified by plotting the residuals vs leverage plots and calculating the Cook’s distance, which considers the influence of the ith case on all fitted values. A threshold of Di > 1/110 is set to raise concern. Then, a model without outliers is established and compared to the original model to indicate the influence of outliers. The influential outliers will be eliminated in the final model. Finally, the multicollinearity is double-checked with the VIF value of each variable (without interaction).

Results

Data exploratory analysis From the two tables (Table 1 and 2), we saw the mean and standard deviation of 13 numeric variables and the frequency of the factor variable region. After plotting a boxplot to check outliers of the target variable (Figure 1), we noticed four counties' CRM_1000 are extremely high, whose names are Kings, Dade, Fulton, and St.Loui respectively. These four outliers should be considered in the subsequent analysis.

From 12 histograms of each numeric variable (Figure 2), we noticed that among them, the distribution of 5 histograms is right-skewed. Therefore, we performed a log transformation for these five variables(area, pop, docs, beds, and totalinc). The improved distribution is shown in figure 3. In order to check multicollinearity among independent variables, we construct a correlation plot (Figure 4) to help us intuitively find the strength of the correlation of independent variables.From the correlation plot,we can find independent variables like poverty(0.47), unemp(0.42), log_docs(0.443), log_beds(0.493) have a high correlation with the target variable CRM_1000. In addition, we also notice that independent variables like log_totalinc, log_beds, log_docs, and log_pop have a high correlation with each other. In order to verify this, we perform a stepwise variable selection according to a general rule of thumb for variance inflation factor(VIF): a value between 1 and 5 indicates a moderate correlation between a given predictor variable and other predictor variables in the model, and a value greater than 5 indicates a potentially severe correlation between a given predictor variable and other predictor variables in the model. At last, we remove four variables log_totalinc, log_docs, bagrad, and log_beds).

Model selection The model was constructed using stepwise selection and 8 variables (Table 3) are included in the model. R square is reported as 0.5579 and adjusted R square is 0.5497, suggesting acceptable fitness of the selected model.

Model interaction We made the interaction terms between the region and other variables and made the interaction plot. From the plot (Figure 5), we could find that the interaction terms between the region and pop18/pop65/log_pop were not significant so we removed them. The interaction between region and poverty/log_area are significant so we left the two models with these two interactions separately. Also, we were interested in the interaction between continuous variables and we kept the model with the interaction term between poverty and log_pop from others because of the larger adjusted r^2. After selecting three models, we used ANOVA to compare the original model and the other three adjusted models with interaction, in which the original model is nested in model 1/2/3. Based on the information, we could reject H0 and conclude that the larger model 1 and 3 are superior. We failed to reject H0 and conclude model 2 is not superior.

Cross validation Two established final models were cross-validated by folding the data 5-fold and testing the model using ⅕ of data. RMSE and R-squares of the two models were compared for evaluation. The model with lower RMSE and higher R-square was selected as optimal due to smaller deviation and better fitness(Table 5).

Model diagnostics We used box-cox transformation to compare whether the selected model needs to be transformed to make the data more “normal”. We graphed a box-cox plot with the selected model [graph 10] and figured that the λ of the distribution is close to 0.5. Therefore, we chose to take the square root of the response value, the crime rate per 1,000 population. By comparing the residual vs fitted plots of the new model and the original model, we saw that the interval of residuals is narrower in the original model (Figure 11) than in the transformed model (Figure 12). When comparing the QQ plots of both models, we found that the original model (Figure 13) had fewer outliers than the transformed model (Figure 14). Overall, we concluded that the selected model was optimal enough and did not require further transformation. To exclude the effect of outliers in this model, we calculated the Cook’s distance (D). The result showed that three cases have extremely large Ds, among which one outlier has an extremely large CRM rate (295.98, outlier of the CRM rate), another has extremely high poverty (36.3, outlier of the poverty variable). Therefore, we excluded these two extreme outliers but kept the remaining. The updated model showed a higher R-squared value (0.57 vs 0.54) which means the CRM rate is better explained by the variables after excluding outliers from the model. The normality of the updated model is also verified as improved in the QQ plots. Besides, no multicollinearity exists in the final model (all VIF < 5).

Model Interpretation Based on the above process, we finalized our model as a multivariable linear regression model with interaction. Specifically, percent of the population aged 18-34, percent of the population aged 65+, percent of the population below poverty level, geographic regions, log transformation of land area, and log transformation of the total population, together with the interaction between poverty and total population, are included as the independent variable in the model. The crime rate per 1,000 population is set as the dependent variable. All variables are significantly correlated with crime rate (P<0.05, table 3), while only land area and percent below the poverty level are negatively correlated with the crime rate.

Conclusions/Discussion

In conclusion, we established a multivariable linear regression model to predict the crime rate in the country based on six populational variables. Interactions are also considered between poverty and regions, which suggests that poverty in different regions may have different effects on crimes. Unconsidered geographical and political reasons may contribute to this interaction process. It is also in accordance with the model estimate result that the region factors showed the most significant relationship with crime rates. The adjusted R2 is barely 0.56, which means this model can only explain 56% of the variance of the dependent variable, a little lower than our expectation and cannot be a precisely predicted model. Through our analysis, we have screened useful variables and built a complete model. In the next step, we want to make our model more efficient through more extensive data and more variables. We want to study other relevant factors that affect the crime rate in the region. By making assumptions, finding relevant data, establishing models, and completing the analysis, our model will become more superior and extensive.

Figures and Tables

Figure 1: Distribution of CRM_1000

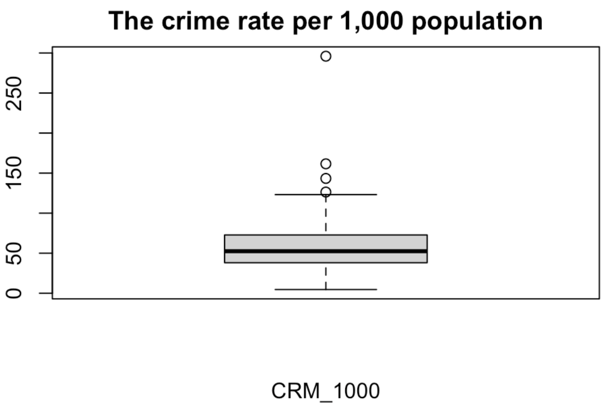


Figure 2: Distribution of variables before log transformation

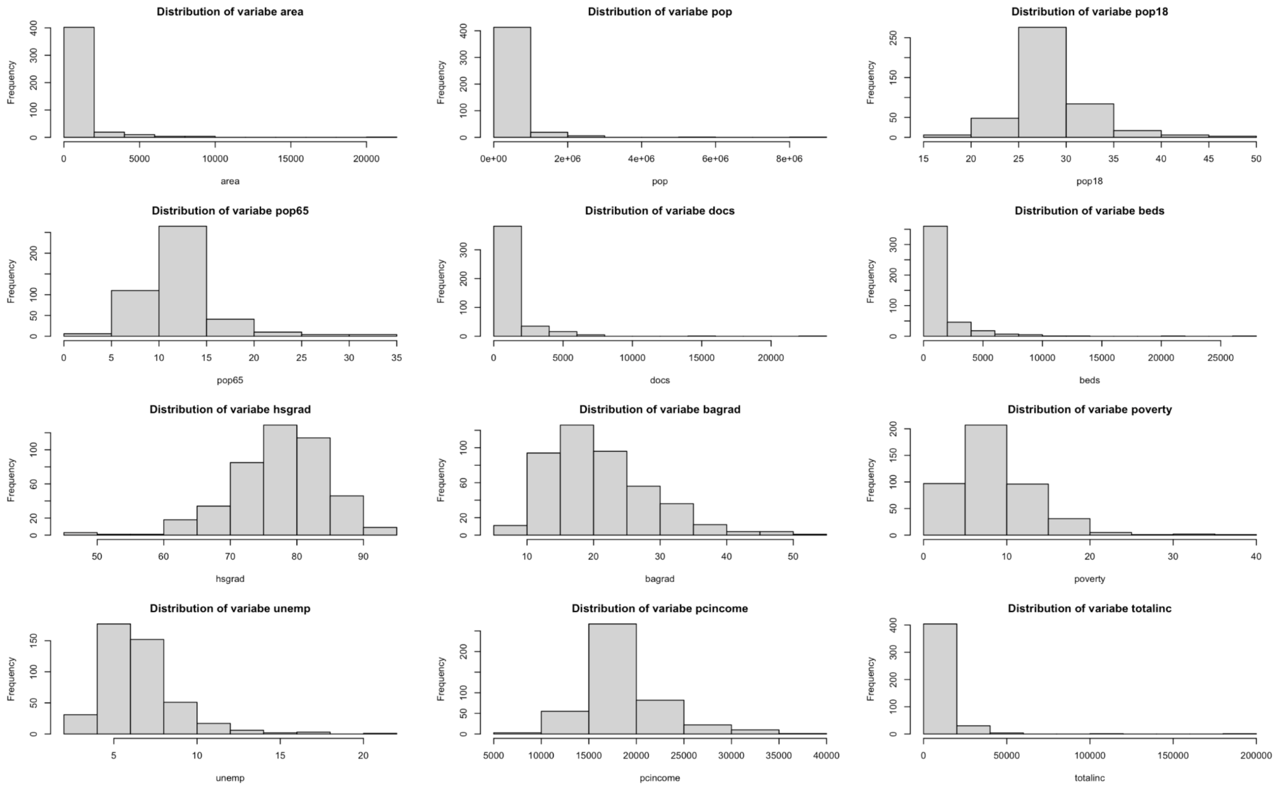


Figure 3: Distribution of variables after log transformation

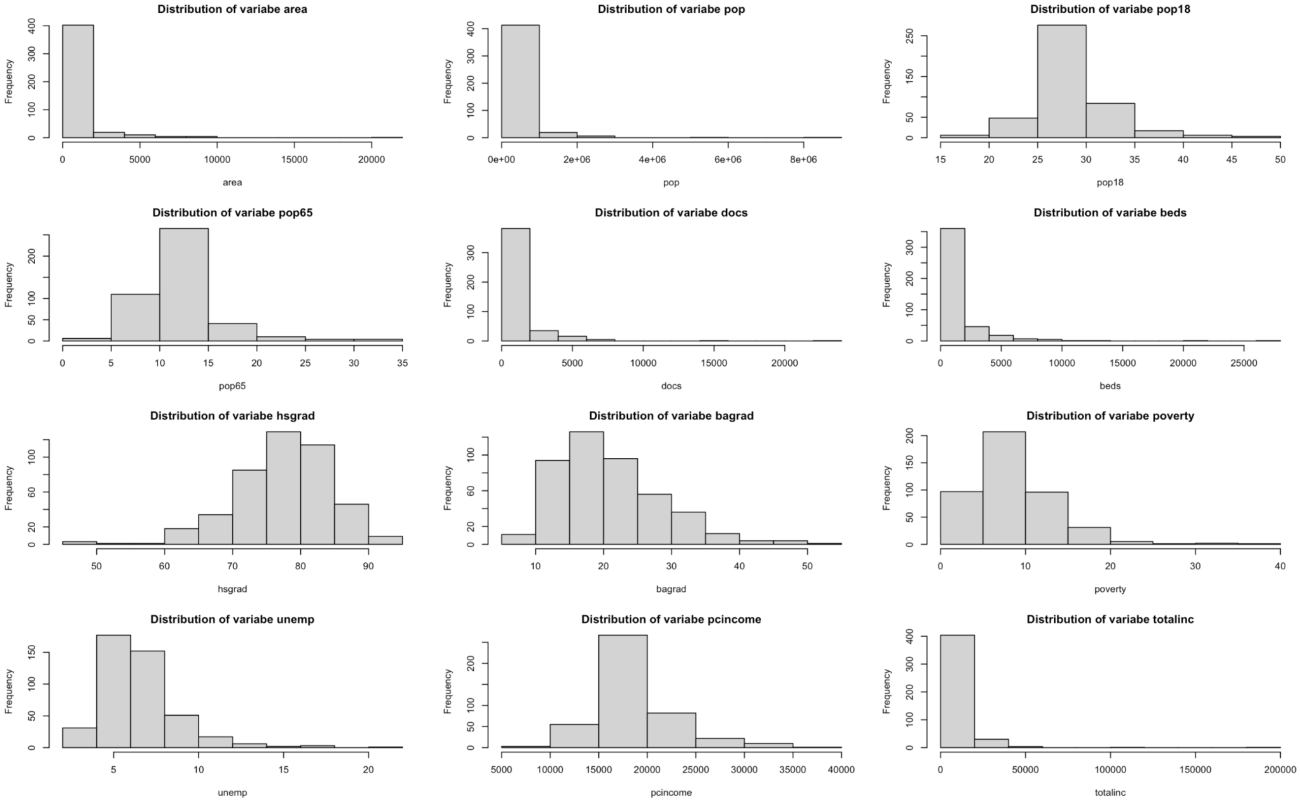


Figure 4: Correlation plot of all variables

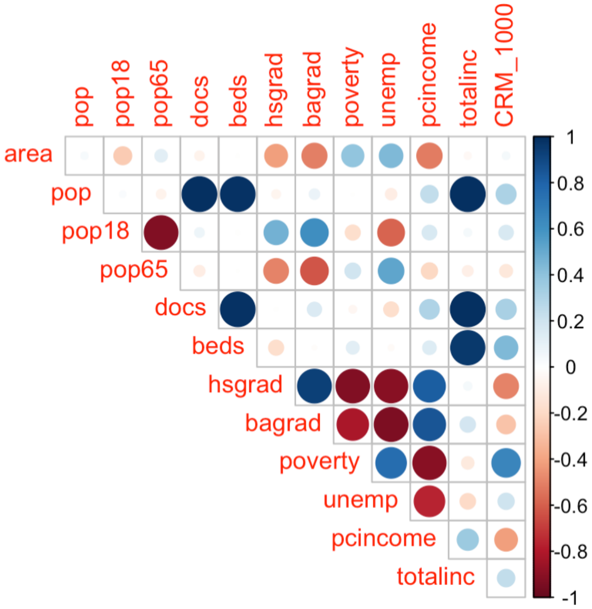


Figure 5: Interaction plot between region and pop18

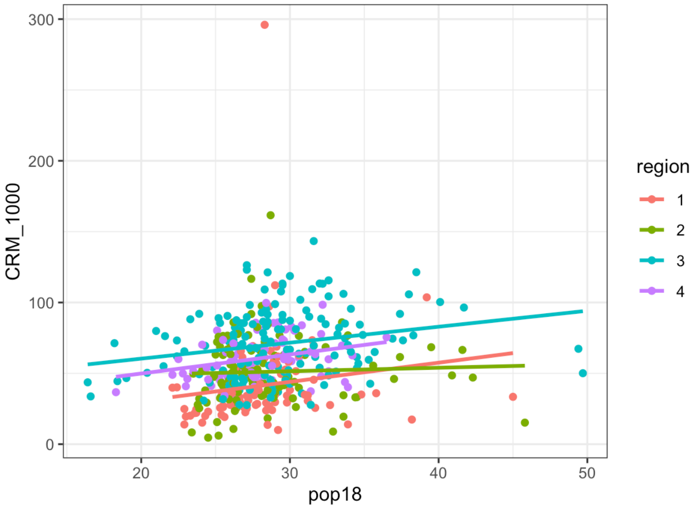


Figure 6: Interaction plot between region and poverty

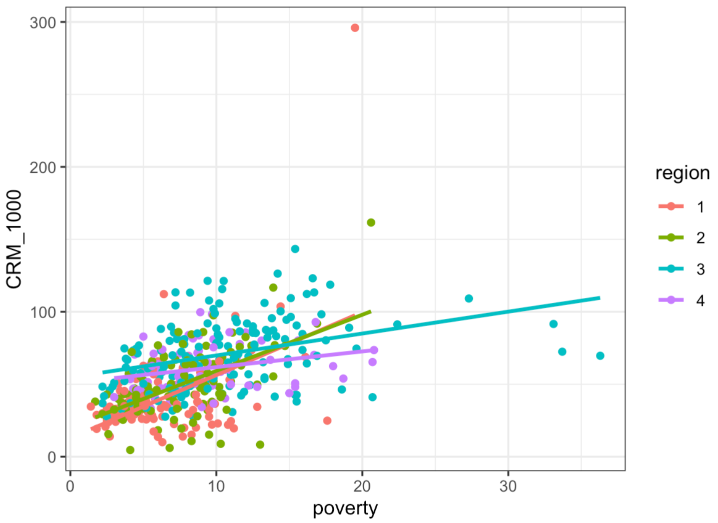


Figure 7: Interaction plot between region and pop65

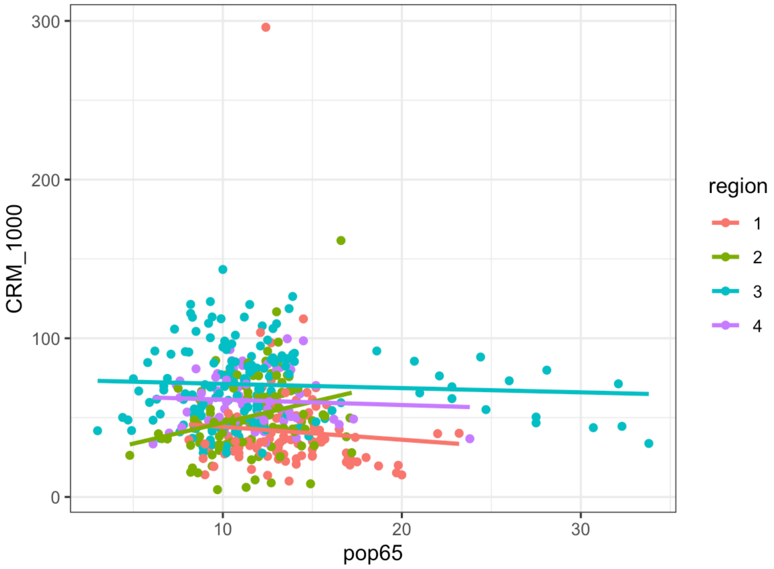


Figure 8: Interaction plot between region and log_area

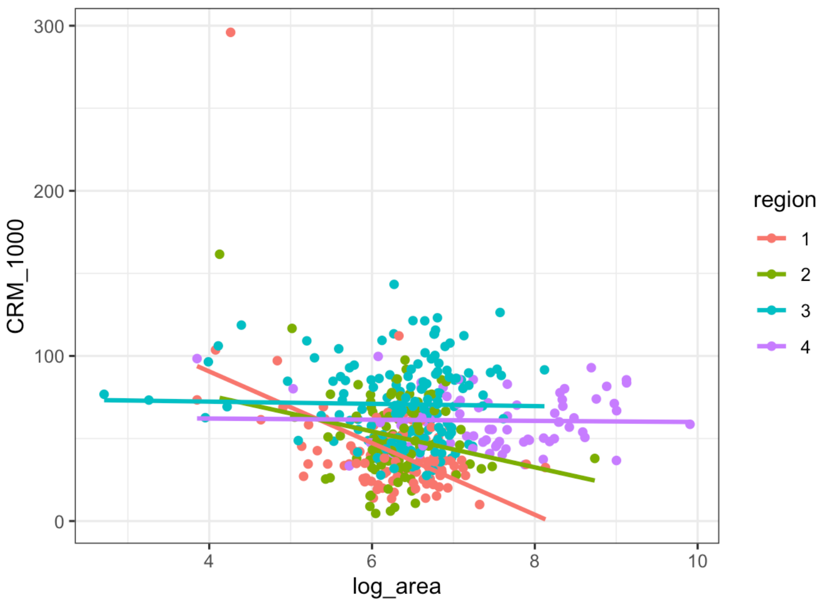


Figure 9: Interaction plot between region and log_pop

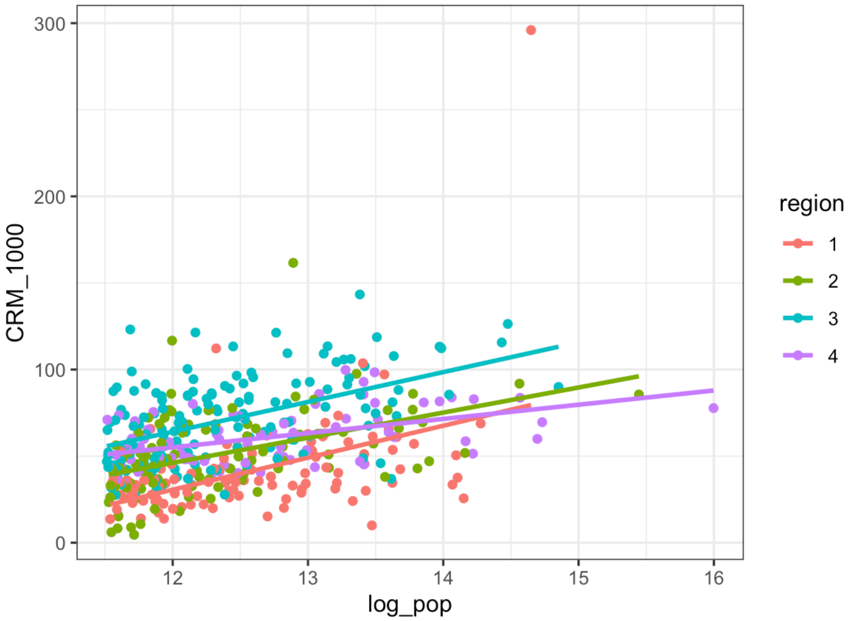


Figure 10: Box-cox plot of the selected model

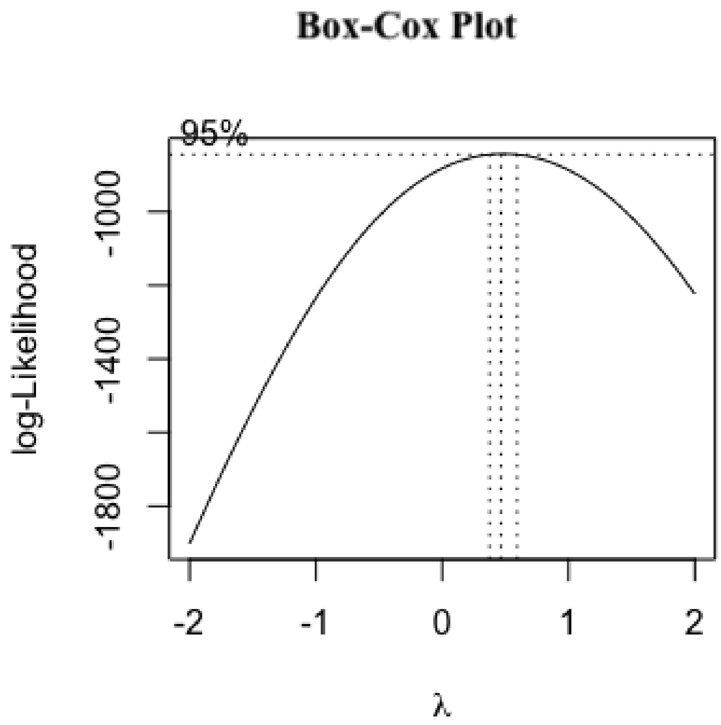


Figure 11: Residuals vs fitted plot of the selected model

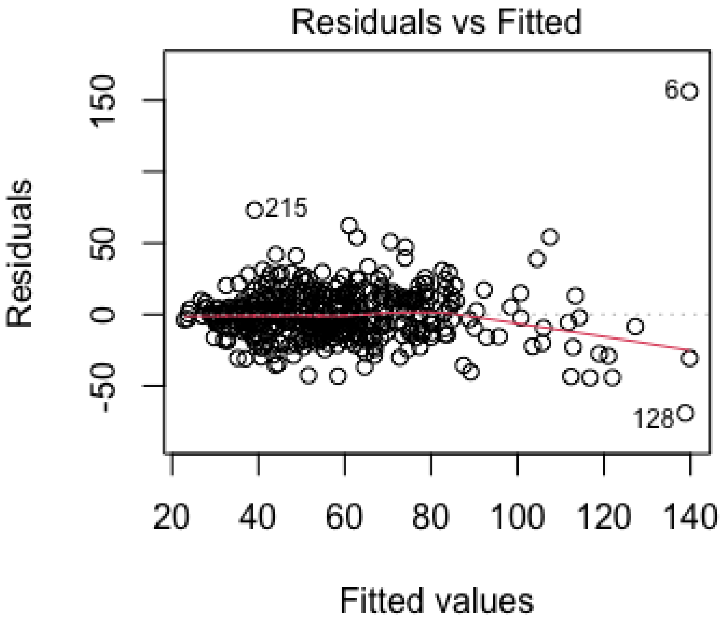


Figure 12: Residuals vs fitted plot of the transformed model

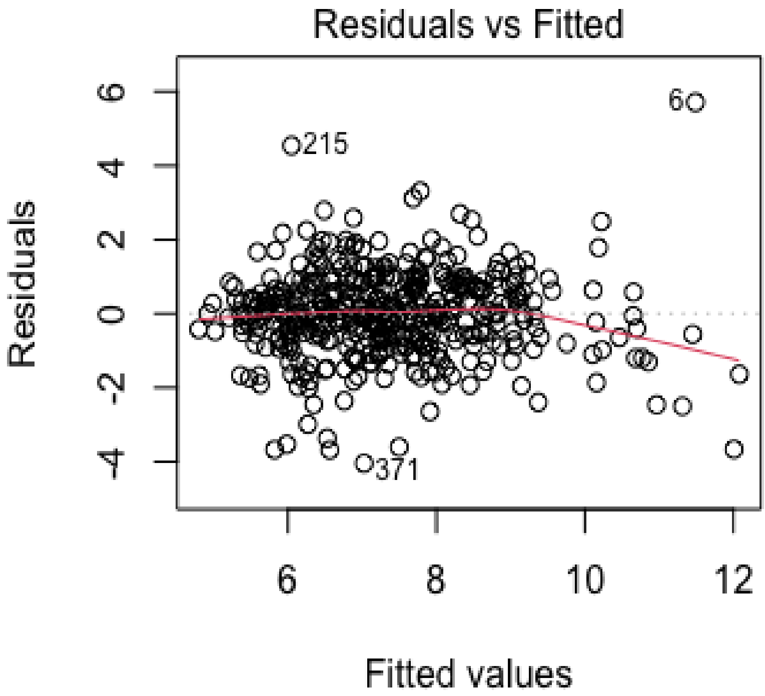


Figure 13: QQ plot of the selected model

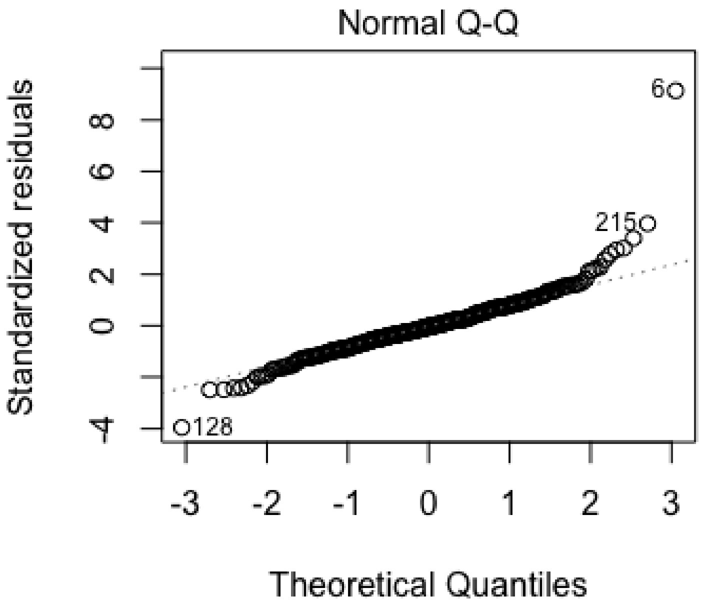


Figure 14: QQ plot of the transformed model

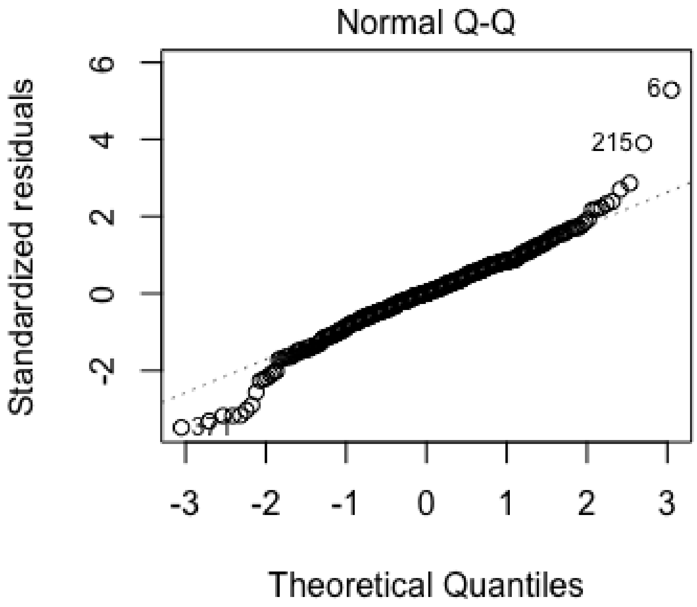
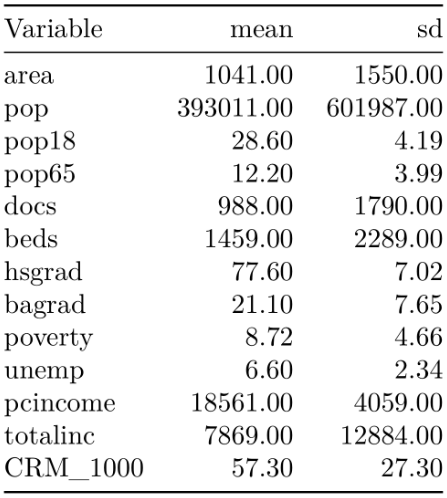


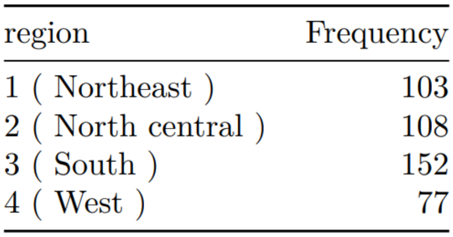
Table 1: A Table for Numeric Variables



Variable	mean	sd
area	1041.00	1550.00
pop	393011.00	601987.00
pop18	28.60	4.19
pop65	12.20	3.99
docs	988.00	1790.00
beds	1459.00	2289.00
hsgrad	77.60	7.02
bagrad	21.10	7.65
poverty	8.72	4.66
unemp	6.60	2.34
pcincome	18561.00	4059.00
totalinc	7869.00	12884.00
CRM_1000	57.30	27.30

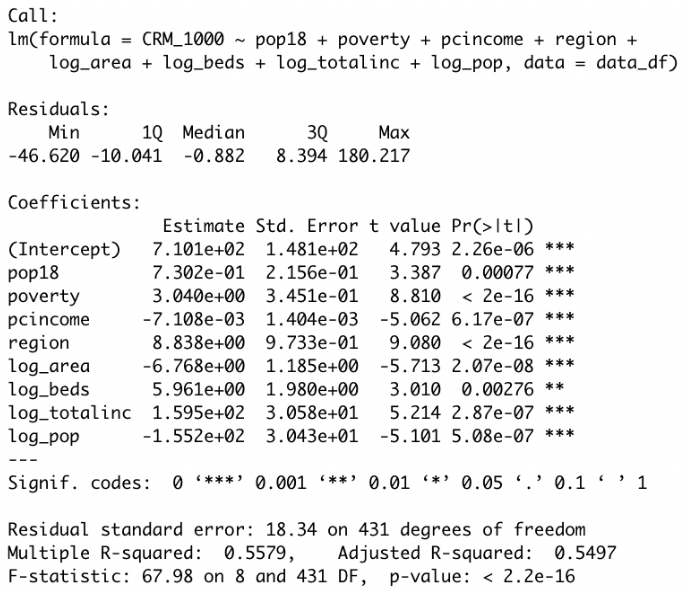
Table 2 Frequency of Different Regions

Table 2: A Frequency Table



region	Frequency
1 (Northeast)	103
2 (North central)	108
3 (South)	152
4 (West)	77

Table 3 stepwise selection of multivariable linear regression model and correlated statistics



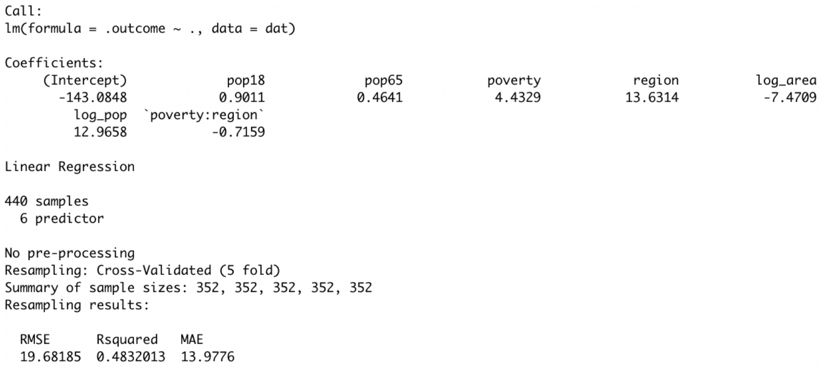
```
Call:
lm(formula = CRM_1000 ~ pop18 + poverty + pcincome + region +
    log_area + log_beds + log_totalinc + log_pop, data = data_df)

Residuals:
    Min      1Q  Median      3Q     Max
-46.620 -10.041  -0.882   8.394 180.217

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.101e+02  1.481e+02   4.793 2.26e-06 ***
pop18         7.302e-01  2.156e-01   3.387  0.00077 ***
poverty       3.040e+00  3.451e-01   8.810  < 2e-16 ***
pcincome     -7.108e-03  1.404e-03  -5.062 6.17e-07 ***
region        8.838e+00  9.733e-01   9.080  < 2e-16 ***
log_area     -6.768e+00  1.185e+00  -5.713 2.07e-08 ***
log_beds      5.961e+00  1.980e+00   3.010  0.00276 **
log_totalinc  1.595e+02  3.058e+01   5.214 2.87e-07 ***
log_pop      -1.552e+02  3.043e+01  -5.101 5.08e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.34 on 431 degrees of freedom
Multiple R-squared:  0.5579,    Adjusted R-squared:  0.5497
F-statistic: 67.98 on 8 and 431 DF,  p-value: < 2.2e-16
```

Table 4 Model Estimates of final multivariable linear regression model with interaction

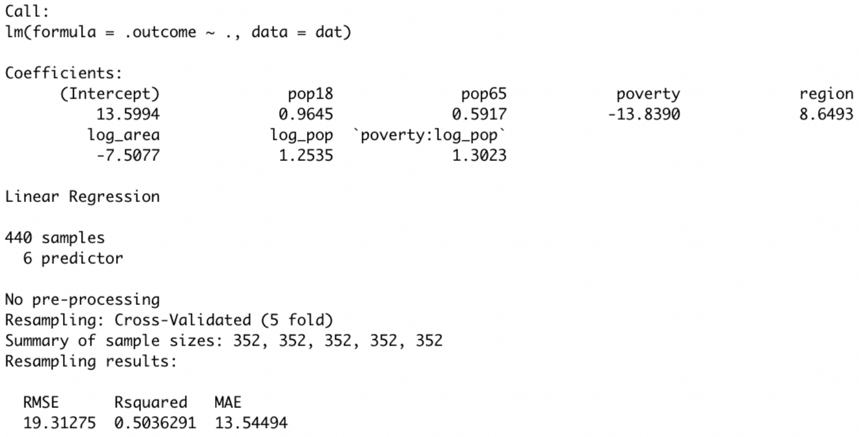
Table 4: Final Model Estimates

term	estimate	std.error	statistic	p.value
Geographic region - South	25.65	2.28	11.27	0.0000
Geographic region - West	20.93	2.86	7.33	0.0000
Geographic region - North Central	11.76	2.35	5.00	0.0000
Log(Land area)	-4.44	1.07	-4.13	0.0000
Percent of population - aged 18-34	0.91	0.25	3.72	0.0002
Interaction between poverty and total population	0.74	0.24	3.03	0.0026
Log(Total population)	5.93	2.42	2.46	0.0145
Percent below poverty level	-7.05	3.03	-2.33	0.0204
Percent of population - aged 65+	0.54	0.26	2.09	0.0375
Residual	-54.58	31.81	-1.72	0.0869

Table 5 Cross validation of final multivariable linear regression model



```
Call:
lm(formula = .outcome ~ ., data = dat)

Coefficients:
     (Intercept)             pop18             pop65           poverty            region          log_area  
       -143.0848            0.9011            0.4641            4.4329           13.6314           -7.4709  
         log_pop  `poverty:region`  
         12.9658           -0.7159  

Linear Regression 

440 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold) 
Summary of sample sizes: 352, 352, 352, 352, 352 
Resampling results:

  RMSE      Rsquared   MAE    
  19.68185  0.4832013  13.9776
```



```
Call:
lm(formula = .outcome ~ ., data = dat)

Coefficients:
      (Intercept)              pop18              pop65            poverty             region  
          13.5994             0.9645             0.5917           -13.8390             8.6493  
         log_area            log_pop  `poverty:log_pop`  
          -7.5077             1.2535             1.3023  

Linear Regression 

440 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold) 
Summary of sample sizes: 352, 352, 352, 352, 352 
Resampling results:

  RMSE      Rsquared   MAE     
  19.31275  0.5036291  13.54494
```

Reference

[1] "Criminal Victimization, 2019" (PDF). U.S. Department of Justice.

https://bjs.ojp.gov/content/pub/pdf/cv19.pdf

[2] "US records highest increase in nation's homicide rate in modern history, CDC says". J.Howard, CNN.

https://www.cnn.com/2021/10/06/health/us-homicide-rate-increase-nchs-study/index.html

[3] Groot, Wim & Maassenvandenbrink, H.. (2010). The effects of education on crime. Applied Economics. 42. 279-289. 10.1080/00036840701604412.

https://www.cnn.com/2021/10/06/health/us-homicide-rate-increase-nchs-study/index.html

[4] Hsieh, C.-C., & Pugh, M. D. (1993). Poverty, Income Inequality, and Violent Crime: A Meta-Analysis of Recent Aggregate Data Studies. *Criminal Justice Review*, *18*(2), 182–202.

https://doi.org/10.1177/073401689301800203