Linear Modeling

Department of Applied and Computational Mathematics & Statistics

The University of Notre Dame

Final Project Report

Criminal Activity and Predictive Variables

Stephen Megos

Linear Models

Professor Alan Huebner

May 12, 2021

**Introduction**

The purpose of this study is to analyze several criminal acts across several communities, as well as analyze variables that can be used to predict the likelihood of the same criminal acts. The data being used is a ‘CommunitiesAndViolence’ text file which has been provided. The set includes several different types of criminal acts and information about the location and community that the act took place in. For the study to take place, two types of variables were identified: four response variables, and one predictive variable. The study involves using the predictive variables to predict the likelihood of the response variable to take place. In this study, the four predictive variables are: ‘Population’ of a community, the ‘Median Income’ of the community, the ‘Number of Homeless Persons’ in the community, and the ‘Racial Match’ of the community. The response variable being studied is ‘Burglaries.’

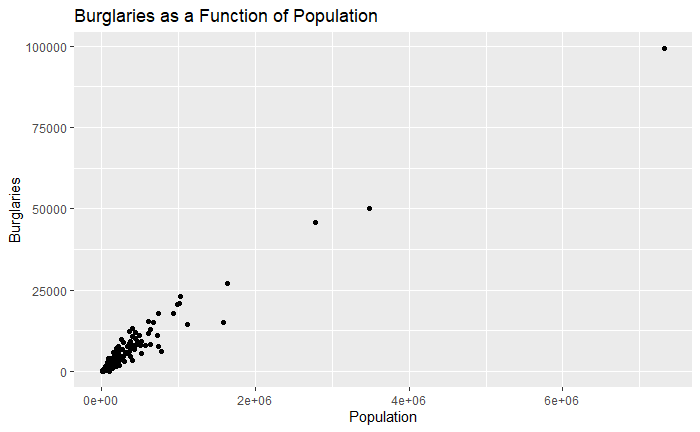
**Step I: Exploratory Data Analysis**

To conduct any kind of analysis on a particular set of data, the data needs to be ‘tidy’ and ‘clean’ to be used efficiently throughout the analysis. This means that the data should not contain any empty or ‘NA’ values that could harm or negatively influence the effectiveness of the model. The first step in this analysis was first to examine the data set itself. Immediately it was determined that most, if not all the values were being stored as character strings instead of numerical values. When a function looking for ‘NA’ (empty) values was run, the function returned ‘0’ indicating there were no ‘NA’ values. Each of the variables that was set to be used in the data set, was then converted from character type to numeric type, then any rows that contained any newly formed ‘NAs’ were deleted leaving only useable numeric values.

By viewing the relationship of the raw data variables and comparing them to the response variable, and ultimately each other, hypotheses were formed regarding the predictiveness of the predictive variables on the response. Plotting each of the predictive variables against the response variable provided some insight into the kind of relationship the values had with each other and helped establish some context for the analyses to follow. It was also noted that there could have been multicollinearity present in the model, and multiple variables could have had similar relationships with each other and then appeared to have a similar relationship with the response variable. For example, a community with a large overall population is more likely to have a large population of homeless people and communities with a large overall population are intuitively more likely to have many burglaries. Therefore, many homeless people (which is a result of the high population) could be seen as being a predictor of burglaries even the homelessness does not directly predict burglaries.

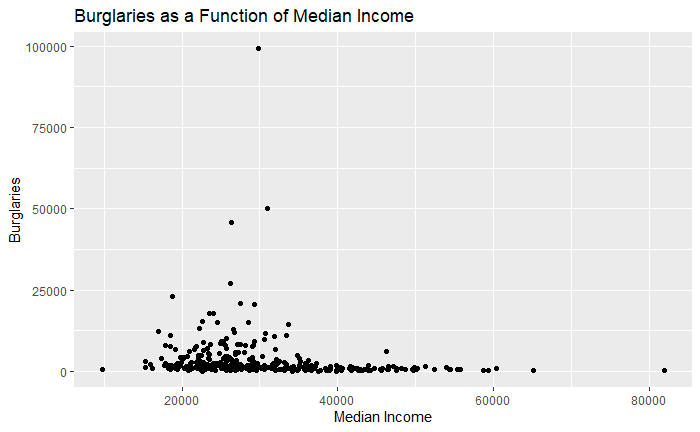
Variable 1: Burglaries as a Function of Population

Judging from the ‘Burglaries as a Function of Population’ plot, the two variables have a clear positive correlation. As the population of a community increases, the number of Burglaries increases as well. This makes sense intuitively and matches with the pre-determined hypothesis. Since burglaries and population have a linear relationship, their relationship can be used to compare and reference other variable relationships later in the analyses.



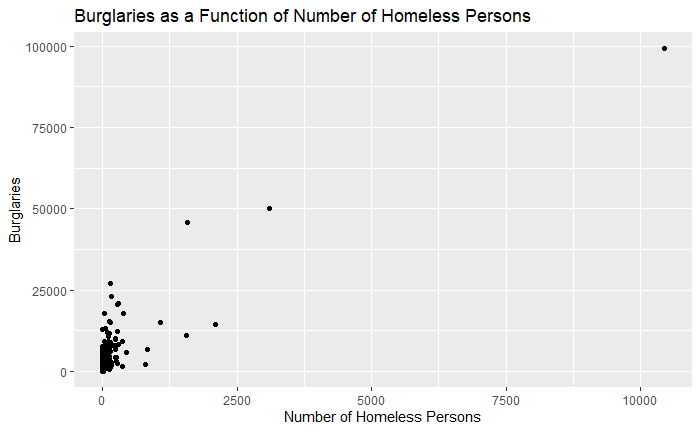
Variable 2: Burglaries as a function of Median Income

Judging from the ‘Burglaries as a Function of Median Income’ plot, the two variables also have a somewhat positive correlation. There could potentially be outliers, but it is somewhat difficult to tell at first glance considering the high concentration of points in one area of the plot.



Variable 3: Burglaries as a function of Homeless Persons

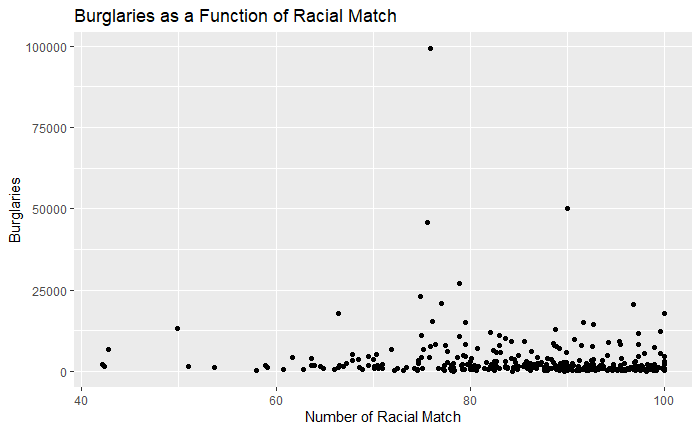
Judging from the ‘Burglaries as a Function of Number of Homeless Persons’ plot, the two variables have a positive correlation. Although there is one value that appears to be an outlier pulling the perspective of the graph to a wider range. In the initial exploratory data analysis portion, it is important to leave all values in the dataset untouched to see all the values and their relationships, without eliminating any outliers or unusual looking values. This variable could be a candidate for multicollinearity with its respect to population, considering the close relationship that number of homeless people has with total population. However, that hypothesis could be unproven considering the lack of a clear relationship between homelessness and burglaries in the same way there is a relationship between population and burglaries.



Variable 4: Burglaries as a Function of Racial Match

Racial match is defined as the ratio of the racial makeup of the police force, relative to the ratio of the racial makeup of the community (i.e., a community with 50% white people, and 50% black people, that has a police force with 50% white officers, and 50% black officers, would have a Racial Match value of 100).

Judging from the ‘Burglaries as a Function of Racial Match’ plot, the two variables seem to have a positive correlation. Although there is a greater concentration of points on the right side of the graph, there does seem to be an upward trend in the points as the Racial Match value reaches 100.



Of the four predictive variables, each variable appears to have a positive or relationship with burglaries to varying degrees. The predetermined hypotheses are upheld for the most part after the initial glance at the data and their relationships with each response variable, but there does not seem to be as close of a relationship between population and burglaries, and homelessness and burglaries as previously speculated.

**Fitting a Linear Model**

Using the pre-determined response and predictor variables that were identified, cleaned, and analyzed, a linear model was fit. Burglaries was set as the response, and the other four were set as the predictor variables. The summary output in R provided details regarding the predictive power of the different variables, as well as gave an idea of the accuracy of the model itself.

The Multiple R-squared value of the model was 0.96 (scale of 0-100), and the Adjusted R-squared was 0.9595. The p-value of the model was 2.2 x 10-16 and the p-values of each predictive variable showed population to be the most predictive variable and racial match to be the least predictive.

Judging from the summary output of the model, racial match is not as critical of a value when determining the number of burglaries in a community relative to the other variables in the same model. Other than population, the number of homeless persons appears to be the most critical of the remaining three predictors. This upholds the original hypothesis claim that number of homeless persons and total populations will have similar predictive power over the model. The predictive power of racial match upholds the observations made in the EDA, as it does not seem to have as much influence in the model as the other three variables.

**Step III: Performing Model Selection**

Once the original model was formed and summary output was analyzed and interpreted, steps needed to be taken to discern how predictive each variable was relative to the response. Using packages available in R, the model can be analyzed, and certain variables might need to be removed to improve the overall predictive power of the model. Having variables included in a model that are not good predictors of the response can decrease the accuracy of the model by overly complicating it. This can especially be the case with a large quantity of predictive variables.

The ‘fastbw()’ function and ‘stepAIC()’ function available in the rms and MASS packages respectively offer a way to discern which variables are valuable to the integrity of the model and which are a hindrance to its effectiveness.

Using hypothesis testing and automated backward selection, ‘fastbw()’ can help determine a more efficient model, which would include having fewer predictive variables than the original. After running the function on the original model, the output shows that none of the variables should be used and that all four of our original predictors should be left in the model to maintain its predictiveness and efficiency.

The ‘stepAIC()’ functions performs an iterative process and fits multiple models, and compares their AIC (Akaike Information Criterion) values. StepAIC() removes a single predictive variable, refits a linear model, and then extracts and compares the AIC values to the previous model’s AIC until the efficiency of the model no longer improves. After running ‘stepAIC()’ on the original model, the final output of the function shows, like the ‘fastbw()’ function, that the original model is the most efficient. The function immediately returns a higher AIC value and therefore terminated its iteration.

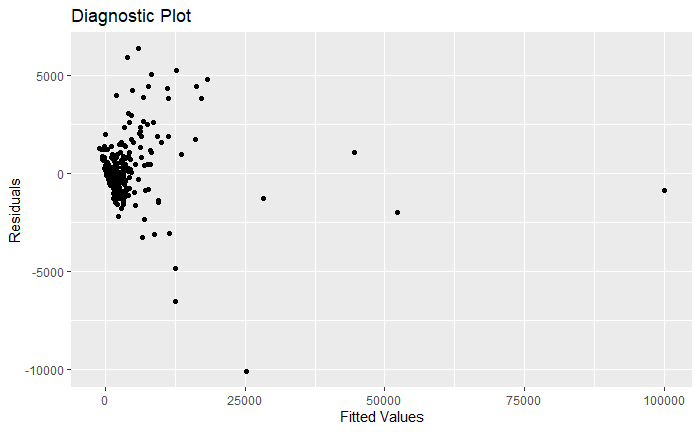
The two functions sharing a result is a strong indication that the model is as efficient as possible. If the models returned a different, and more efficient model, then that new model would replace the old and the next steps in the study would resume using the newer model.

**Step IV: Apply Diagnostics to the Linear Model**

Performing Model Diagnostics helps understand the output of the model. By running multiple diagnostic tests on the final model derived after performing model selection, information regarding the accuracy and variance of errors of the model can be accurately identified. The diagnostic tests in this study included creating a diagnostic plot, Q-Q plot, and lagged residual plot. These plots can be used to determine whether the model upholds the assumption of constant variance among errors.

Diagnostic Plot

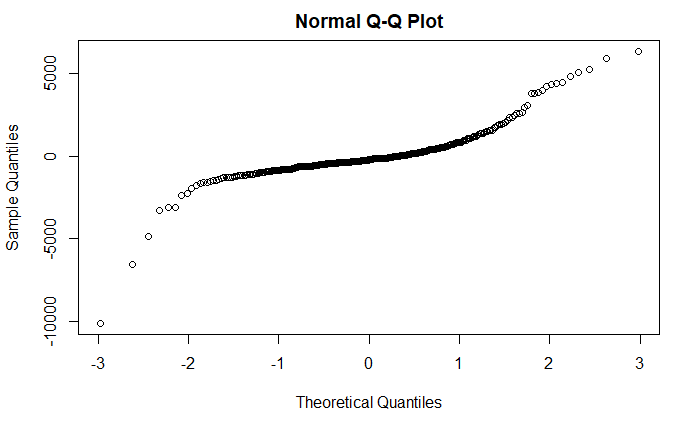
A basic diagnostic plot involves plotting the final model’s residuals against the same model’s fitted values. This plot provides perspective on the assumption of constant error variance. Ideally, a model’s diagnostic plot will show a random pattern amongst its points which will uphold the error of constant error variance assumption. For the final model in the study, the residuals were plotted on the y-axis, and the fitted values were plotted on the x-axis. The plot failed to uphold the assumption as most of the points were very clearly gathered in almost a ‘cone’ shape.



Q-Q Plot

Another diagnostic test available is the Q-Q test which can be executed using the ‘qqnorm()’ function available in the ‘stats’ package in base R. The Q-Q plot plots the residual’s sample quantiles on the y-axis, and the x-axis displays the quantiles under the assumption that the data came from a normal distribution. Using this normal distribution assumption, if the line formed by the points on the plot forms a linear pattern, this will uphold the assumption of normal errors.

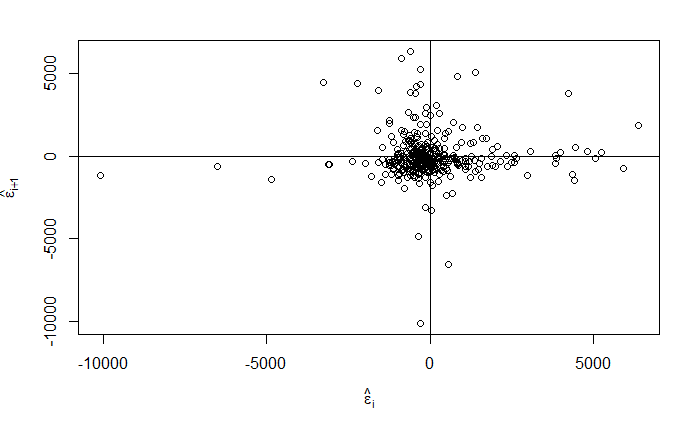
In the Q-Q plot executed for the final model, the Q-Q plot very clearly forms a non-linear pattern, thus failing to uphold the assumption of normal error variance. This is consistent with the assumption formed after analyzing the results of the diagnostic plot.



Lagged Residual Plot

The lagged residual plot is a plot formed by plotting each residual value against the successive residual. If the errors are scattered, this shows error independence and would uphold the assumption of distributed error variance.

In the plot used for this analysis, the residual is plotted on the x axis, and the following residual is plotted against the y-axis. The points in the following plot are very clearly centered around a similar point which, like the other two diagnostic tests, fails to uphold its assumption.



**Step V: Investigate Fit for Individual Observations**

Understanding if any specific points are skewing the results of a model can be done by searching for outliers and influential points in a model. By using the ‘rstandard()’ function, each output’s standardized residual can be determined, the closer to 0 a value is, the less of an outlier it is. The further from 0, regardless of whether the value is positive or negative, the more of an outlier it is. Although there is no definite, mathematically derived threshold, a rule of thumb that is generally accepted by statisticians is that if the absolute value of a standardized residual is greater than three, then it is considered an outlier.

Understanding the influence that a point has on a model is also necessary to determine whether certain points should be discarded to increase the effectiveness of a model. By using the ‘cook.distance()’ function available in the ‘car’ package in base R the influence of a point can be derived by comparing the value to a derived threshold. Some analysts, compare the cook’s distance value of points in a model to 1. If the cook’s distance value is greater than 1, than one can assume that the point is an influential point. To find a more precise threshold relative to the model being assessed, an f-statistic threshold can be derived in R using the size of the original data set, the degrees of freedom of the set, and the 50th quantile of an f-distribution for that model. In this case the ‘F Threshold’ is derived and identified as 0.872.

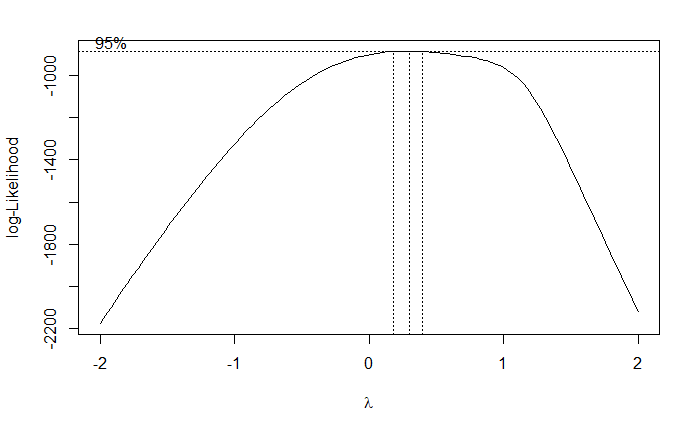
Out of approximately 340 values, only ten are considered outliers after analyzing the residual standards. After deriving the cook’s distances of the model, only one point is determined to be considered an influential point. Although in many instances, it is acceptable to remove or even adjust the outliers to make the outputs more predictive, in this case because of the small number of outliers, and the even smaller number of influential points, the outliers and influential point was left in the model to provide a more wholistic view of the model output.

**Step VI: Applying Transformations to the Model as Needed**

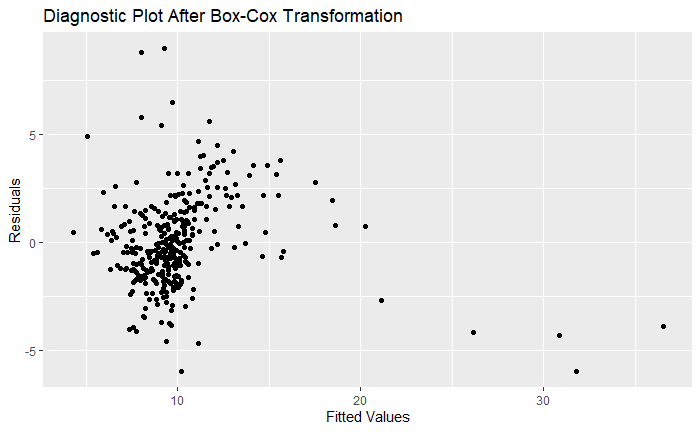
After analyzing the results of the mathematical assumptions in step IV, it is necessary to appropriately transform the model to account for the lack of distribution of errors. To perform the transformation, the results of the diagnostics tests need to be reviewed. All three of the tests concluded that there was no appropriate distribution of errors present in the output of the model.

A Box-Cox plot can be created to determine the lambda of the model which is needed to determine which kind of transformation needs to take place to correct the model. The ‘boxcox()’ function is available in the MASS package. The Box-Cox method is a more objective way of determining the transformation method necessary since it requires less individual interpretation. The plot shown below is a normal distribution-like pattern with three dotted lines. The outermost dotted lines show the 95% confidence interval of the optimal value of lambda. The centermost line is the most optimal value.

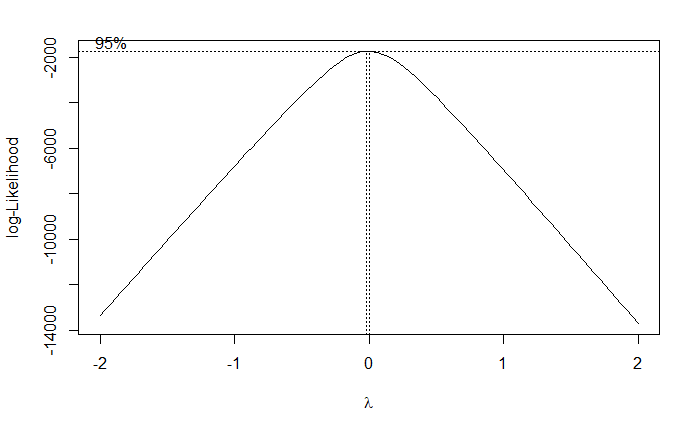
Since the 95% range of lambda does not include 0, the transformation includes taking the final model, assigning the lambda to the exponent of the response variable, and then rerunning the model to apply the transformation.



Once the transformation is applied, a new diagnostic plot can be applied with the new model’s residuals and fitted values, and then it can be compared to the original. In this case, the diagnostic plot does show a more spread-out distribution of errors, relative to the first plot.



The new box-cox plot also shows a narrower 95% interval which also now includes 0.



**Part VII: Report Inferences and Make Predictions Using the Final Model**

Once all the necessary analyses, diagnostics, and transformations have been concluded, the results of the model can be derived. The coefficients and final p-values are all shown and seem to have become more precise than the original model first reported except for racial match. Although racial match’s p-value is still less than 0.05, it is reassuring that it remains as the least predictive quality of determining burglaries.

The R-squared of the new model is lower than that of the original plot, indicating at face-value that it is now less predictive than the original model. However, a lower R-squared value is not always worse. Since an R-squared value is calculated by dividing, “the variance explained by the model” and “the total variance” it is important to analyze the diagnostic plots and the rest of the model summary output to determine the quality of the model.

A determined prediction variable was used to compute a 95% confidence interval for the model by using the variable’s estimate value, and the standard error value. Because median income was one of the more interesting variables and was still a strong predictor, it was used as an indicator for the confidence interval.

**Conclusion**

Using the medians of the prediction variables as predictors, the model predicts with 95% confidence that a community with 94,886 people, a median income of $27,295.50, seven homeless people, and a racial match of police to the community of 87.9%, to experience one burglary in a year. The same model is confident that all communities with the same attributes are likely to experience approximately one burglary per year.

In the final step of the study, after using the model to test potential indicator variables to predict the likelihood of a burglary to take place, the model does not appear to be exceptionally effective. Initially, the transformation step was assumed to be the most likely location for a user error to have taken place; however, it appears that was not the case. After revising the original data again, it appears that the large number of communities with low populations and low numbers of homeless people, caused there to be a lack of an even distribution of variables. This forced there to be a heavy skew, where the number of burglaries taken place in a community was artificially low. This was surprising, since it appeared in the data that there could have been a heavier skew towards the other side. This proves the importance of carefully examining the data and making sure that the data set is not only clean and usable, but also contains a distribution of data values.