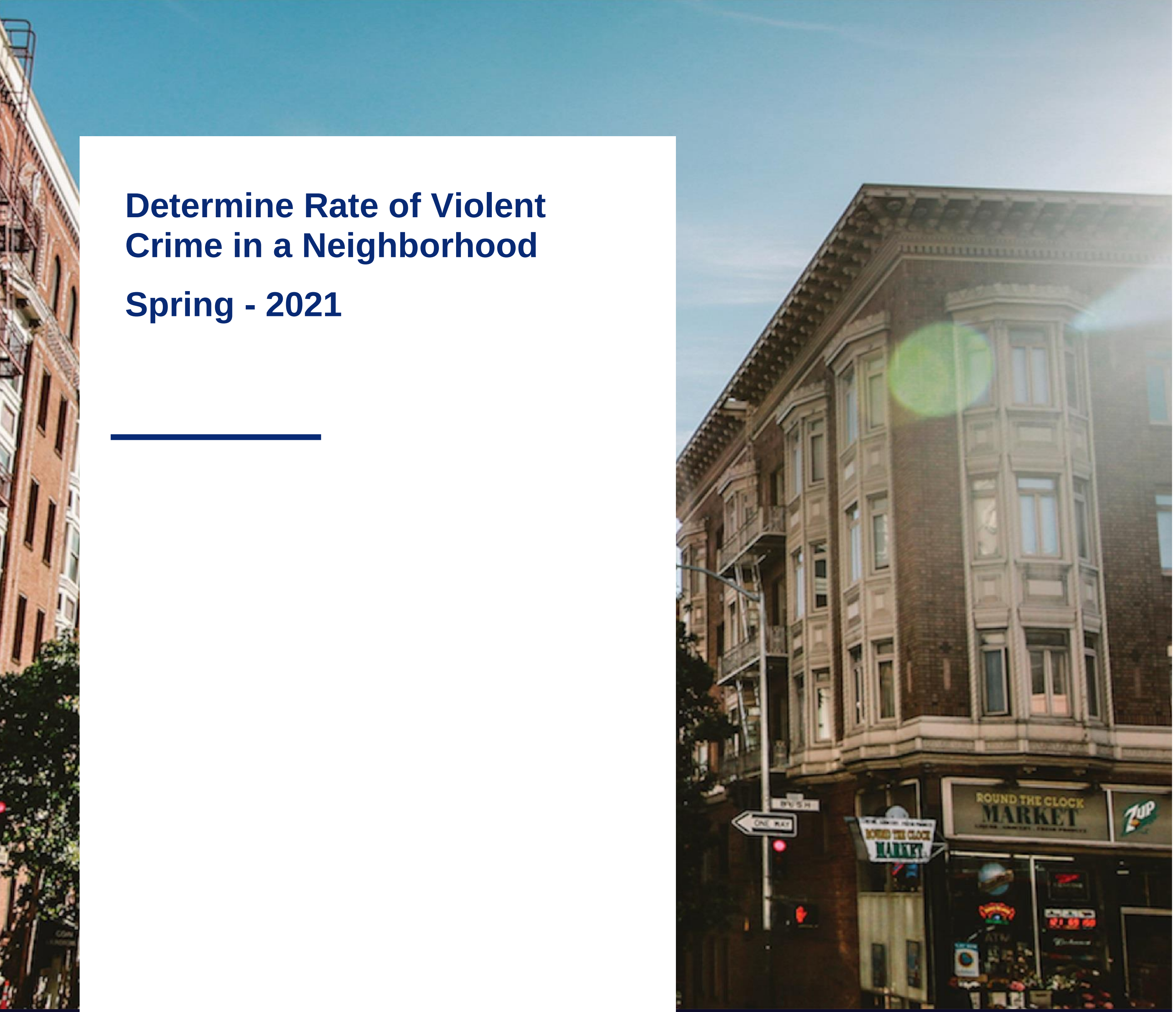


Determine Rate of Violent Crime in a Neighborhood

Spring - 2021

MAY 12

Notre Dame MS - ACMS
Authored by: Jonathan Gragg

Predicting # of Violent Crimes

Introduction

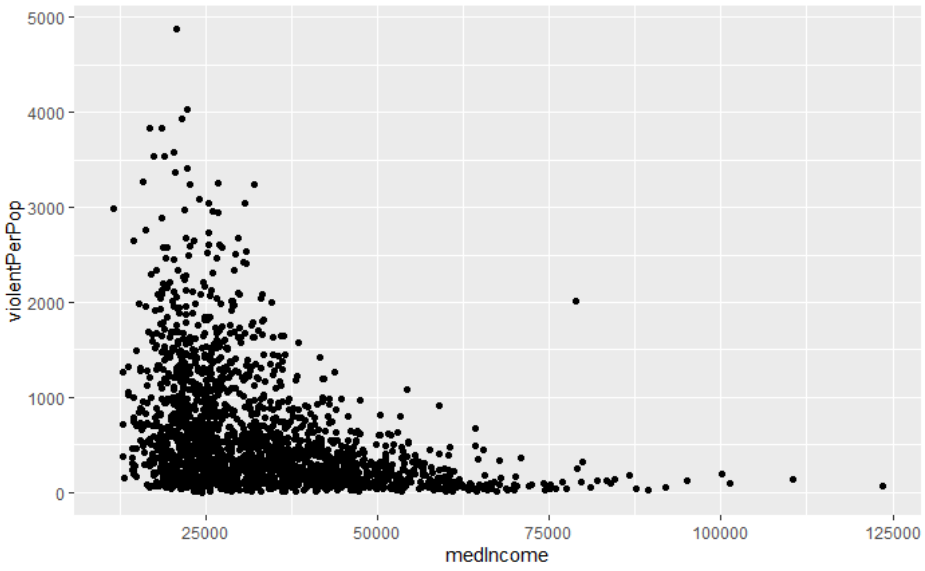
This evaluation was started with the goal of writing a model to predict the number of violent crimes per 100,000 people. This began with exploring the *Communities and Crime* data set. This data set is a large collection of data coming from the 1990's census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats, and finally crime data from the 1995 FBI UCR. With these three data sets we can determine demographic data about individual communities, any law enforcement statistics from those communities and finally the number of different levels of crime for each of these communities.

To predict number of violent crimes we can create a linear model with number of crimes per 100,000 as the response variable. For this model to be as accurate as possible we will need to make sure that we are picking the correct variables that have the largest impact on our response variable.

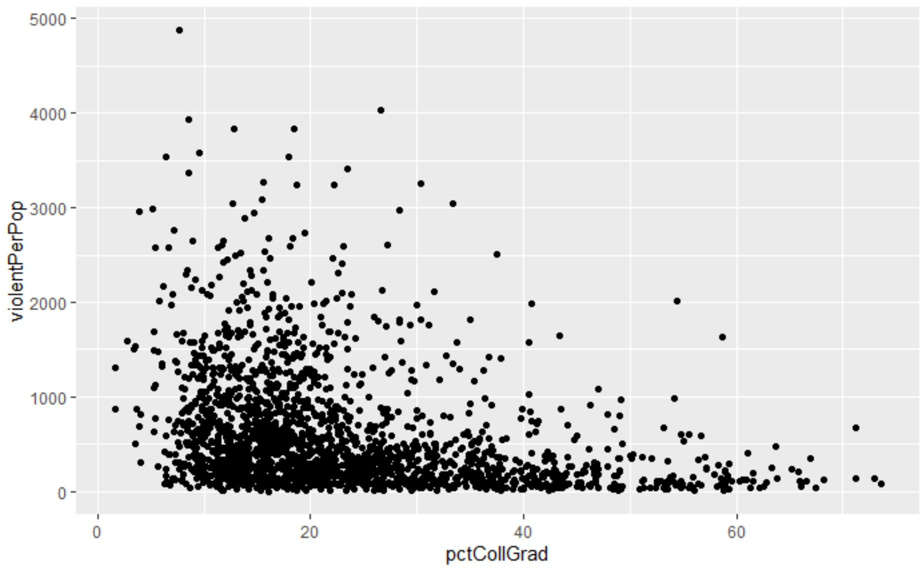
This analysis will go into depth of the initial exploratory analysis and dive deeper into fitting an accurate model to the data in hopes of creating an algorithm that will predict the most accurate outcome for the general population.

Exploratory Data Analysis (EDA)

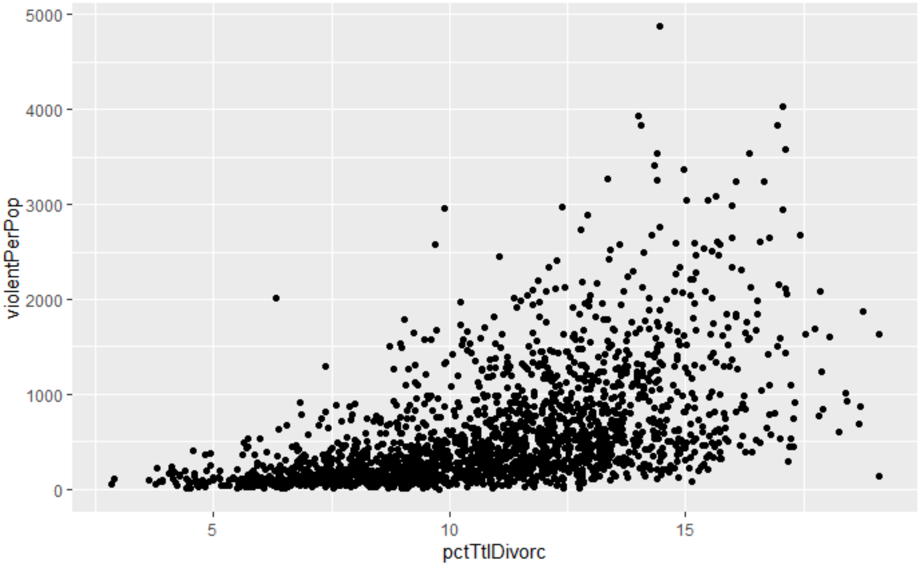
First step in exploring what factors may be correlated with response variable. The fastest and easiest to determine an immediate correlation is a scatterplot with the response variable as the y axis and the variable as the x axis. When the two variables are compared a correlation will be present if there is an easily noticeable pattern in the points. Here are 4 examples of variables that showed strong correlation with our response variable (violent crime per 100,000)



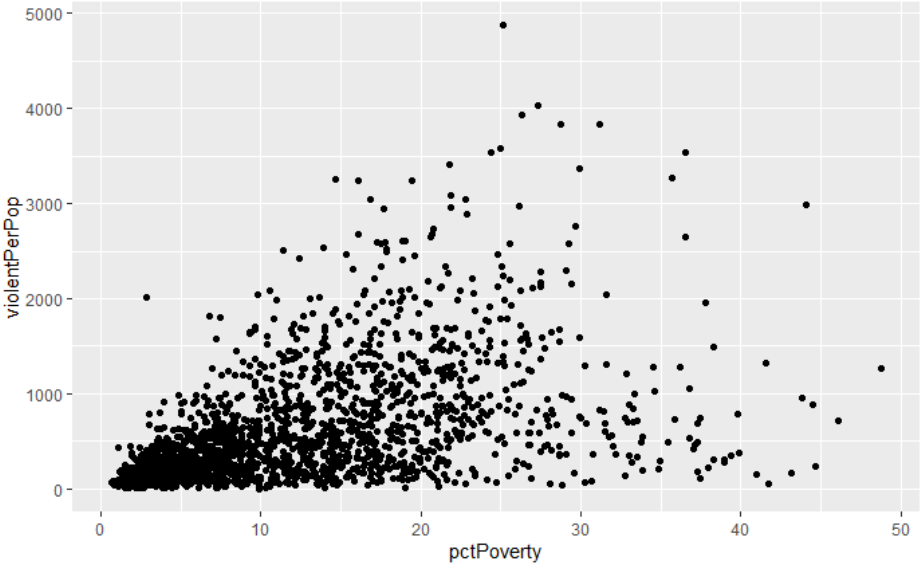
Above we see how violent crimes per 100,000 people compares to the median income in a community. You can see that there is an inverse correlation between the two variables leading the assumption that as median income increases in a neighborhood the less violent crimes are also committed.



In the previous graph we can see another inverse correlation between violent crimes per 100,000, but in this example the variable is the percent of college graduates in each community. This relationship is not as sharp and we see a little less of a clear correlation than we did with median income but still seeing a correlation leading the assumption that on average as the percent of college graduates in a community increases, then the number of violent crimes should decrease.

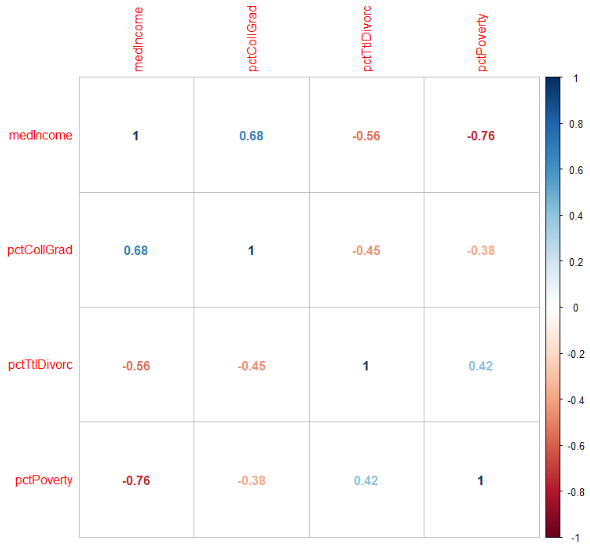


The next variable explored in the analysis was the percent of people divorced in the community. As the divorce rate increases so does the number of violent crimes within a community. In the chart above you can see the relation visually, and one major call out is after about 12.5% total divorce rate there is not a single county without a violent crime.



Another variable that has a positive correlation with violent crime is the poverty rate within a community. This variable was expected after the analysis of the median income variable showed an inverse correlation. It would make sense that the poverty rate would have a positive correlation.

Let’s check to see how correlated these variables are with each other using a function known as corrplot.



As we could have potentially guessed we are seeing a correlation between poverty and median income. Something else to note is correlation between percent of college graduates and median income.

In the last four graphs we have evaluated the correlation between them and violent crimes per 100,000 people. The data is showing that there is a correlation whether positive or negative with all four so can be reasonable to test these four in the next steps to see how sufficient of predictors these are in determining number of violent crimes per 100,000.

Fitting a Linear Model

Using the information from the exploratory analysis we can move on to creating the linear model. We will set violentPerPop as our response variable. The four predictors chosen were Median Income, percent of college graduates, percent of people divorced, and percent living in poverty. The summary of the model shows that all 4 predictors have a p value of almost 0 and do have statistical significance. This is later confirmed by the F Statistic p value of almost 0 confirming that at least one of the predictors is statistically significant. The value for R^2 is not very high showing that only about

41% are accurately explained by the predictors. The median residual of -56.4 shows that our model often underestimates violent crime per 100,000.

Performing Model Selection

Now that the initial model has been created, we can start to run tests to see if the variables not only significant impact has but if the combination of variables provides the best prediction for the variables. This analysis is an important step in the process to make sure that the model is not over saturated, and you are only choosing necessary variables when making predictions.

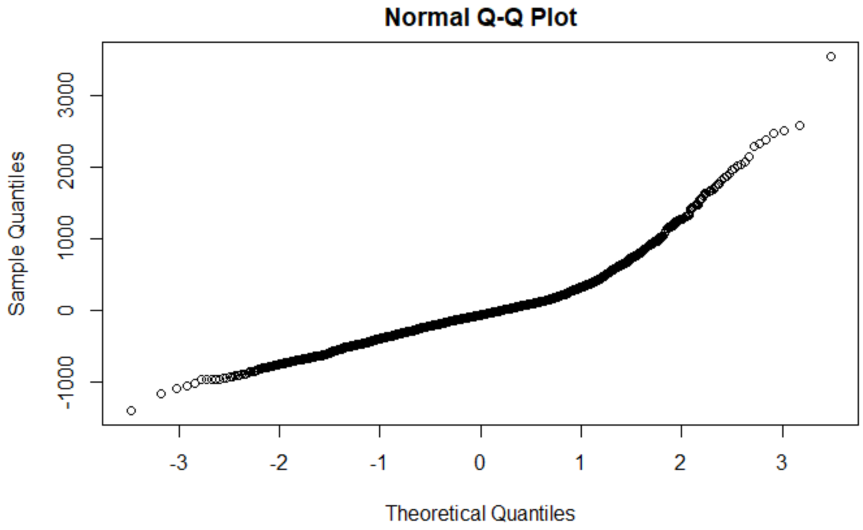
There is two methods to determining this; fast backward variable selection and to perform a stepwise algorithm using the AIC of each coefficient. Under both analysis' the model keeps every initial variable meaning that our model is performing best with all four of our currently chosen predictors. We can conclude that our initial assumption of the correlation of the variables is correct and that all 4 do in fact have a significant impact on determining the number of violent crimes per 100,000

Apply Diagnostics to the Model

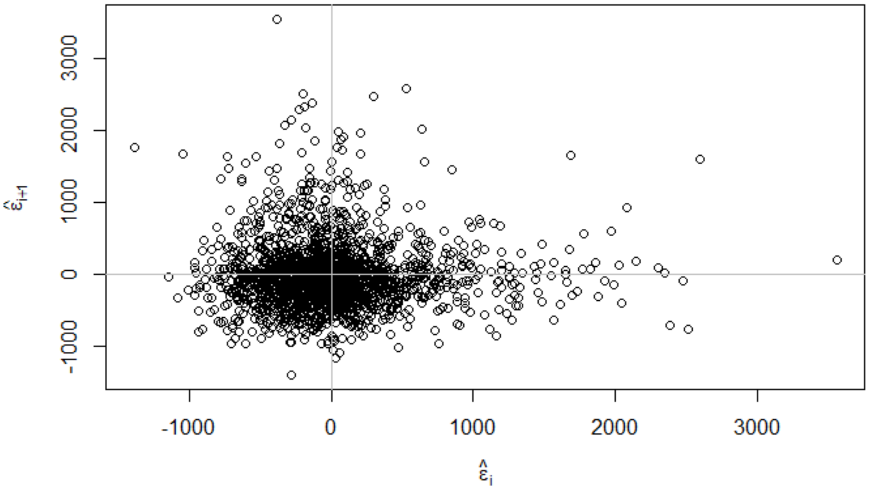
One important aspect of evaluating linear models is to make sure that all error assumptions are met. Because if these error assumptions are not met it could lead to inaccurate confidence intervals for our predictions making our model less accurate. The 4 main assumptions we will be evaluating are:

1. Errors are normally distributed
2. Errors have constant variance
3. Mean of zero
4. The errors are independent of each other

First we'll take a look to see if the errors are normally distributed. There is several different ways to test for normal distribution. I personally prefer the visual evaluations by using plots, the third is a test known as the Shaprio-Wilks test which is more technical but can be easily influenced. First lets take a look at the Q-Q plot. This test will plot the residual distribution of our model vs the residuals that are normally distributed. Ideally we should see a solid straight line.



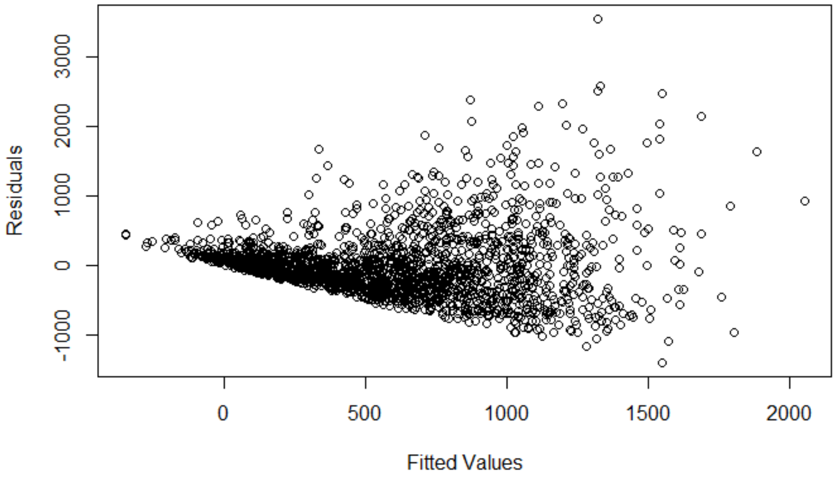
We’re seeing some curvature for this line which initially fails the eye test of our model. When the line is curved towards the theoretical quantiles we may be able to conclude that our errors are right skewed. Secondly lets looks how the errors themselves are distributed. Ideally we will see a mean around 0 with a nice round cloud shape.



The previous assumption of a slight right skew seems to be true. We can see above that most of the dots are centered around 0 but the distribution seems to be farther on the right than it does on the left. In conclusion the errors do seem to be slightly off from being normally distributed but not significantly enough to greatly impact our confidence intervals for the predictions.

Next, we will check to see if the errors have constant variance. This is simple, we will plot the predicted values for the model against the residuals (distance between the predicted and the actual).

To determine the error variance we would expect to see a nice round cloud with no correlation. The appearance to avoid would be a cone shaped distribution.



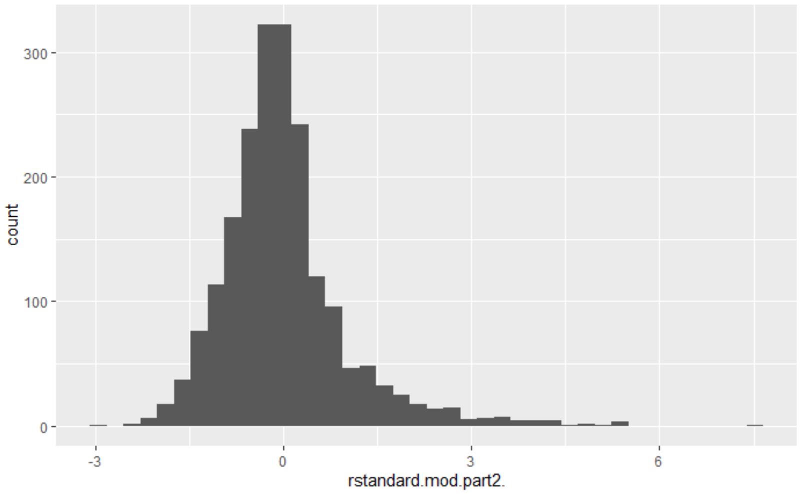
The graph above looks textbook cone shaped so we can confidently conclude that this model does not have constant error variance and their may be some heteroskedasticity in the model.

Finally let us see if the errors are independent. There is not an easy graph for this to be evaluated, but a statistical test known as the Durbin-Watson test which test the null hypothesis that the errors are uncorrelated. If we fail to reject this assumption, then we can conclude that the errors are independent of each other.

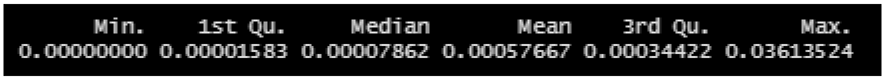
When running the Durbin-Watson test we can set a threshold of alpha = 0.05 meaning that if the Durbin-Watson returns a p-value greater than 0.05 we can fail to reject the null hypothesis and assume that the errors are independent. In this model the p value was 0.5585, this is well above the threshold and can safely assume independence.

Investigate Fit for Individual Observations

To make sure the model isn't being dramatically skewed by observations that are out of the ordinary we can evaluate the residual standards of the each observation in the model. In this way we can standardize the comparison and make it easier to identify errors. The distribution of the residual standards appears as follows.



This looks normally distributed, but we are seeing one observation on the far right that may be skewing or model. To investigate further we will use the cooks.distance test. This will set a threshold that we can use a line where if an observation is above this line, then can be assumed that the observation is in fact having a significantly higher impact on the model than the other observations. When running our threshold, we get a value of 0.8706, so now we know if any observation has a cooks distance higher than 0.8706 that observation should be removed from our training data. After running an evaluation for all observations cooks distance, we can see a summary of the output.

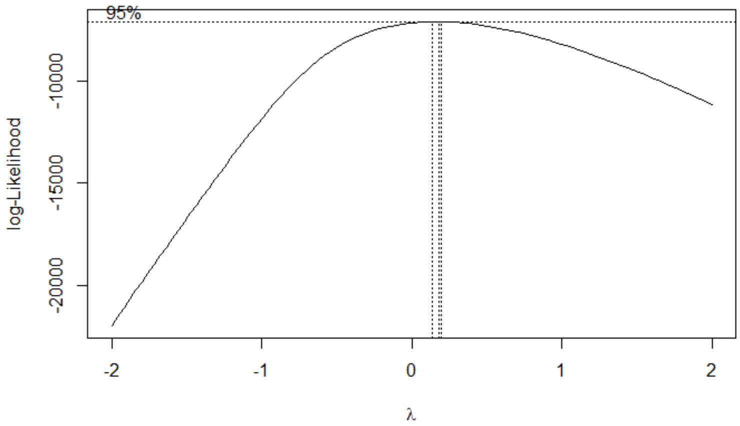


```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
0.00000000 0.00001583 0.00007862 0.00057667 0.00034422 0.03613524 
```

With a max cooks distance being 0.036, it is easy to conclude this is far below the threshold of 0.8706 and we can conclude that all observations can stay in the model.

Apply Transformations to the Model

Previously in our test for model error assumptions we did detect that there may be some heteroskedasticity in our model. This could invalidate our tests for statistical significance which could lead to inaccurate conclusions on the effectiveness of the model. One way to potentially solve for this issue would be to apply transformations to the response variable to account for heteroskedasticity. There is a test known as a Box-Cox plot that could show how to transform the information or if not transformation of the response variable is necessary. Box-Cox will give a lambda value and a confidence interval for that value, we are assuming that 1 should be in the confidence interval. If 1 is not in the confidence interval for lambda then we can conclude that some transformation may be necessary for the model. The box-cox plot for our model comes out to:



The confidence interval for lambda is in near 0 leading the assumption that some transformation is needed. The first option is to do a generic log of our response variable. This would improve some of the heteroskedasticity but is most effective if lambda has 0 within it's confidence interval. The other option is the raise our response variable to the lambda power. The downside to the later is that is more difficult and abstract to make predictions off of these types of transformations. When applying both transformations the r squared alpha is within 0.01 with the lead going to the more abstract version. For the reason previously stated it may be easier and more replicable to use the log transformation of our response variable rather using lambda if the difference in effectiveness is so small.

Making Predictions

Evaluating the coefficients we can see what variables have the largest impact on our predictions. Looking back on our model before transformation we can see the table below for coefficients for each predictor.

Predictor	**Parameter Estimae**	**p-value**
medIncome	0.01540305	<0.000002
pctCollGrad	-6.07933618	<0.000002
pctTtlDivorc	92.02741315	<0.000002
pctPoverty	38.00029775	<0.000002

Based on the coefficients of our first model we would see an linear model algorithm of

*Violent crime per 100000 = -1239.414832 + 0.015403*medIncome - 6.079336*pctCollGrad + 92.027413*pctTtlDivorc + 38.000298*pctPoverty*

Taking one example, lets say averages for all predictor variables, what would the expected number of violent crimes per 100,000 people be. When running these variables using the algorithm above we would get a confidence interval of violent crimes per 100,000 people to be 500-524 violent crimes.

This tells us that on average we are seeing ~500+ crimes per 100,000 people within a community that has average traits for all predictor variables.

Conclusion

The initial goal for this analysis was to try and determine what characteristics of a community can give us an indication of how much violent crime will occur in that neighborhood. We initially started out with some exploratory data analysis to see if there are any clear data correlations with violent crimes. Once a significant level of exploratory analysis was done we could narrow our focus down to four predictor variables; median income, percent of college graduates, rate of divorce, and finally the percent of people living below the poverty line.

After our focus was narrowed onto a couple of predictor variables we began to create linear models and evaluate performance. We found that all of our predictor variables have statistical significance and therefore cannot be removed. There was a conclusion that some transformation of the response variable may be needed and we walked through and example of how to use the linear model algorithm to predict future neighborhoods violent crime per 100,000.

For our specific model, I would say that it is a good predictor of violent crime but may be too simple to provide extremely valuable results. A good follow up to this analysis would be to perform a Stepwise analysis using AIC on the entire model and see what those predictors were chosen and see how far off our initial exploratory data analysis may have been off.