Winter 2020 TMATH 410

Explaining Crime Rates

Sarah McKinlay, Kaitlyn Jones

March 20, 2020

1 Introduction

Unlike the laws of science, governmental laws are broken regularly. This negatively impacts the national economy, as well as the well-being of individual persons. No location or subgroup of humanity is immune; damages range from financial to physical, including loss of life.

For this project, we are interested in crime rates in the various regions of the United States and how they may be correlated with unemployment rates, population density, and per-pupil levels of government spending on K-12 education. We believe that these predictor variables may have a relationship with the changes in crime rates. During the research process we discovered that there are articles regarding crime rates and other types of explanatory variables such as alcohol abuse, income inequality, and deviant peer affiliations, but there are no articles that examine the exact variables and relationships that we are considering. However these articles also include multiple types of analysis that allow us to gain a broad perspective on the process.

To evaluate the effects of state and federal policy on crime rates, we have isolated four civil policy indicators as predictor variables. Each of these four predictor variables were meant to address a different facet of civil policy. The geographic region may indicate local cultural differences. Unemployment rates are an indicator of individual financial comfort levels. Population density has been correlated with many different social behaviors in the past, and is influenced heavily by municipal codes and legislation on zoning and housing projects. Lastly, we consider annual pre-college educational spending, per-student, for a focus on the level of support that a student receives before entering the workforce.

The objective of our project is to determine which, if any, of these factors play a role in the direction and level of change of crime rates. We use each state in the US as an individual sample, and take the average values as reported by government resources for each predictor variable. We would like to note that data from Washington D.C. has been removed from the dataset and is not factored into this project. Due to the the fact that Washington D.C. is not an actual state, along with the small land area and high population that were skewing the plots of the data we found it best to remove it altogether. This decision did not affect the model. Should we find correlations that warrant further study, we hope to realize findings that could speak to changing civil policies in order to minimize crime rates.

We wish to study the effects of civil policies on crime, as measured by the yearly total crime rate per thousand people in each state. We have not seen any previous research that attempts to correlate population density and per-student spending simultaneously with crime rate outcomes, but it should be noted that ”Research in the fields of criminology and economics suggests that inequitable allocations of resources can incite criminal activity.” (Brush 2007) This observation means that there may be several other variables that all contribute to crime rates, and many that may be confounding variables with one another. All data was able to be collected through official government websites and is assumed to be reliable and accurate. We expect that the data will show a relationship between population density, per-student spending, and yearly percentage change of total crime.

2 Methods

The data for this project was collected online from the US Census website and the Federal Bureau of Investigation website. Most of the data was usable as is, but we did generate a column for population density for each state, calculated as the ratio of state population to state land area. We also created a column for crimes per 1000 people, from the state population and the total crimes committed per state.

Our exploratory data analysis consisted of multiple scatterplots and boxplots, summaries of the response and explanatory variables, and an interpretation of these items. The scatterplots showed the quantitative explanatory variables plotted against the response variable and then each of the explanatory variables plotted against each other. The boxplots show our qualitative explanatory variable plotted against our response variable and then plotted against the other explanatory variables.

We are fitting a multiple linear regression model. We chose to start by analyzing the plain data values, rather than converting any of them to a logarithm. We selected our variables based on a lack of existing analyses that factor in educational spending. We then tested our model, with all possible factors, and examined it with some variables omitted to find the best possible fit based on our evaluation. In the end we kept all the original variables in the model. The final model took the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ where the coefficients β_n are: β_0: the intercept, β_1: population density, β_2: annual education spending, β_3: region, and lastly β_4: unemployment rate.

During the research phase we came across a type of analysis called Geographically Weighted Regression(GWR), which ”...estimates a local model, producing a set of mappable parameter estimates and t-values of significance that vary over space.” (Cahill and Mulligan 2007) In the article, it was used in addition to traditional methods of regression to compare findings. While this was not a method we were able to use at this stage in the project, it is of great interest to us as a future method that could be used on this model.

We assumed that the data collection would be in an unbiased fashion. We also assumed that sampling errors in the data would be independent and identically distributed. We verified that the data was accurate by selecting reliable sources – the 2010 census data, especially, was obtained with an incredible number of individual, on-site samples taken to form the statistics for each state. The data gathering methods are uniform. The crime data is self-reported by agencies throughout the United States, but we chose to consider that a relatively unbiased source as well. To verify that our errors are independent and identically distributed, we examined the residuals plots for significant patterns to the contrary. Residual diagnostics were also used to assess normality, linearity, constant variance, and influential observations. We also used the VIF (variance inflation factor) to detect whether or not we had severe multicollinearity. Lastly, we inspected the Cook’s distances to reveal any highly influential data points.

3 Results

After performing the aforementioned tests and analysis we found that there is some strong individual correlation between some of the explanatory variables and the response variable. Specifically annual amount spent on education per-pupil and region, the corresponding r values from the linear model being .517 and .726, respectively. The full model for the multiple linear regression shows a strong correlation between the explanatory variables and the response variable, with an r value of .831. Shown below in Figure 1 are the plots for our variables. From these plots we can visually assess any possible relationships. Also seen below in Table 1 are the summaries for each variable.

Figure 1. Shows the scatterplots for our response variable and each quantitative explanatory variable. Also shown below is the boxplot for our response variable and our single qualitative explanatory variable.

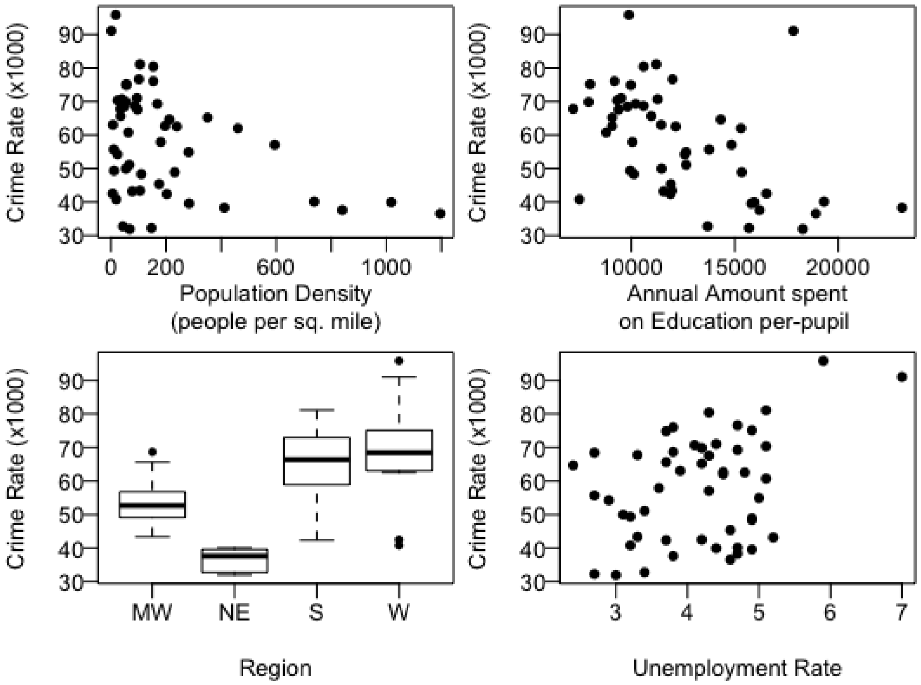
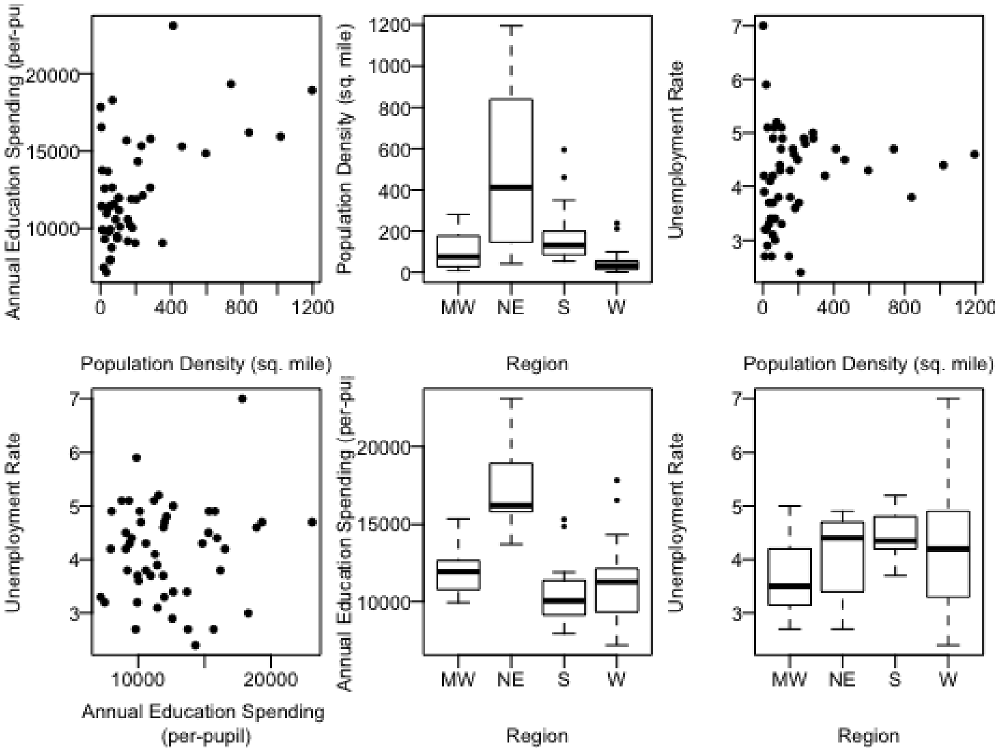


Table 1. Shown for the full model are the values for the regression coefficients, their standard errors, p-values for hypothesis tests, and conclusions. Note that the full linear model can obscure significance of certain variables if they have relationships with other variables.

Coefficient	Estimate	Standard Error	P-Value	Conclusion
Intercept	4.594e+1	9.559e+0	1.82e-5	Significant
Population Density	5.999e-3	1.387e-3	8.63e-5	Not significant
Per-pupil Spending	-9.068e-4	6.762e-4	1.868e-1	Not significant
RegionNE	-1.644e+1	5.739e+0	6.36e-3	Significant
RegionS	5.372e+0	4.539e+0	2.4296e-1	Not significant
RegionW	1.118e+1	4.504e+0	1.693e-2	Significant
Unemployment Rate	4.901e+0	1.893e+0	1.302e-2	Significant

During the analysis phase we also observed possible correlation between our explanatory variables. The results indicated a strong correlation between annual amount spent on education per-pupil and region with an r value of .717. Following closely behind is the correlation between population density and region with an r value of .629 and lastly, the correlation between population density and annual amount spent on education per-pupil with an r value of .549. The plots for these relationships can be seen below as well as the other less correlated variables.

Figure 2. Shows the scatterplots and boxplots with our explanatory variables plotted against each other. The boxplots correlate to our qualitative explanatory variable.

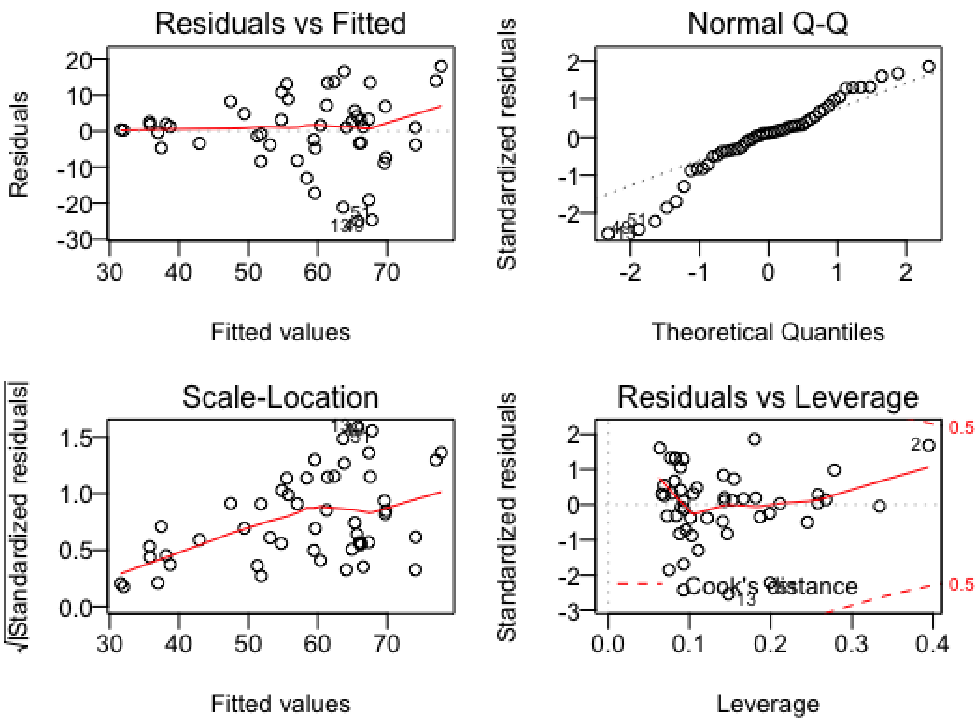


We used the VIF (Variance Inflation Factor) to test for multicollinearity and found that none of the variables have an individual VIF > 10. The mean VIF across all predictors is 1.822 (which is less than 2), and therefore we do not detect any simultaneous multicollinearity. After running the residual diagnostics, seen below in Figure 3, we were able to assess our model and data more clearly by examining linearity, constant variance, normality, and influential observations.

One method of analysis that supports our decision to omit Washington D.C. as an outlier is to observe that it exhibits high leverage simultaneously with high influence. The Cook's Distance for D.C. is greater than 1, which indicates that it will skew our results. Because D.C. is not a state, but rather a single city, it will have a disproportionately high population density.

At first, we were concerned that California may be another outlier; it is held up by politicians throughout the United States as a high-crime center, despite its high social spending. While we did examine it carefully, the Cook's Distance for California fell well within our comfortable range, and below .5. Thus, we feel entirely confident in retaining California in our model, and not only because it is an actual state instead of a city.

Figure 3. Shows the residual diagnostics performed on the data to assess normality, linearity, constant variance, and possible influential observations.



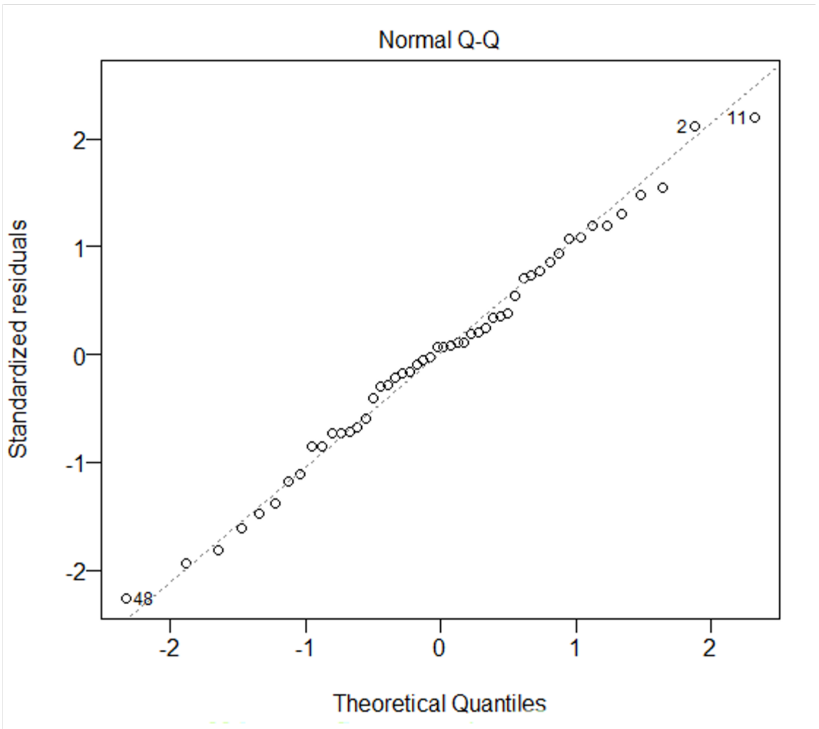
In the Residuals vs Fitted plot we see a violation of the constant variance assumption and a slight violation of the linearity assumption. In the Normal Q-Q plot we notice the tails that trail away from the dashed line, indicating a skewed distribution. Transformations were performed on the model in an attempt to fix the linearity and normality issues with some success. We used a log transformation to adjust linearity, and experimented with removing region from the model to fix our normality.

All graphs are completed with D.C., and US territories omitted. All US states were included in our data sample. While the VIF for the regional and per-pupil spending predictors was low, which eased our minds about collinearity, we still see the potential for them to confound one another. One consideration is that the cost of living in the NE and W regions is higher, and thus, per-pupil spending amounts are diluted by the actual purchasing power each dollar possesses. Below in Table 2 we see the summary table for the linear model without region included. Also below in Figure 4, we see the Normal Q-Q plot of the residuals without region included.

Table 2. Shown for the full model, without region included, are the values for the regression coefficients, their standard errors, p-values for hypothesis tests, and conclusions.

Coefficient	Estimate	Standard Error	P-Value	Conclusion
Intercept	5.356e+1	10.09e+0	3.09e-6	Significant
Population Density	1.239e-2	7.78e-3	1.18e-1	Not significant
Per-pupil Spending	-2.0515e-3	5.855e-4	1.03e-3	Significant
Unemployment Rate	7.7262e+0	1.900e+0	1.85e-4	Significant

Figure 4. Shows the Normal Q-Q plot of the standardized residuals plotted against the theoretical quantities with region removed.



4 Discussion

In conclusion, we see that, while educators may prefer the more parsimonious model, it is eclipsed by the model that accounts for the percentage of crime rate that is predicted by other variables. In other words, crime rates have a significant correlation to the predictor values of geographic region and unemployment rate. They also appear to have a standard ”baseline” rate of crime, regardless of external factors.

It is important to note that there are many other societal factors that may affect crime rate that were not studied during this project. A future direction could be to examine these other variables either in tandem with the current variables examined in this project, or as part of a completely new set of predictor variables.

One could go even further with this project by following in the footsteps of our sources with a more localized approach. As this model was based on data per state, a future model could be based off of city or even county lines which would allow for a closer look at the small details, interactions, and influential factors.

Some of the key questions that we would like to propose for further study are intended to untangle the relationship between educational spending and region, as well as create a more granular idea of what each region's differences may be.

5 Questions For Further Study

- ○ Are the effects of educational spending mediated by cost-of-living? Purchasing power may influence the benefit actually received by students.
- ○ What results would be obtained by applying Geographically-Weighted Regression? Given the crucial role that the geographic region played in our outcome, this method may be particularly fruitful to explore.
- ○ What other factors go into a geographic region that influence the crime rate? Policy differences may be one difference, but access to a frequently-travelled port, local business types, and even the climate may all play a role.
- ○ What happens when we examine data sourced after 2010? The 2020 census provides a golden opportunity for statisticians to find out what may be different from the previous decade.

6 Sources

FBI - Table 2. [Internet]. c2017. Washington (DC): Criminal Justice Information Services Division; [cited 2020 Feb 5]. Available from: https://ucr.fbi.gov/crime-in-the-u.s/2019/preliminary-report/tables/table-2/table-2.xls/@@template-layout-view?override-view=data-declaration.

Bureau of Justice Statistics (BJS) - National Crime Victimization Survey. [Internet]. 2018. Washington (DC): Bureau of Justice Statistics; [cited 2020 5 Feb].
Available from: https://www.bjs.gov/index.cfm?ty=dcdetail&iid=245#.

Brush J. 2007. Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties. ScienceDirect. Economics Letters 96(2): 264-268.

Cahill M, Mulligan G. 2007. Using Geographically Weighted Regression to Explore Local Crime Patterns. Social Science Computer Review 25(174).

United States Census Bureau. [Internet]. c2011. Decennial Census Datasets. Available from: https://www.census.gov/programs-surveys/decennial-census/data/datasets.2010.html.