STATISTICS 378
Using Regression to Predict Total Crime

December 8, 2023

Jacob Winch, Parteek Baidwan, Bennett Piercy

"All models are wrong, but some are useful" - George Box[10]

Abstract

It is self-evident that the effective handling of crime is paramount to the well-being of any community. As such, the practice of "crime analysis" has proved to be an enduring field of statistical analysis, and recent developments in the world of police management have increasingly shifted the crime analysis focus towards predictive metrics at a rapid pace. In this paper, we explore statistics pertaining to major metropolitan areas in the US during the 1970s and aim to identify relationships between total serious crime and several other statistics. To achieve this goal, we use multiple linear regression analysis in R and ultimately fit a final model using Ridge regression. We split the dataset into a prediction and estimation set using the DUPLEX algorithm. The initial model that we fit did not satisfy the assumptions of linear regression, so we had to consider various transformations to satisfy the assumptions. In addition, we added interaction terms to the model. We eventually used forward stepwise regression to conclude the final model. For our results, we will first check the validity of the model using our prediction data to see if the model is valid, then we will fit our model to the full data set and check to see if we pass our assumptions. Afterwards, we will check for influential and leverage points in order to see if we require robust regression. Additionally, we will check if our model has a multicollinearity problem and if we do we will find our final model by doing penalized regression. Our final model features the variables land area, total population, percentage of population 65 or older, number of active physicians, percentage of high school graduates, civilian labour force, total person income, a set of indicator variables corresponding to the broader region, and several interaction terms.

Contents

1. Introduction

1.1 Motivation

Occurrences of serious crime are a top concern of governing officials and private citizens alike. Accurately predicting the frequency and modes of future criminality is especially of interest to lawmakers seeking to craft public policies that best protect the citizenry. As such, studies identifying factors associated with local criminality are highly sought-after, and the "crime analysis" industry has proven to be an evergreen field of study in the world of statistical modeling and machine learning. Crime analysis dates back to at least the 1820s[4], but the practice has recently seen an explosion in interest and funding, with the estimated market size of the controversial "predictive policing" industry surpassing $5 billion USD in 2023[5].

1.2 Aim

This project aims to provide insight into the possible associations between crime and other major statistics of large cities and the areas immediately surrounding those cities. This is a popular area of research. An advantage of our project is that it offers an opportunity to introduce new perspectives to an area of statistical research for which model adequacy, model validity, and proper interpretation are imperative. Crime statistics and models predicting crime are an essential tool for police, and as such improper or incorrect modeling practices may result in serious effects on the public's safety and liberty.

1.3 Research Question

Our research question is as follows: what factors about population centers show a significant association with the amount of serious crime recorded in said population centers?

Stated in terms of the dataset available to us for this paper, our research question is as follows: is there a significant association between total serious crimes and land area, total population, percent of population in central cities, percent of population 65 and over, number of active physicians, number of hospital beds, percent high school graduates, civilian labor force, total personal income, and geographic region (here represented by to the classification according to the US Bureau of the Census)?

1.4 Approach

In order to address our research question, in this paper we will use multiple linear regression analysis to explore the relationship between crime and 10 other relevant statistics pertaining to population centers around cities. Here the dependent variable total serious crimes is provided by reports from law enforcement agencies and consists of crimes of a violent or otherwise major nature.

1.5 Related Work

Here we discuss three studies relevant to our paper, and elaborate on findings that affect the approach and interpretation of our paper.

1.5.1

"Analysis of the Factors Influencing Crime Occurrence by Commercial through Application of the Spatial Geography Weighted Regression Model" by Yeo (2023).

This study examines the crime rate in several of Seoul's commercial districts. According to the study, "the factor with the greatest influence on the occurrence of crime in the commercial districts of Seoul was the number of stores in the hops and snack bar industries" (Yeo). In this paper, a spatial geography weighted regression model was used, which differs from our multiple regression model.

1.5.2

"Analysis of thefts in Colombia during 2017 using multiple linear regression models and geographically weighted regression" by Lopez, Aceros, and Luzardo Briceño (2019).

In this study, the number of thefts in different Colombian municipalities were modeled using multiple linear regression. This study differs from our own principally in that it focuses on instances of theft, while our follows all occurrences of serious crime.

1.5.3

"The Impact of Measurement Error in Regression Models Using Police Recorded Crime Rates" by Pina-Sánchez, Buil-Gil, Brunton-Smith, and Cernat (2023).

This study provides a review of other papers using regression analysis to predict crime rates. Specifically, the study investigates the extent to which measurement error by police affects the coefficients of regression models predicting crime. Their findings were that the effect of this kind of error is "highly variable across different settings" and "biases [of this nature] could range from negligible to severe".

1.5.4

The findings presented in this section have direct applicability to our paper in numerous ways. Firstly, the papers discussed in sections 1.5.1 and 1.5.2 have bearing on our effort here since they comprise other cases of regression modeling being used to predict the occurrence of crime. The paper from Yeo found that the number of retailers, or more broadly the extent of commerce in an area, is an important factor in this type of analysis. It is noteworthy that no variables in the dataset available to us represent this effect.

The study from Lopez et al. is related to our paper since it aims to identify common factors among different areas insofar as those factors are related to the rate of a type of crime in

that area. This study also demonstrates some data transformations and other analysis techniques which may be relevant to our paper.

Although the analysis from Pina-Sánchez et al. discussed in section 1.5.3 is not principally focused on a novel prediction model, this work is extremely relevant to our paper nonetheless. Their findings suggest that there is a potential data integrity issue that warrants consideration when analyzing models of the sort that we present here. This is further discussed in section 4.2.

2. Methodology

2.1 Methodology Overview

All analysis in this paper was conducted in R (visuals and graphical outputs included). All code used in our analysis is available in a GitHub repository linked in the appendix.

2.2 The Data

The dataset used in this paper was provided in "Applied Linear Regression Models" by Neter, Kutner, Nachtsheim, and Wasserman (1996). It consists of eleven variables and an identification number for 141 Metropolitan Statistical Areas (MSAs). Metropolitan Statistical Areas are geographic areas of the United States as classified by the United States Bureau of the Census. They principally correspond to major population centers that include a densely populated core area that represents a local hub of social and economic activity[6]. The variables included in the dataset were collected from multiple sources each originating from a specific year between 1970 and 1977. At the time these data were collected, MSAs were referred to as Standard Metropolitan Statistical Areas (SMSAs).

2.3 Splitting the data with the DUPLEX algorithm

We split the original 141 rows of the SMSAs dataset into an estimation and prediction set. The prediction set has a size of 25 data points and the rest is in the estimation set. The estimation and prediction sets can be found in tables 2.3.1 and 2.3.2 respectively in the appendix. We split the data using our implementation of the DUPLEX algorithm[9]. The data was split since collecting new data for validation purposes was not possible. The estimation data will be used to build the model and the prediction data will be used to understand the predictive ability of the model.

2.4 Fitting the Regression Model

In this paper, we will fit a multiple linear regression model to describe the relationship between total serious crime and several other factors. We fit the following preliminary multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12}$$

where:

y = Total Serious Crimes, Total number of serious crimes in 1977, including murder, rape, robbery, aggravated, assault, burglary, larceny-theft, and motor vehicle theft, as reported by, law enforcement agencies.

x_1 = Land Area, In square miles.

x_2 = Total Population, Estimated 1977 population (in thousands).

x_3 = Percent of population in central cities, Percentage of 1979 SMSE population in central cities.

x_4 = Percentage of population 65 or older, Percentage of 1979 SMSA population 65 years old or older.

x_5 = Number of active physicians, Number of professionally active nonfederal physicians as of December 31, 1977.

x_6 = Number of hospital beds, The number of beds, cribs and bassinets during 1977.

x_7 = Percentage of high school graduates, Percentage of adult population (person years old or older) who completed 12 or more years of school, according to the 1970 Census of Population.

x_8 = Civilian labor force, Total number of persons in civilian labor force (person 16 years old or older classified as employed or unemployed) in 1977 (in thousands).

x_9 = Total person income, Total current income received in 1976 by residents of the SMSA from all sources, before deduction of income and other personal contributions to social security and other social insurance programs (in millions of dollars).

x_{10} = NE, Indicator variable as to whether an individual lives in the North East.

x_{11} = NC, Indicator variable as to whether an individual lives in the North Central.

x_{12} = S, Indicator variable as to whether an individual lives in the South.

The initial regression model when fit to the estimation data using R is:

$$\hat{y} = -11630 + 0.8563x_1 + 58.29x_2 + 52.33x_3 - 2.288x_4 + 5.245x_5 - 2.496x_6 + 124.3x_7 - 5.873x_8 + 2.221x_9 - 8233x_{10} - 5302x_{11} + 92.08x_{12}$$

Which has $R^2 = 0.9845$ and $R^2_{Adjusted} = 0.9827$, therefore 98.27% (adjusted) of the variation in total serious crimes is explained by the preliminary regression model. However, in order for the results of this model to be valid we need to check if the model satisfies the assumptions.

2.5 Residual analysis

If we consider Figure 2.5.1, the Normal Q-Q Plot for the sample values vs the theoretical values, we can see that our sample quantiles show severe deviation from the straight line. This indicates that the assumption of normal populations is violated. In order for a linear regression model to be valid, for each value of the predictor variable the conditional distribution of the response variable should follow a Normal distribution[8].

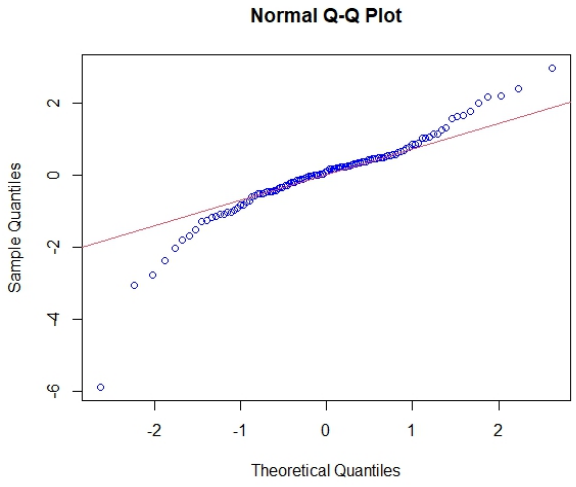

Figure 2.5.1: A Normal Q-Q plot for the preliminary data

Also considering Figure 2.2, the plot of residuals vs fitted values, the plot shows a clear pattern of an outward funnel shape indicating that the assumption of equal standard deviations is also violated, in addition to the violation of the assumption of normality. In order for a linear regression model to be valid the conditional standard deviations of the response variable should be the same for all values of the predictor variable[8].

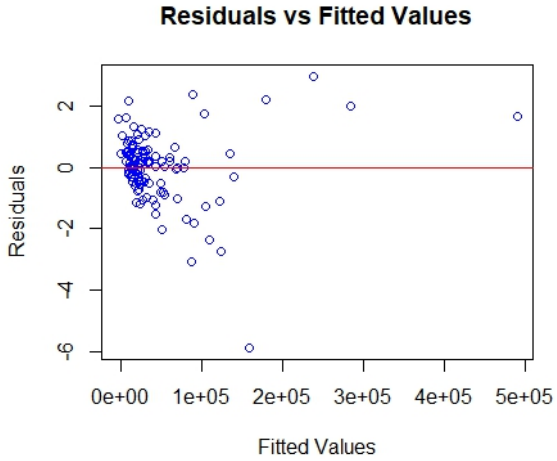

Figure 2.5.2: A plot of the residuals vs fitted values for the preliminary model

The results from these plots, as well as the residuals vs regressor plots, the partial-regression plots, and the partial-residual plots (available in the appendix), indicate that data transformations are necessary to ensure the validity of a final model.

2.6 Transformations

After applying transformations identified through an ad-hoc process, we finally concluded a transformed linear regression model of:

$$\frac{1}{lny} = \beta_0 + \beta_1\frac{1}{\sqrt{x1}} + \beta_2\sqrt{x2} + \beta_3x_3 + \beta_4\ln x_4 + \beta_5\ln x_5 + \beta_6\sqrt{x6} + \beta_7x_7 + \beta_8\frac{1}{lnx8} + \beta_9\sqrt{x9} + \beta_{10}x_{10} + \beta_{11}x_{11} + \beta_{12}x_{12}$$

Note: for the remainder of this paper, we will refer to the transformed regressors using the notation of x^*. Therefore the transformed model is:

$$y^* = \beta_0 + \beta_1x_1^* + \beta_2x_2^* + \beta_3x_3 + \beta_4x_4^* + \beta_5x_5^* + \beta_6x_6^* + \beta_7x_7 + \beta_8x_8^* + \beta_9x_9^* + \beta_{10}x_{10} + \beta_{11}x_{11} + \beta_{12}x_{12}$$

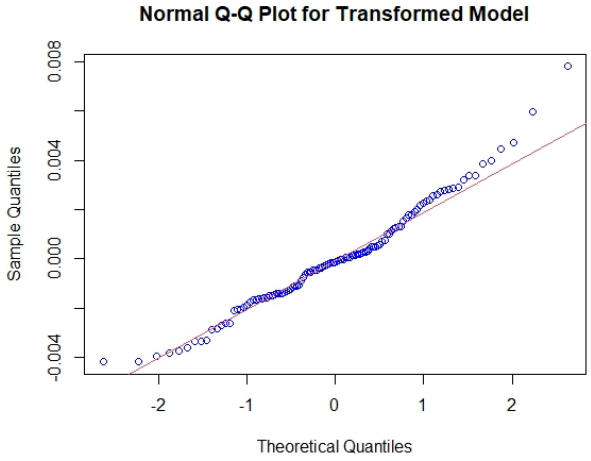

Figure 2.6.1: Q-Q plot for the transformed model

Referring to Figure 2.6.1 (Normal Q-Q plot after data transformation), we see that slight deviation from the theoretical quantiles exists, but it is not significant enough to violate the normality assumption. Therefore, we conclude the transformations have satisfied the normality assumption. This conclusion is supported by the results of the Shapiro-Wilk test for normality. The p-value for the Shapiro-Wilk test on the transformed y is 0.3997. Therefore, we have little to no evidence against the null hypothesis, and thus we conclude that the assumption of normality is satisfied.

The residuals vs fitted values plot, Figure 2.6.2 shows that assumption of equal variance is now satisfied as well, since there is no significant pattern displayed in the figure.

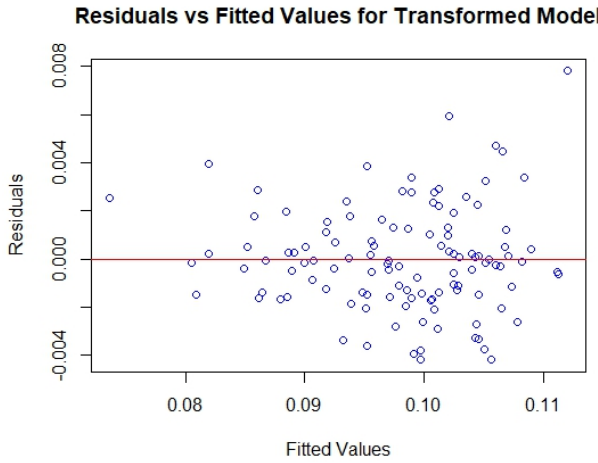

Figure 2.6.2: Residuals vs Fitted values plot for the transformed model

The Breusch-Pagan test, which tests for heteroskedasticity in a linear regression model has a p-value of 0.2101 from R, therefore we fail to reject the null hypothesis, thus our transformation has satisfied the assumption of equal variances.

2.7 Possible Interactions Between Regressors

Guided by intuitive analysis through a trial-and-error approach informed by detailed ANOVA tables, our research identified significant interaction terms that merit inclusion in the full model. These ANOVA tables are provided in the appendix for reference. The complete model, which forms the core of our discussion in this paper, is as follows:

$$y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3 + \beta_4 x_4^* + \beta_5 x_5^* + \beta_6 x_6^* + \beta_7 x_7 + \beta_8 x_8^* + \beta_9 x_9^* + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_5^* x_{12} + \beta_{14} x_7 x_{11} + \beta_{15} x_1^* x_{10} + \beta_{16} x_2^* x_{10} + \beta_{17} x_3 x_{10} + \beta_{18} x_4^* x_{10} + \beta_{19} x_5^* x_{10}$$

2.8 Variable Selection

Considering the comprehensive nature of the full model as described in section 2.7, we implemented various computational techniques for variable selection. These included forward elimination, backward elimination, and stepwise selection. Ultimately, the forward selection process proved most effective, culminating in the smallest model that still achieved the highest adjusted R^2. Therefore, we selected the output of the forward selection for our study. The forward elimination process concluded with:

$$y^* = 0.064 + 0.0640x_1^* - 0.00047x_2^* + 0.00107x_4^* - 0.0004181x_5^* - 0.000132x_7 + 0.230x_8^* + 0.00013x_9^* + 0.0177x_{10} - 0.0082x_{11} + 0.00017x_7x_{11} - 0.1230x_1^*x_{10} + 0.00039x_2^*x_{10} + 0.0091x_4^*x_{10} + -0.0063x_5^*x_{10}$$

Which has an $R^2 = 0.9457$ and an adjusted $R^2 = 0.9382$. Therefore 93.82% (adjusted) of the variation in the transformed total serious crimes is explained by this regression model.

3. Results

3.1 Results Overview

In order to provide a comprehensive analysis of our model, we will address the validity of the model, the presence of influential points in the data to see if we need robust regression, the presence of multicollinearity among predictor variables, and finally the use of penalized regression in relation to our final model if we have multicollinearity. We will then present our findings for which variables are the best predictors for total crime.

3.2 Model Validation

We will now be fitting our model with the prediction data from the DUPLEX algorithm splitting to check the validity of our model. With R we find the PRESS statistic is 160270573 and that our $R^2_{\text{prediction}} = 0.9457$ using the prediction data. Since we have an $R^2 = 0.9432$ from our estimation data we can see that the difference between these is ($|0.9432 - 0.9457|$) which is less than 0.1, which suggests that our model is valid.

3.3. Fitting the model to the full data

Since we concluded that our model was valid in section 3.2, we fit our model to the full SMSAs dataset. After we fit the model to the full data we get this model:

$$y^* = 0.072 + 0.024x_1^* - 0.00050x_2^* + 0.00020x_4^* - 0.00048x_5^* - 0.000125x_7 + 0.206x_8^* + 0.0001297x_9^* + 0.01956x_{10} - 0.0040x_{11} + 0.00010x_7x_{11} - 0.0946x_1^*x_{10} + 0.000208x_2^*x_{10} + 0.00623x_4^*x_{10} + -0.00487x_5^*x_{10}$$

Which has an R^2 of 0.9432 and an R^2_{adjusted} of 0.9369. Therefore 93.69% (adjusted) of the variation in the transformed total serious crimes is explained by the preliminary regression model.

The assumptions for the full model are still satisfied as shown in Figures 3.3.1 and 3.3.2

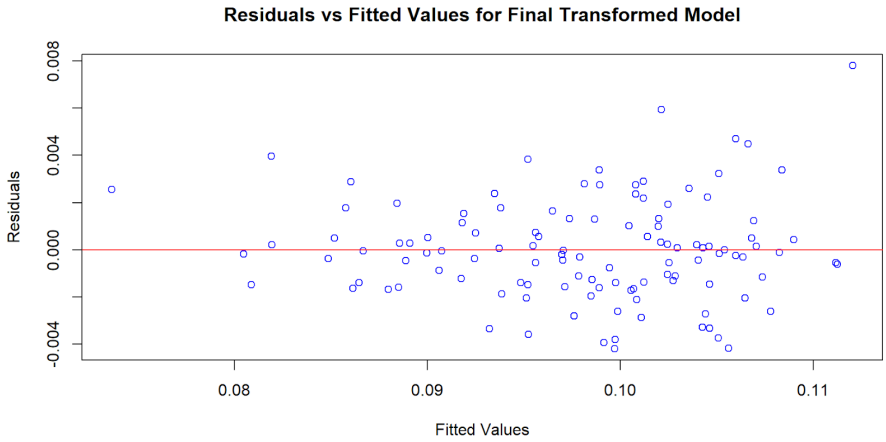

Figure 3.3.1: The residuals vs fitted values plot for the final transformed model indicating no clear pattern, so the assumption of constant variance is satisfied.

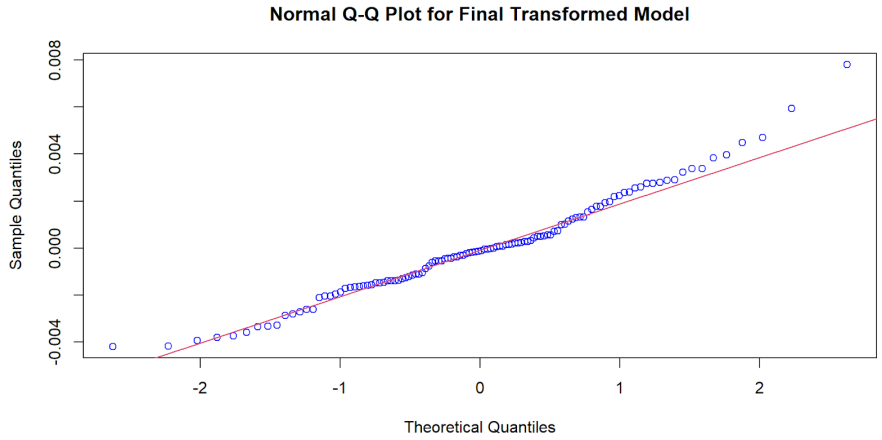

Figure 3.3.2: The Q-Q plot for the final transformed model indicates no severe deviation from the line, so the assumption of normality is satisfied.

3.3 Influential & Leverage Points

From the appendix 3.3.1 we can see Cook's distance and DFFITS, we get influential points 1, 2, 8, 40, 59, 62, 69, 73, 82, 127, 128 and 141. From R we also get leverage points 1, 2, 3, 4, 8, 54, 59, 62, 104, 123, 133 and 141. The presence of influential points leads us to conclude that we need to employ robust regression methods to fit our final model.

3.4 Multicollinearity

As we can see from the appendix figure 3.4.1, R outputs a VIF of greater than 10 for variables x_2, x_4, x_6, x_7, x_8, x_9 and the interaction terms. In R we also get a condition index greater than 30, which means there exists a collinearity problem which may threaten the integrity of the model.

Many strategies are available to combat this problem, such as model respecification, collecting additional data or using penalized regression. In our paper we will consider running a penalized regression model that performs ridge regression.

3.5 Penalized Regression Model

Since collinearity exists in our data, we will now use penalized regression in order to resolve this issue. We do this using the variables we have in our transformed model from above and we use Ridge regression to find our new model. For Ridge regression we will add the penalty term to the least squares estimator. With that we get the final model:

$y^* = 0.08889648 - 0.001955627\, x_1^* - 0.00008478262\, x_2^* + 0.0006244373\, x_4^* - 0.00195231\, x_5^* - 0.00008026747\, x_7 + 0.1487878\, x_8^* - 0.00002085494\, x_9^* + 0.002469305\, x_{10} + 0.0004822058\, x_{11} - 0.00002571456 x_{10} x_2^* - 0.03715647 x_{10} x_1^* + 0.00002145102 x_{11} x_7 + 0.00003187017 x_{10} x_5^* + 0.001335878 x_{10} x_4^*$

3.6 Interpretation of the Final Model

Our final model gets us a an $R^2 = 0.9222516$ and a $R^2_{\text{Adjusted}} = 0.9129218$ which means we can explain 91.29% of the variation of total serious crime with our final model, which is worse than our initial model without transformations, but nonetheless preferable since that model did not adhere to the assumptions of linear regression. Our final model clears all necessary assumptions such as normality and homoscedasticity and is a valid model. Our model has also adequately addressed the collinearity problems present in our attempts prior to using Ridge regression. One possible issue that still exists in our model is that we have influence and leverage points suggesting we may benefit from the use of robust regression techniques. We can see in the final model that total crime is predicted by these factors: land area, total population, percentage of population 65 or older, number of active physicians, percentage of high school graduates, civilian labor force, total person income, and whether the person lives in the northeast or north central part of the United States of America. We also include the interactions between living in the north east and total population, living in the north east and land area, living in the north east and number of active physicians, living in the north east and percentage of population 65 or older and finally living in the north central and population 65 or older. Since the penalized model and the model we had before isn't that different we can use the other model for the prediction for our betas. So from figure 3.4.2 from at 95% confidence we can see that regressors x2,x7,x8,x9,x1:x10,x5:x10 are significant for predicting our total crime. So the significant factors are total population, number of hospital beds, percentage of high school graduates, civilian labor force, total land area north east USA interaction, and the number of active physicians north east USA interaction.

4. Conclusion

4.1 Summary

We observe that the overall incidence of total crime can be anticipated based on various factors, namely: land area, total population, the percentage of the population aged 65 or older, the count of active physicians, the percentage of high school graduates, civilian labor force, total personal income, and geographical location, specifically whether the individual resides in the northeast or north-central region of the United States of America. Additionally, our analysis incorporates the interactions between residing in the northeast and total population, residing in the northeast and land area, residing in the northeast and the number of active physicians, residing in the northeast and the percentage of the population aged 65 or older, and finally, residing in the north-central region and the population aged 65 or older. However we can see that the significant factors for predicting total crime is total population, number of hospital beds, percentage of high school graduates, civilian labor force, total land area north east USA interaction, and the number of active physicians north east USA interaction.

4.2 Limitations

The limitations of the model presented here are mainly related to the dataset the model was trained on, and the model's ability to produce useful estimates of crime in novel spatial and temporal contexts. The generalizability of these findings would be improved if data from other countries, instead of just the United States, were available. This is specifically relevant when it comes to the definition of individual observations as MSAs (and SMSAs) and the nature of the 'Geographic region' variable, since both of these factors are defined according to the US Bureau of the Census. Furthermore, the time at which the data were recorded may pose further issues to the external validity of the model. The regression fit in this paper was trained on data recorded over 40 years prior to the analysis presented here, and as such it may not perform as well on more recent data.

Certain characteristics of the dataset raise other concerns that may create limitations for the model. Firstly, the data comprising the dependent variables were collected from different sources, and not all of the variables correspond to the same year. For instance, the total population figures used in this model are based on estimates from 1977, while the percent high school graduates figure was instead derived from a census in 1970. Additionally, there was no apparent correction for non-standard changes in population and other statistics between MSAs that occurred in the interim between the earliest-reported statistics (1970) and the latest (1979). Secondly, some of the predictor variables were recorded after the response variable. Percent of

population in central cities and percent of population 65 or older were recorded in 1979, two years after the response variable total serious crimes was recorded. This poses an issue to model generalizability and also may have had a confounding effect on the analysis of those variables, or the model as a whole.

As presented by Pina-Sánchez et al., it is also relevant to mention the issue of using police-reported crime statistics in analysis of this nature. In the dataset used here, the total number of serious crimes was derived from reports from law enforcement agencies. This figure does not represent an unbiased account of all occurrences of crime, since over-policed areas will report a comparatively higher rate of crime than under-policed areas. The effect of this may be that factors that are generally believed to be associated with increased crime rate may become over-policed for that reason, leading to a heightened number of crime reports in those areas. This issue is especially salient due to the social context surrounding this issue.

Finally, the presence of influential points presents another limitation of our model. Future research to investigate this data analysis problem using robust regression analysis may be prudent.

4.3 Future Research

Given these factors outlined in the Limitations section, we propose various possible avenues for future research to build upon this paper. Firstly, a follow-up study analysis addressing some issues with the dataset used in this paper may produce a more effective final model. Secondly, based on our review of relevant literature, we have identified that the standard approach to using multiple regression modeling to predict crime is to use crime rate as the dependent variable. As such, future work may explore fitting a model to predict serious crime rate instead of total serious crimes. Finally, future work exploring some more complex modeling techniques including Ridge regression, LASSO regression, or neural networks would expand on the scope of this paper and may yield promising results.

Bibliography

1. Yeo, K. (2023, June). Analysis of the Factors Influencing Crime Occurrence by Commercial through Application of the Spatial Geography Weighted Regression Model. *The Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography* Vol. 41, No. 3, pp.153-165.
2. Lopez, N., Aceros, M., Luzardo Briceño, M. (2019, December). Análisis de los hurtos en Colombia durante el año 2017 mediante los modelos de regresión lineal múltiple y la regresión ponderada geográficamente (Analysis of thefts in Colombia during 2017 using

multiple linear regression models and geographically weighted regression). *Revista Criminalidad*, 61(3): 141-163.

3. Pina-Sánchez, J., Buil-Gil, D., Brunton-Smith, I. et al. The Impact of Measurement Error in Regression Models Using Police Recorded Crime Rates. *J Quant Criminol* 39, 975–1002 (2023). https://doi.org/10.1007/s10940-022-09557-6
4. Hunt, J. (2019, July). From Crime Mapping to Crime Forecasting: The Evolution of Place-Based Policing. *US Department of Justice | Office of Justice Programs*. https://www.ojp.gov/ncjrs/virtual-library/abstracts/crime-mapping-crime-forecasting-evolution-place-based-policing#:~:text=The%20history%20of%20this%20evolution,oriented%20policing%2C%20and%20many%20other
5. 360iResearch (2023, October). Predictive Policing Market by Technology (Crime Analysis, Facial Recognition Technology, Geo-Location Technology), End-Use (Detective Agencies, Law Enforcement Organization, Military & Defense) - Global Forecast 2023-2030. *360iResearch*. https://www.360iresearch.com/library/intelligence/predictive-policing
6. United States Bureau of the Census (2020). Our Surveys & Programs | Metropolitan and Micropolitan | About. *United States Census Bureau.* *ht*tps://www.census.gov/programs-surveys/metro-micro/about.html
7. Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996). Applied Linear Regression Models, *IRWIN*, pp. 1367–1368.
8. Simchi. A (2023), Stat 378 course materials
9. Snee, R. D. (1977). Validation of regression models: methods and examples. *Technometrics*, *19*(4), 415-428. https://doi.org/10.1080/00401706.1977.10489581
10. Box, George E. P. (1976), "Science and statistics" (PDF), *Journal of the American Statistical Association*, 71 (356): 791–799, doi:10.1080/01621459.1976.10480949.
11. Montgomery, C. D. Peck, A, P, and Vinning, G., Introduction to Linear Regression Analysis, Wiley

Appendix

```
> prediction_points
 [1]   1   4   7  24   3  19 105  10  12 133  40  73 141  21  15  92  14  78
42 137  74  59  17 100
[25]  64
```

Table 2.3.1: The points in the prediction set chosen by the DUPLEX algorithm

```
> estimation_points
  [1]  27   1  18   5  69   2   8  49   6 138   9 104  11 139  62 112  13 124
88  52  68  35  67 113
 [25] 107  16  20  22  23  25  26  28  29  30  31  32  33  34  36  37  38  39
41  43  44  45  46  47
 [49]  48  50  51  53  54  55  56  57  58  60  61  63  65  66  70  71  72  75
76  77  79  80  81  82
 [73]  83  84  85  86  87  89  90  91  93  94  95  96  97  98  99 101 102 103
106 108 109 110 111 114
 [97] 115 116 117 118 119 120 121 122 123 125 126 127 128 129 130 131 132 134
135 136 140
```

Table 2.3.2: The points in the estimation set chosen by the DUPLEX algorithm

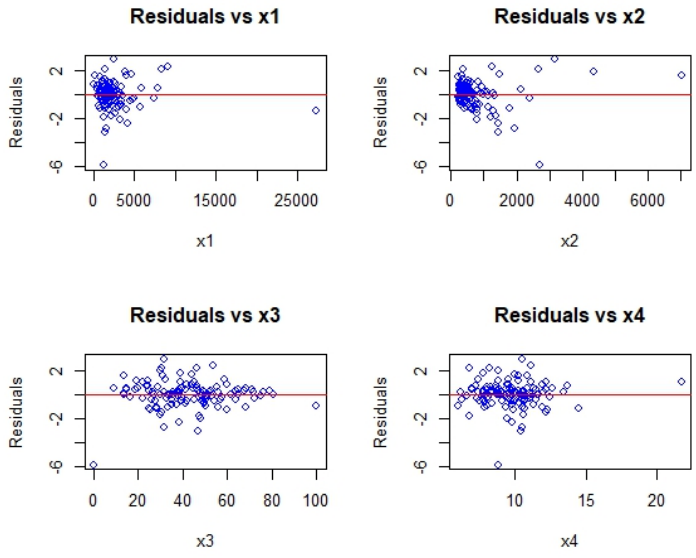

Figure 2.5.3: Plots of regressor vs residual plots for the initial model

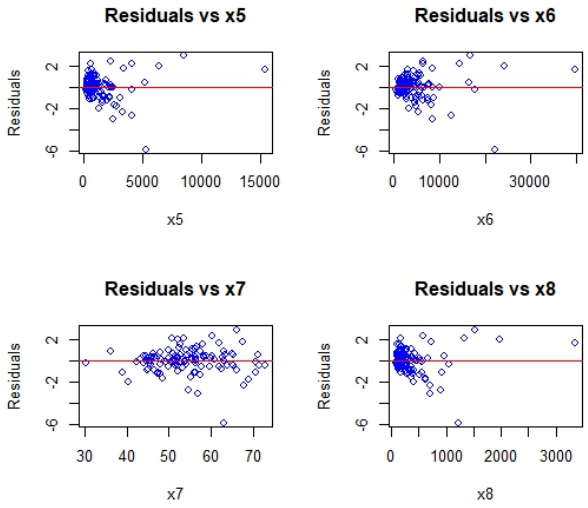

Figure 2.5.4: Plots of regressor vs residual plots for the initial model

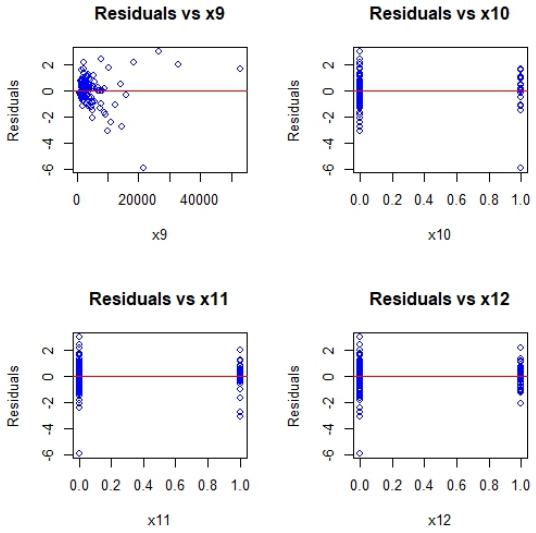

Figure 2.5.5: Plots of regressor vs residual plots for the initial model

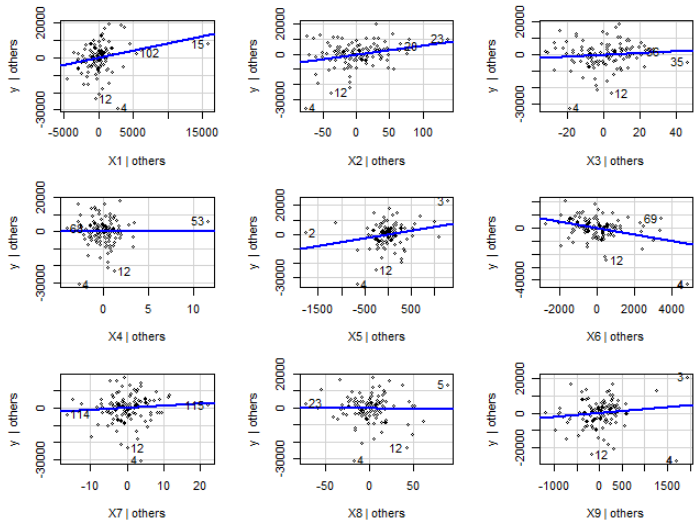

Figure 2.5.6: Partial regression plots for the initial model

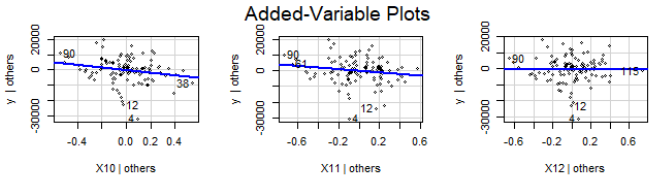

Figure 2.5.7: Partial regression plots for the initial model

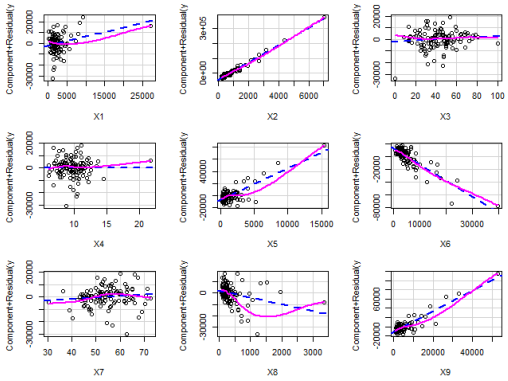

Figure 2.5.8: Partial residual plots for the initial model

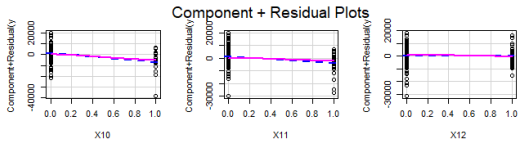

Figure 2.5.9: Partial residual plots for the initial model

Analysis of Variance Table

Response: transformed_y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
transformed_x1	1	0.0003007	0.0003007	76.0966	9.501e-14	***
transformed_x2	1	0.0050178	0.0050178	1269.6739	< 2.2e-16	***
transformed_x3	1	0.0001500	0.0001500	37.9449	1.790e-08	***
transformed_x4	1	0.0000502	0.0000502	12.7048	0.0005754	***
transformed_x5	1	0.0004978	0.0004978	125.9516	< 2.2e-16	***
transformed_x6	1	0.0001229	0.0001229	31.1007	2.342e-07	***
transformed_x7	1	0.0000066	0.0000066	1.6762	0.1985988	
transformed_x8	1	0.0001365	0.0001365	34.5499	6.290e-08	***
transformed_x9	1	0.0000251	0.0000251	6.3503	0.0134224	*
transformed_x11	1	0.0000002	0.0000002	0.0421	0.8379301	
transformed_x12	1	0.0000955	0.0000955	24.1578	3.750e-06	***
transformed_x10	1	0.0000768	0.0000768	19.4394	2.759e-05	***
transformed_x1:transformed_x10	1	0.0000284	0.0000284	7.1811	0.0086993	**
transformed_x2:transformed_x10	1	0.0000373	0.0000373	9.4313	0.0027894	**
transformed_x3:transformed_x10	1	0.0000198	0.0000198	5.0146	0.0274922	*
transformed_x4:transformed_x10	1	0.0000257	0.0000257	6.4955	0.0124352	*
transformed_x5:transformed_x10	1	0.0000427	0.0000427	10.7971	0.0014297	**
transformed_x6:transformed_x10	1	0.0000009	0.0000009	0.2210	0.6393368	
transformed_x7:transformed_x10	1	0.0000068	0.0000068	1.7241	0.1923614	
transformed_x8:transformed_x10	1	0.0000044	0.0000044	1.1149	0.2937355	
transformed_x9:transformed_x10	1	0.0000002	0.0000002	0.0594	0.8080480	
Residuals	94	0.0003715	0.0000040			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 2.7.1 ANOVA table to determine possible interactions between regressors and x_{10}

```
                                Analysis of Variance Table

                                 Response: transformed_y
                                        Df    Sum Sq   Mean Sq  F value    Pr(>F)
 transformed_x1                       1 0.0003007 0.0003007  57.4396 2.418e-11 ***
 transformed_x2                       1 0.0050178 0.0050178 958.3813 < 2.2e-16 ***
 transformed_x3                       1 0.0001500 0.0001500  28.6417 6.127e-07 ***
  transformed_x4                      1 0.0000502 0.0000502   9.5899  0.002579 **
 transformed_x5                       1 0.0004978 0.0004978  95.0714 6.234e-16 ***
 transformed_x6                       1 0.0001229 0.0001229  23.4756 4.975e-06 ***
   transformed_x7                       1 0.0000066 0.0000066   1.2653  0.263523
 transformed_x8                       1 0.0001365 0.0001365  26.0791 1.709e-06 ***
  transformed_x9                      1 0.0000251 0.0000251   4.7934  0.031045 *
 transformed_x10                      1 0.0001356 0.0001356  25.8950 1.842e-06 ***
  transformed_x12                     1 0.0000263 0.0000263   5.0209  0.027398 *
   transformed_x11                      1 0.0000106 0.0000106   2.0241  0.158128
   transformed_x1:transformed_x11       1 0.0000054 0.0000054   1.0268  0.313516
   transformed_x2:transformed_x11       1 0.0000026 0.0000026   0.4890  0.486100
   transformed_x3:transformed_x11       1 0.0000014 0.0000014   0.2605  0.610979
   transformed_x4:transformed_x11       1 0.0000000 0.0000000   0.0040  0.949753
   transformed_x5:transformed_x11       1 0.0000045 0.0000045   0.8512  0.358569
   transformed_x6:transformed_x11       1 0.0000002 0.0000002   0.0436  0.834994
  transformed_x7:transformed_x11        1 0.0000218 0.0000218   4.1703  0.043941 *
   transformed_x8:transformed_x11       1 0.0000027 0.0000027   0.5165  0.474101
   transformed_x9:transformed_x11       1 0.0000069 0.0000069   1.3237  0.252851
             Residuals                     94 0.0004922 0.0000052
                                          ---
         Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2.7.1 ANOVA table to determine possible interactions between regressors and x_{11}

```
Analysis of Variance Table

Response: transformed_y
                                  Df    Sum Sq   Mean Sq  F value    Pr(>F)
transformed_x1                     1 0.0003007 0.0003007  56.6744 3.080e-11 ***
transformed_x2                     1 0.0050178 0.0050178 945.6141 < 2.2e-16 ***
transformed_x3                     1 0.0001500 0.0001500  28.2602 7.128e-07 ***
transformed_x4                     1 0.0000502 0.0000502   9.4622  0.002747 **
transformed_x5                     1 0.0004978 0.0004978  93.8049 8.577e-16 ***
transformed_x6                     1 0.0001229 0.0001229  23.1629 5.667e-06 ***
transformed_x7                     1 0.0000066 0.0000066   1.2484  0.266707
transformed_x8                     1 0.0001365 0.0001365  25.7317 1.968e-06 ***
transformed_x9                     1 0.0000251 0.0000251   4.7295  0.032159 *
transformed_x10                    1 0.0001356 0.0001356  25.5500 2.119e-06 ***
transformed_x11                    1 0.0000351 0.0000351   6.6196  0.011651 *
transformed_x12                    1 0.0000018 0.0000018   0.3316  0.566102
transformed_x1:transformed_x12     1 0.0000006 0.0000006   0.1164  0.733733
transformed_x2:transformed_x12     1 0.0000008 0.0000008   0.1476  0.701690
```

```
transformed_x3:transformed_x12  1 0.0000014 0.0000014   0.2669  0.606661
transformed_x4:transformed_x12  1 0.0000005 0.0000005   0.0908  0.763798
transformed_x5:transformed_x12  1 0.0000151 0.0000151   2.8387  0.095335 .
transformed_x6:transformed_x12  1 0.0000039 0.0000039   0.7291  0.395338
transformed_x7:transformed_x12  1 0.0000107 0.0000107   2.0136  0.159205
transformed_x8:transformed_x12  1 0.0000059 0.0000059   1.1113  0.294499
transformed_x9:transformed_x12  1 0.0000000 0.0000000   0.0032  0.954693
Residuals                      94 0.0004988 0.0000053
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2.7.1 ANOVA table to determine possible interactions between regressors and x_{12}

```
> ols_coll_diag(final_transformed_model)
Tolerance and Variance Inflation Factor
------------------------------------
                       Variables   Tolerance        VIF
1                  transformed_x1 0.151671903   6.593179
2                  transformed_x2 0.004375779 228.530736
3                  transformed_x4 0.752865843   1.328258
4                  transformed_x5 0.060120184  16.633349
5                  transformed_x7 0.570728050   1.752148
6                  transformed_x8 0.074303236  13.458364
7                 transformed_x10 0.001303438 767.201643
8                  transformed_x9 0.004801993 208.246886
9                 transformed_x11 0.009836482 101.662365
10 transformed_x2:transformed_x10 0.021904894  45.651900
11 transformed_x1:transformed_x10 0.076044137  13.150258
12 transformed_x7:transformed_x11 0.010033727  99.663860
13 transformed_x5:transformed_x10 0.002275008 439.558975
14 transformed_x4:transformed_x10 0.002749395 363.716367
```

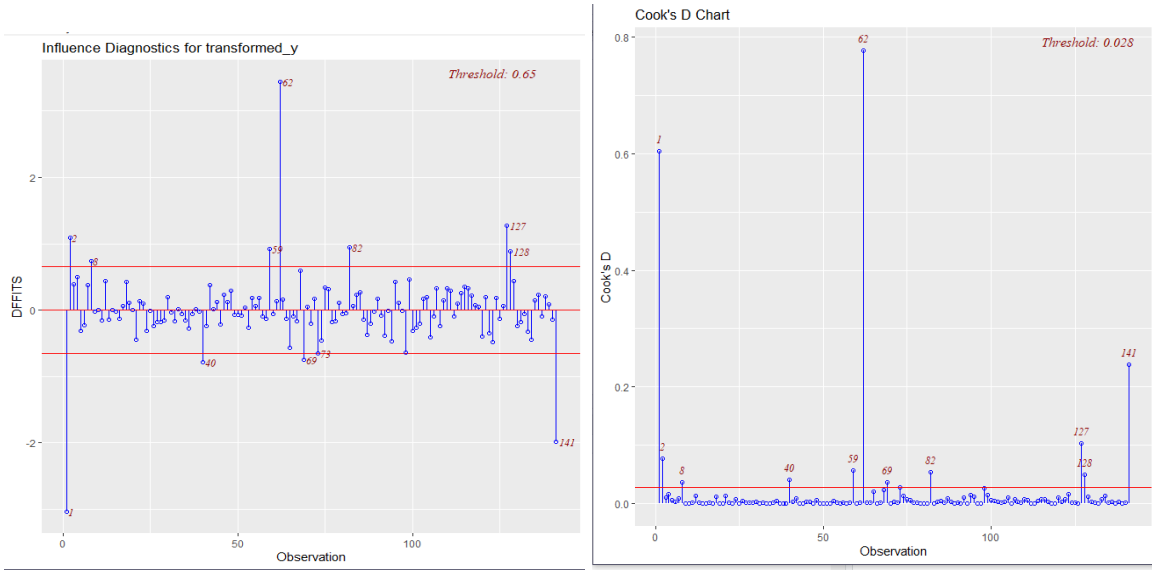

Figure:3.3.1 cook D's plot and DFFITS plots for influential points.

Figure 3.4.1 Table showing VIF and tolerance to determine multicollinearity

```
> confint(final_transformed_model)
                                        2.5 %        97.5 %
(Intercept)                      5.399055e-02  0.0902038345
transformed_x1                  -3.950716e-02  0.0880707968
transformed_x2                  -8.733553e-04 -0.0001329631
transformed_x4                  -1.491056e-03  0.0018973126
transformed_x5                  -1.896134e-03  0.0009344238
transformed_x7                  -1.833830e-04 -0.0000668633
transformed_x8                   1.540741e-01  0.2579718503
transformed_x10                 -4.846367e-03  0.0439735376
transformed_x9                   5.178769e-06  0.0002542177
transformed_x11                 -1.205285e-02  0.0041342665
transformed_x2:transformed_x10   4.216949e-05  0.0003742719
transformed_x1:transformed_x10  -1.651584e-01 -0.0241150487
transformed_x7:transformed_x11  -4.136628e-05  0.0002415743
transformed_x5:transformed_x10  -7.481933e-03 -0.0022566983
transformed_x4:transformed_x10  -7.393151e-04  0.0132046664
```

Figure 3.4.2 R output for the confidence interval for the betas.