What Causes Violent Crime?

Robert Williams

Alaina Spicer

Tadas Vilkeliskis

Ian Cordasco

December 21, 2010

Abstract

The amount of violent crime in a community affects all aspects of life in that community – the economy, the mood of residents, the careers of the politicians in power, etc. Unfortunately, violent crime is a hard variable to measure because many such crimes go unreported. The goal of this study is to find a statistically significant model that can predict the violent crime in a community using explanatory variables that are readily available in public data sets. In doing this, the resulting model will allow law makers and police to understand if their communities exhibit unexpectedly high or low crime – this could be used to prompt an investigation and get a better understanding of crime in that community. Through linear and logistic regression, two useful models were found that can be used for these purposes. Additional experimental models such as neural networks were investigated with lesser success.

1 Introduction

The purpose of this study is to analyze the causes of violent crime. There are over 11 million crimes committed each day in the United States, and that is only the amount which are reported. There are studies [1] [2] which suggest that a large amount of crimes, especially violent crimes, are never reported. This means that law makers and police forces may not know how bad crime truly is in their communities. Therefore, the main reason to analyze the causes of crime is to accurately predict the amount of violent crimes which are occurring.

Another reason to analyze crime is to show law makers where laws may have to be enforced or where some laws are not needed. People believe there are many different causes of crime but with this analysis the number of variables can be reduced. The fewer the amount of explanatory variables, the better chance those looking at large data sets will have at finding useful trends. In short, a simple model allows humans to process lots of information better. By pointing out which variables are significant, a law maker or the police can be pointed in the right direction on where to set their investigations.

Everyone would like a safe place to live so the more humanitarian goal of this study would be to help police departments, legislators and community groups know how to reduce crime. Judging from the United States's short but colorful past, reducing crime is not easy. One city may benefit from one solution while another requires a radically different solution. For this reason, the goal is to devise some model from statistical analysis that could serve as a base template for answering these difficult questings for all cities.

Certain variables like race, location, and employment are often attributed as explanations of crime, but this study unveils additional areas which affect crime.

The data set chosen for this study was from the University of California Irvine [3] which was a compilation of data collected from the Department of Commerce, Bureau of Census, Department of Justice and the Federal Bureau of Investigation. For the purposes of this study, the year 2007 was chosen, even though almost twenty years of data were given. The decision to focus on a single year was because time-series analysis is beyond the scope of this study.

[1] http://www.highbeam.com/doc/1G1-64114364.html

[2] http://www.cbsnews.com/stories/2002/09/09/national/main521212.shtml

[3] http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

The data has 124 quantitative variables, in addition to some observational variables which were pruned from the data set for the purpose of statistical analysis. Each of the nearly 2000 observation rows is a different community in the United States, ranging from small towns to large cities. The response variable is Violent Crime Per Population, which is the amount of reported violent crimes in a community divided by the population.

The strength of this data set is the amount of communities sampled. Undoubtedly, there are communities in the set which have a large number of unreported crimes and therefore the response variable may not be accurate. Just as likely, there are communities where every single crime is reported. Since the analysis will be done on all rows in the data set, the problem of unreported crimes should not affect the outcome too drastically.

Analyzing the violent crime in all of these cities will hopefully lead to answers to the following questions:

- Can crime be predicted?
- Which variables are significant to an increase in crime?
- Do different cities have different significant variables?

2 Preliminary Model

The original dataset used had 124 quantitative variables that could be used as explanatory variables. This amount of variables was too large for any in-depth study and so it had to be narrowed down. R's step-wise function was used for this purpose which arrives at a linear regression model with the minimum (optimal) AIC value.

The step-wise function returned a model with 13 variables, all of which showed a significant relationship with violent crime rate, and are listed below:

Variable Name	Explanation
PctKids2Par	The percentage of children in family housing with two parents
HousVacant	The number of vacant households
pctUrban	The percentage of people living in areas classified as urban
PctWorkMom	The percentage of mothers in the labor force
NumStreet	The number of homeless people counted in the street
MalePctDivorce	The percentage of males who are divorced
PctIlleg	The percentage of children who born out of wedlock
numbUrban	The number of people living in areas classified as urban
PctPersDenseHous	The percentage of persons living in dense housing
raceptctblack	The percentage of the population that is African American
MedOwnCostPctIncNoMotg	The median owner's cost as a percentage of household income
RentLowQ	Lower quartile rent
MedRent	Median gross rent

Table 1: Thirteen variables in all three models

Out of these 13 variables we expected percentage of kids born to never married parents and people living in areas classified as urban to be the most significant. If a child s parents were never married, may mean that the parents aren t together, possibly signifying an unstable household. Crime may also be significant to urban areas because they are usually densely populated (more people in a condensed implies more crime). On the other hand, the variable of percentage of mothers of children under 18 in the labor force should have little effect on crime. If children are in the labor force they might not have a reason to commit crime.

The initial regression model we constructed from those 13 explanatory variables is as follows:

$$\begin{aligned} ViolentCrimesPerPop &= -0.291 \times PctKids2Par + 0.304 \times HousVacant + 0.046 \times pctUrban \\ &- 0.084 \times PctWorkMom + 0.200 \times NumStreet + 0.168 \times MalePctDivorce \end{aligned}$$

$$
\begin{aligned}
&+ \quad 0.176 \times PctIlleg - 0.239 \times numbUrban + 0.203 \times PctPersDenseHous \\
&+ \quad 0.179 \times racepctblack - 0.062 \times MedOwnCostPctIncNoMtg - 0.233 \times RentLowQ \\
&+ \quad 0.277 \times MedRent + 0.226 \qquad (1)
\end{aligned}
$$

The corresponding plots for this model's redisuals is shown in Figure 1. Judging from the plots, this model is far from perfect. The deviations of the normality curve in the Q-Q plot show that the model is lacking in normality, which is an important criterion for correct models. The plot of residuals vs fitted values is an appropriate curve for the preliminary model. There is only a small recess in the fitting curve; however, the data points seem to have a funnel effect which suggests that a model transformation is needed. Similarly, the scale-location and residuals vs leverage plots show some deviations. Nevertheless, the model provides a relatively good prediction with adjusted R^2 of 0.6643.

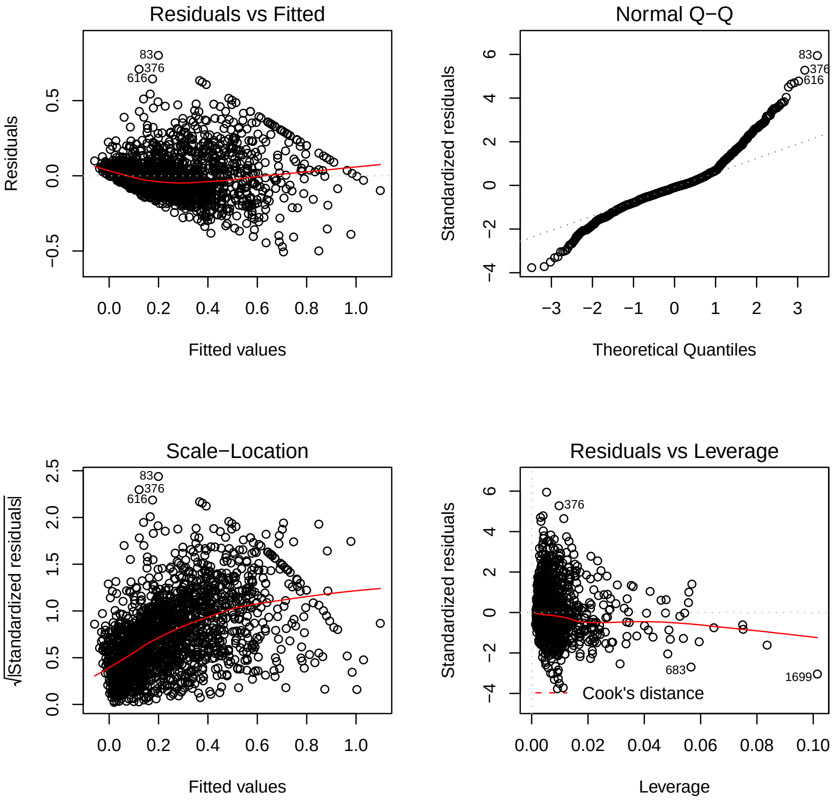

Figure 1: Initial linear regression model.

2.1 Possible Improvements

To improve the initial model two methods were used: the power transformation on the response variable, and logarithm and polynomial transformations for the explanatory variables. In order to obtain the transformation power, a Box-Cox plot was created for the initial model, shown in Figure 2 to suggest a transformation power of 0.27. Creating the Box-Cox plot was not feasible right away because our data set contained some zero values for the response variable. A common practice applied in statistics for dealing with zero values was employed, which is replacing zeros in the data by half of the smallest observation for that variable. Such data modification allowed the Box-Cox plot to work.

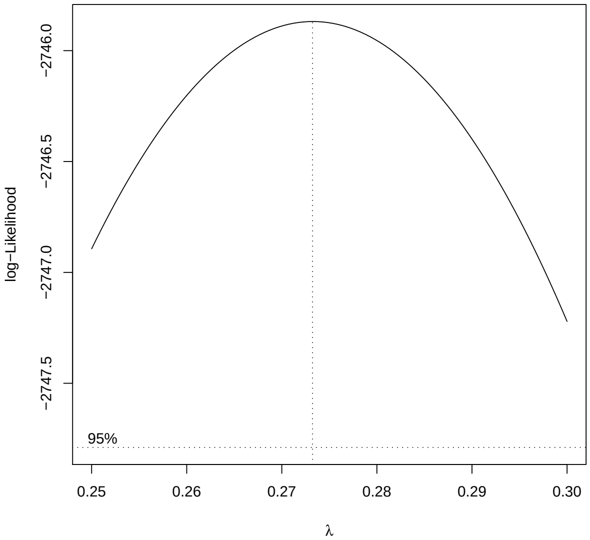

Figure 2: Box-Cox plot for the initial model.

For the explanatory variables we referred to GAM plots. GAM plots give an indication of the type of relationship an explanatory variable has on the response. This provides good insight on how to transform the explanatory variables. For instance, as shown in Figure 3, the curves of `racepctblack` and `PctIlleg` look quadratic in nature, so they should be transformed with a second degree polynomial. The `NumStreet` variable needs a fourth degree polynomial transformation and `pctUrban` needs a fifth degree polynomial. The rest of the variables do not show easily recognizable polynomial shapes, making it hard to identify which transformation should be applied.

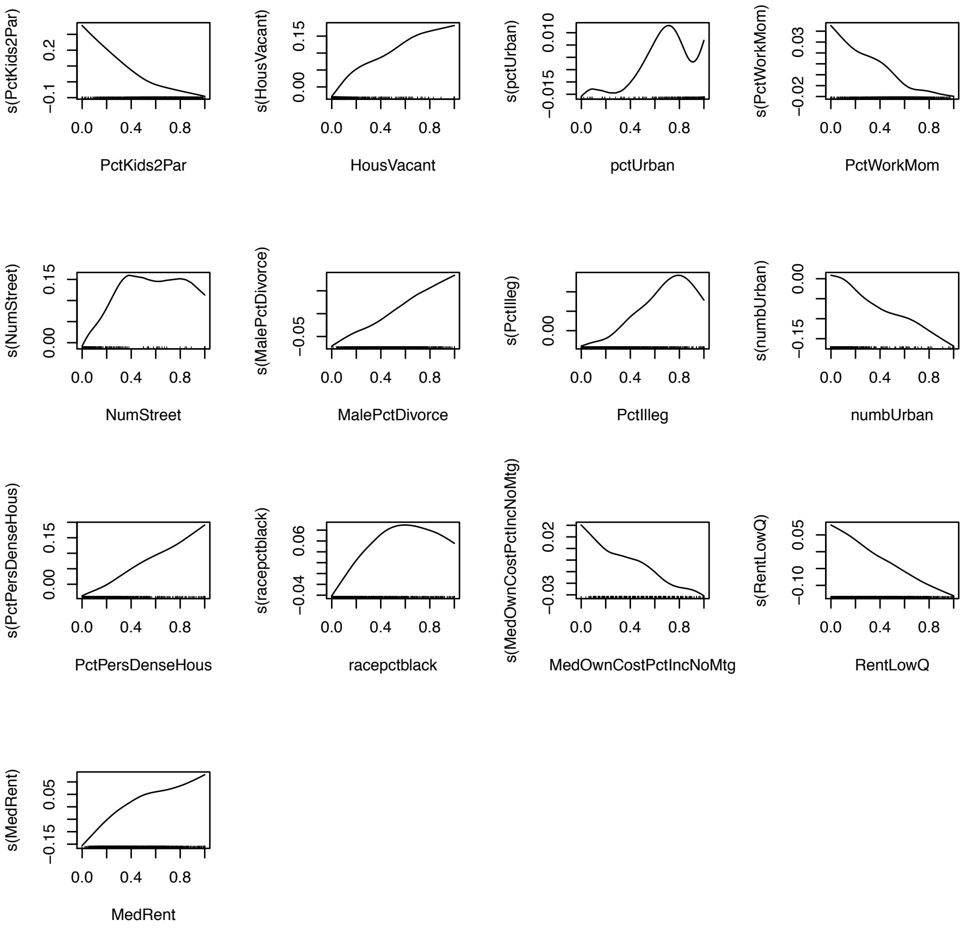

Figure 3: GAM plots for the initial model.

To alleviate this problem the variable transformation was split into two parts. The first part included only the polynomial transformations mentioned above and a logarithm transformation for the `HousVacant` and `MedRent` variables. In the second part, all variables transformed in the first part were transformed in the same way but then all the rest of the variables were transformed logarithmically. These transformations both improved the model, resulting in the following adjusted R^2s:

- 0.6873 for the first transformation.
- 0.6742 for the second transformation.

Even though the second transformation decreased the value of adjusted R^2, it was chosen as the superior model because it increased the significance for some variables. The final model is shown[4]:

$$\begin{aligned}
\frac{ViolentCrimesPerPop^{0.27}-1}{0.27} &= \beta_1 \times log(PctKids2Par) + \beta_2 \times log(HousVacant) \\
&+ \beta_3 \times poly(pctUrban, 5) + \beta_4 \times log(PctWorkMom) \\
&+ \beta_5 \times poly(NumStreet, 3) + \beta_6 \times log(MalePctDivorce) \\
&+ \beta_7 \times poly(PctIlleg, 2) + \beta_8 \times poly(numbUrban, 2) \\
&+ \beta_9 \times log(PctPersDenseHous) + \beta_{10} poly(racepctblack, 2) \\
&+ \beta_{11} \times log(MedOwnCostPctIncNoMtg) + \beta_{12} \times log(RentLowQ) + \\
&+ \beta_{12} \times log(MedRent))
\end{aligned}$$

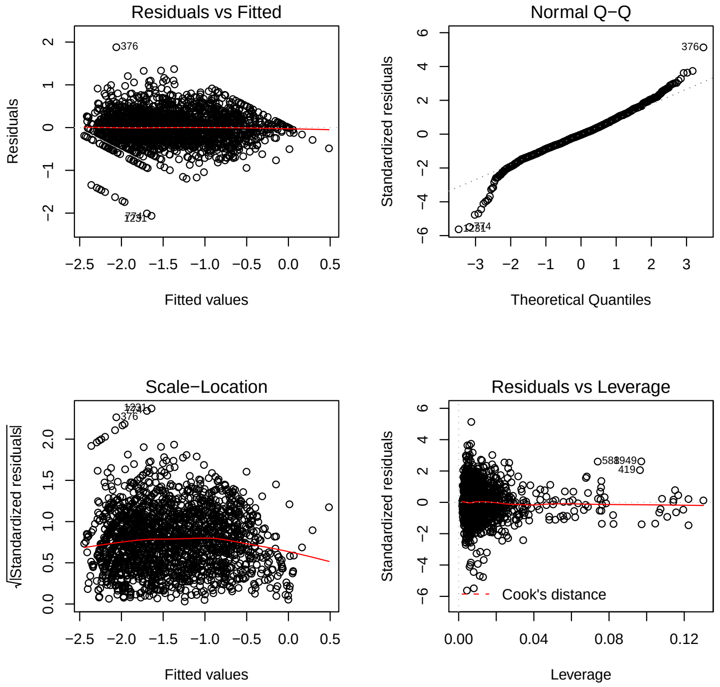

Figure 4: Linear model after polynomial and logarithmic transformations.

Figure 4 shows the residual plots of the final transformed model. It is clear that the model fit improved and is nearly perfect; additionally, the funnel effect is gone. There seems to be an improvement in the scale-location plot because the values are better distributed and the prediction

[4] $poly(X, Y)$ represents a Y^{th} degree polynomial for the variable X.

curve became much closer to a straight line. The normality of the model also improved; however there is still a slight deviation from the normality due to outliers.

2.2 Problems and Outliers

The residuals of the transformed model look much better than the original model but there is still a deviation from normality (see the Q-Q plot). It makes sense that there would be outliers in this study because there are almost two thousand observations, all of which are communities in America – these communities range from small towns to large cities. To attempt to address this, a thorough analysis of outliers was conducted. The goal was to make sure the model was not being adversely affected by outliers, either in terms of the residuals or the model's R^2 fitness.

The first outlier test performed comes from the `car` package in R, called simply `outlierTest`. This test uses the Bonferonni adjustment to report those data observations which fall outside of the normal range for the data. When running this test on the crime dataset, it reported the following six cities as outliers:

1. Vernon, TX
2. La Canada Flintridge, CA
3. Glens Falls, NY
4. Mansfield, TX
5. West Hollywood, CA
6. Plant City, FL

To attempt to interpret these results, the basic demographics of each of these communities was investigated. It turns out that they all have a relatively small population (between 10,000 and 50,000) while also having very high violent crimes per population (a standardized value greater than 0.83 in all cases – only about 4% of the observations are above this amount).

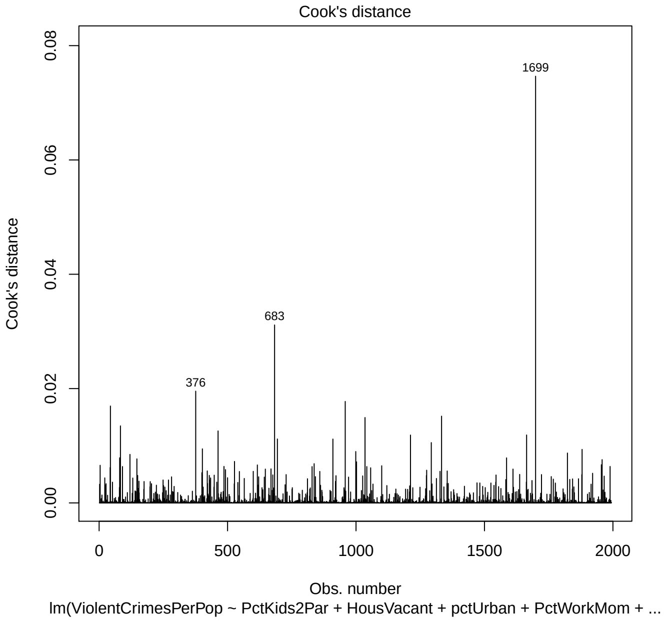

Figure 5: Cook's Distance Plot

While not directly related to outliers, influential points were also considered using the Cook's Distance Plot, which can be seen in Figure 5. Point 376 is La Canada Flintridge, which was found in the outlier test above. 683 is Philadelphia, PA and 1699 is Fort Lauderdale, FL. It is unclear why a diverse set of communities such as these three would all be highly influential.

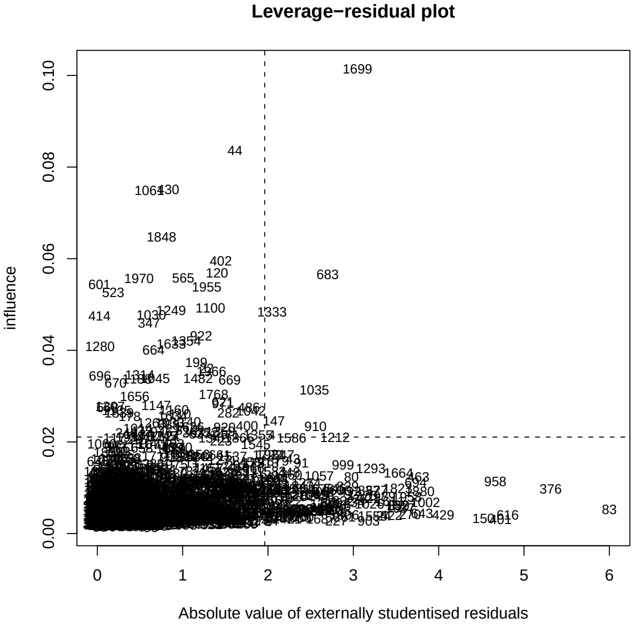

Figure 6: Leverage-Residual (lr) Plot

To combine the concept of outliers and influence – influential outliers – a function called `lrPlot` from Alan Lee of the University of Auckland, New Zealand [5] was used. This plot, in Figure 6, shows four quadrants, each with some combination of leverage (influence) and residual strength (outliers). The upper right quadrant contains the influential outliers which are of concern for this analysis – those points which have a large effect on the model but also don't fit in with the other data observations. The points in the quadrant for influential outliers which were not already identified by prior methods described above are Ocean City, NJ and Gatesville, TX. These two are both very low violent crime per population (standardized value less than 0.10).

Removing all of the outliers (total of ten) found with the methods described above, the new model gets an adjusted $R^2 = 0.6899$, compared with the the $R^2 = 0.6742$ for the original transformed model. This is a slight improvement to the R^2; however, the original concern was with the normality of residuals. Unfortunately this outlier removal did not help that plot at all. Removing only the three influential outliers results in an $R^2 = 0.6733$ which is even worse than the first try, while also not helping the residual plots. Finally, it should be noted that this outlier removal was performed recursively. That is, after the outliers were removed, the outliers in the resulting model were also

[5] http://www.stat.auckland.ac.nz/ lee/330/R330.txt

removed. This was repeated for three iterations, until it was clear that no depth of outlier removal would produce different results.

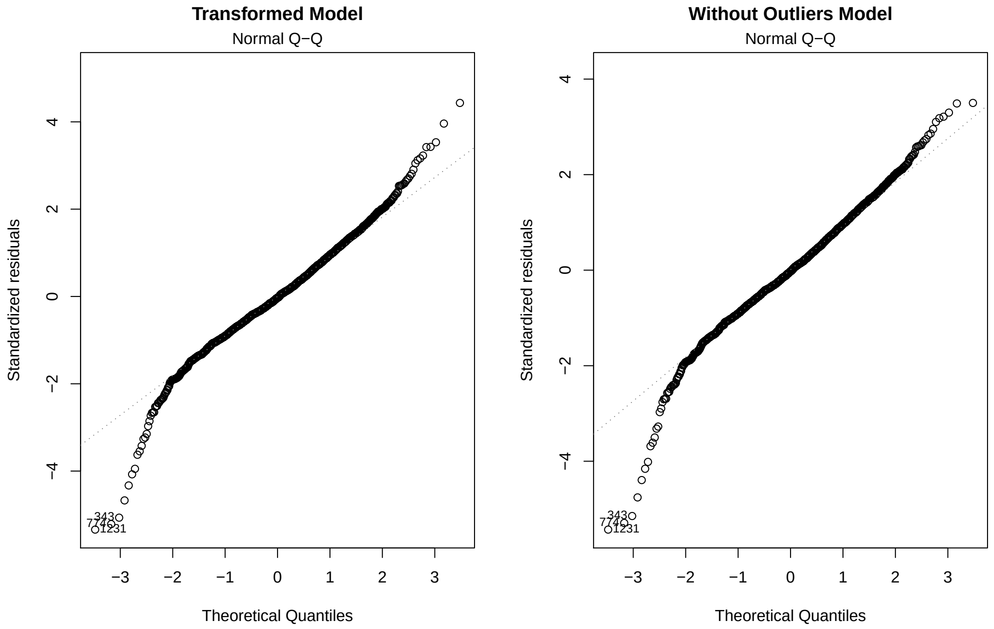

Figure 7: Q-Q Plots of Transformed Model With and Without Outliers

Outlier removal is a controversial topic in the mathematical community, and it requires extensive research in order to justify removal. The problem with this data set is that it was collected by third-party researchers, and the data set gives no indication of whether there may have been measurement error or other understandable causes of outliers. Additionally, the dataset was pre-standardized, meaning that each column was converted to a number which preserves the integrity of the data but makes it impossible for human analysis. For example, it may have been desired to analyze the populations of the cities which were outliers to get a better idea of why they were affecting the overall model. This is not possible when the data is in standardized form because the actual number for population is not recorded in the data set, only the standardized value. Even if these concerns are ignored and the outliers are indescriminately removed, it still does not help the normality of the residuals, which was the primary goal for this analysis. As seen in Figure 7, the Q-Q plots of the original transformed model with outliers and without outliers are nearly identical. Although there was a slight improvement to R^2, removing outliers is not warranted in this case due to all the reasons described. It was decided to keep the outliers in the data set.

2.3 Preliminary Conclusions

The overarching theme among the explanatory variables in the model seems to be related to poverty either directly or indirectly.

Of the thirteen variables, the most significant in the model was undoubtedly the percentage of persons living in dense housing (`PctPersDenseHous`). Dense housing is defined by living quarters where there is more than one person per room. This can indicate a few things but most often suggests a poorer family or tenant. The other natural association with dense housing is prostitution and drug dens. The cheaper the housing, the more profitable it becomes to run one of the two above mentioned “businesses”. Violent crime is often linked to these two kinds of illegal activity.

This is also directly correlated to another of the variables, lower quartile rent in rental housing (`RentLowQ`), which is self-explanatory. Its importance follows from the assumption that denser housing is often cheaper, therefore the more dense housing that exists in a community the lower the lower quartile rental housing.

Dense housing is also related to the number of people living in urban areas (`numbUrban`). The smaller the area classified as urban as well as the more people living in that area, the greater the need for dense housing will be. A less pessimistic observation is that the less space a person has to themself, the more quickly they usually become agitated. It makes sense that the more agitated someone is, the more likely they are to make rash decisions which could lead to violence.

After `PctPersDenseHous`, the next most important variable was the percetage of the population that is African American (`raceptctblack`). There are several sociological reasons why this may be the case. White flight was first observed in the 1950s after the decision by the Supreme Court in the *Brown v. Board of Education* to desegregate schools. This occurrence, which is seen less often today, is a major factor in what is termed environmental racism. The creation of the suburbs led to a concentration of African Americans in urban areas. With the understanding that two of the other significant variables deal with the urban population, this variable acts as a supplement to them. Given that explanation, this variable may seem to be redundant (having both a black population and urban population variable) given that African Americans tend to be associated with urban areas; however, removing either variable from the model hurt the significance of all variables and lowered the R^2, suggesting that both need to be present.

Next, the number of vacant households (`HousVacant`) are very significant to this model. Vacant housing can be the result of a few serious issues. The household could be a building shutdown by the health department for sanitation reasons. Often, when a household is closed by the state the household consists of poorer residents and as has already been stipulated, the poorer the neighborhood the higher the number of violent crimes. Additionally, abandoned houses usually become the dwelling places for criminals with no better place to stay.

The percentage of children living with two parents (`PctKids2Par`) is also logically important to the model. Children with two parents have a more stable evironment to mature in, and stability is an important part of childhood. If a child grows up in an unstable home, a variety of negative outcomes can be linked to these events. One result is the lack of academic success, which could then affect the child's prosperity and emotional and mental health as an adult. Also, having only a single parent means that the child can roam freely around the neighborhood and may get involved in criminal activity.

Median gross rent (`MedRent`) seems to be tied as the next best predictor with the percentage of people living in urban areas and the percentage of males who are divorced (`MalePctDivorce`). The median gross rent is the amount of rent a landlord received before taxes and utility costs are deducted. This and the percentage of people living in urban areas are likely correlated to each other. Depending upon the community the percentage of people living in urban areas could have either a positive or negative effect on rent, although economics would suggest it to be a positive relationship.

The percentage of males who are divorced is clearly correlated with the percentage of children living with two parents. This relates back to the issue of stability. What might have been interesting to study with this variable is the percentage of these males who have custody of their children versus those who do not.

Next, the number of homeless people in a community (`NumStreet`) is an indicator of quite a few important elements. One such element is the overall wealth of a community. Wealthier communities often have the police usher the homeless out of the area to ensure "beauty" and "dignity" of the area is maintained while poorer communities are often home to several homeless shelters. In some of these areas, gang violence targets homeless people since they have less credibility than others. Due to the violence the homeless endure and their living conditions, physical and mental illness are common among them. Between the violence against the homeless and violence which may stem from their mental illnesses, it is clear how the amount of homeless people can affect violent crime. The significance of this variable corellating with violent crime should encurage law makers to do whatever necessary to help homeless people rise above their situation.

The percentage of working mothers (`PctWorkMom`) is important for obvious reasons. If both parents are working (the father is assumed to also work, or may be absent altogether), they cannot likely afford childcare for the hours they are at work leaving children vulnerable to be victims of violence or to be recruited by gang members. If the working mother is in fact a single mother as well, the above situation is more likely to occur since she may be working more than one job and will not have time to monitor her children. These children may also contribute to the percentage of illegitimate children (`PctIlleg`) unless the parents were married.

Overall, these variables provide an understanding of what may be causing crime in the communities from our data set. As is common with statistics, correlation does not imply causality, so it is possible that there are some lurking variables. However, since so many of the significant variables are related to one another, and since they all were chosen from a large pool of possible explanatory variables, the odds are quite good that these variables accurately reflect the causes of violent crime. At the very least, they will be a valuable starting point for law makers and police, since not all of them are the usual suspects which would be included in most state surveys of crime causes.

3 Experimental Models

In an attempt to find models which are more optimal than the preliminary model, which had the weakness in its residual normality and an R^2 that could be higher, some experimental models were investigated. The first experiment was to split the data into high and low crime communities to see if the explanatory variables were different. The second experimental model was to use logistic regression to predict a response of whether a community had high crime or not (a binary response). Finally, a neural network model was constructed to try to predict the binary response better than logistic regression.

3.1 Splitting the Data

Since the investigation of outliers showed that communities with high crime behave differently from communities with low crime, it was suggested to split the data into two sets. One set would repesent communities with high crime, and the other set would represent communities with low crime. The idea was that these two groups might be modelled best with different sets of variables appropriate to high or low crime. For example, a large percentage of densely populated houses in a community may lead to high crime, but the absense of such densely populated houses may not cause low crime.

The large question when splitting data is which threshold to use for the split. Since the data set was already standardized and the response variable (`ViolentCrimePerPop`) was measured as a percentage ranging from 0% to 100%, the obvious place to split would be 50%, and this was indeed the first threshold chosen.

Fifty percent was used as the upper limit for "low" crime, and so any communities with `ViolentCrimePerPop` greater than or equal to fifty percent was put in the set of "high" crime communities. In order to find the best models for the two split data sets, **R**'s `step` function was used. In the case of the low crime data set, the better model was found `step`ping backwards. The AIC was -8679 with an R^2 of 0.593, which is significantly worse than the preliminary non-split model. This model for the low crime data set barely predicts 60% of the data points given to it and has *significantly* more variables than the preliminary model. The residuals do look good but not nearly as good as those in the transformed preliminary model. For the high crime data set, the better model resulted from backward stepping. The AIC was -1122, *far* worse than the normal model, and the R^2 was 0.371, worse even than the low crime model. Overall, the models found by splitting the data at the 50% mark were unsatisfactory.

As a second try, the data was instead split in a way where the two resulting data sets would have similar sizes (in the previous try, the "low" data set was considerably larger than the "high" data set). It turns out that 13.3% caused 900 items to be in the "low" data set and approximately 1100 to be in the "high" data set. Both models in this experiment also resulted from `step`ping backwards. The model to predict a low crime that the step function produced relied on 40 variables, not all of which were significant to the model, with an AIC of -6433 and an R^2 of 0.38. To try to improve the model, the statistically insignificant variables were removed iteratively, but each removal caused more variables to become insignificant and none of these attempts improved the model. Clearly this was an unacceptable model which would not be useful in the least. **R**'s step function did, however, return a slightly better model for predicting "high" crime. This model contained 44 variables, again not all of which were significant, with an AIC of -3989 and an R^2 of 0.534. These residuals looked worse than the ones obtained from the low model. Again, manual modification of the model was attempted but this only served to hurt the model – even more than with the low crime model.

Given the poor characteristics of the split data models, despite two very different strategies for choosing the threshold, it was decided to not continue experimenting. The models that resulted from this study were not useful as they were not as efficient or accurate as the original non-split model. One useful takeaway from this analysis was that the areas with more violent crime have much worse residuals than areas with low crime. This does not help improve the preliminary model, though, because residuals are tied to the models and the models in this analysis were completely different from the original model.

3.2 Logistic Regression

Although the data split experiment did not produce a more useful model, it did suggest that the low and high crime communities behave differently. The next experiment to build off this result was to test if a logistic model might produce better results. Logistic models have a response variable which is binary – that is, either 0 or 1. In this case, a new variable was defined as `HighCrime` which was 1 for communities with a `ViolentCrimePerPop` value of ≥ 0.5 and 0 for communities with a `ViolentCrimePerPop` value of < 0.5. After the data was transformed in this manner, it was possible to use the `glm` function of **R** to create logistic regression models.

As a starting point, the transformed model found in section 2 was used, except instead of using the quantitative response value of violent crimes, the new binary `HighCrime` variable was used. The resulting logistic model had an AIC value of 529.56 and pseudo-R^2 of 0.563. Unlike the linear model, not all of the variables used in the model were significant. Unfortunately, it is not possible to compare this model to the linear model since the R^2 calculation is only an approximation and not calculated the same way. However, the logistic model can be tested for fitness independently.

In order to test the logistic model, the fitted points returned from the regression were compared with the actual values for `HighCrime`. The amount of points which were correct were 89.8% which is very good. The problem with comparing this to the linear model is that they are modeling much different responses. The linear model is predicting an exact number for violent crimes per population, while the logistic model is predicting a much broader response which is whether the crime will be high (≥ 0.5) or not. Because of this, it makes sense that the logistic model will be more accurate. This does not, however, mean that it is a stronger model.

Since the logistic model is weaker, it might make sense that different explanatory variables would be warranted. In order to investigate this, an R function was created that performs a stepwise search for the best logistic regression based on AIC. It used a foreward stepwise search which, during each round, added the variable which improved the AIC the most. It concludes when it converges to some user-configurable difference between rounds. Using this function, the logistic model that was discvered had an AIC of 525.17. The psuedo-R^2 was 0.534. The AIC was only slightly better than the original logistic model while the R^2 was slightly worse. Many of the variables are also the same. The following variables are in the logistic model found via stepwise search:

$$\begin{aligned} HighCrime &= \frac{1}{1+e^{-z})}, \text{where } z = 0.505 \times PctIlleg + 0.468 \times NumStreet \\ &- 0.193 \times agePct12t29 - 0.208 \times racePctWhite - 0.113 \times PctHousOccup \\ &+ 0.060 \times pctUrban - 0.219 \times PctKids2Par + 0.436 \end{aligned}$$

The interesting part about the variables used in the step-wise model is that some of them are the opposites of the ones in the linear model. For example, the amount of occupied houses is used instead of vacant houses, while the percentage of white people is used instead of percentage of black people. Testing this model with the above test method, the 89.3% which is also slightly worse than the original logistic model. The deviance plots are also very similar.

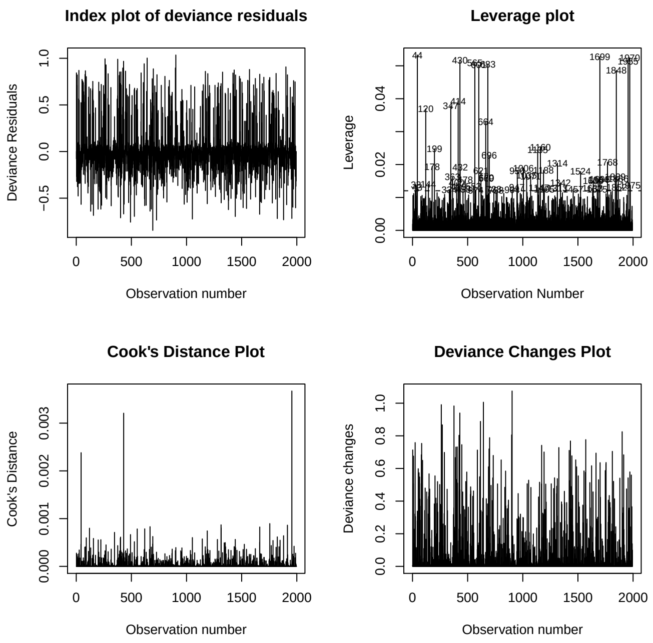

Figure 8: Diagnostic Plots for Stepwise Logistic Model

The diagnostic plots for the logistic regression model are shown in Figure 8. The deviance residuals are quite large, but since most of them are below 0.5, when rounding the data to 0 or 1 it still comes out to the correct value.

A better test method was then used, which involved splitting the data into two parts. The first part had only 100 observations, which the second part had the remaining observations. The latter part of the data was used to train a logistic regression using the `glm` function, while the former 100 observations would be used to test the data for fitness. This splitting technique is used because sometimes regressions tend to fit the data so perfectly that it is dependent on that specific data – any new data will cause it to perform poorly. This technique was performed on the data, and a logistic regression was found with the stepwise function to fit the training data observations. The resulting model fit 86% of the points with its predictions, which is again a very good model.

Overall, the logistic model shows that the explanatory variables included in this data set are excellent for describing the macro trends regarding violent crime. They can very accurately predict whether a community will have high crime or low crime. Depending on the needs of the policymakers or law enforcement members interested in these models, the logistic regression may prove valuable. If they want to determine which communities are at risk of crime, the accuracy of these models will be useful. However, if they want to hone in on the particular effects of any single explanatory variables, the binary response of logistic models does not give a granular enough indicator; a linear model will be a better choice for such a purpose. In short, the utility of a logistic model depends on the goals of those analyzing this data.

3.3 Neural Networks

Due to the success of the logistic model in attributing a community to either low crime or high crime, the next step for an experimental model was a neural network. Neural networks take a much different approach to predicting data which emulates the functions of the human brain. In short, a neural network is a group of connected "neurons". The connections between these are destroyed and created chaotically during the training phase of the neural network in order to arrive at an optimal model. In statistics, the traditional approach when creating a neural network model is to split the data into two pieces. The first piece – encompassing a vast majority of the data – is fed into the neural network to "train" it. The second piece of the data is used after the neural network is formed in order to test how well it predicts the response variable.

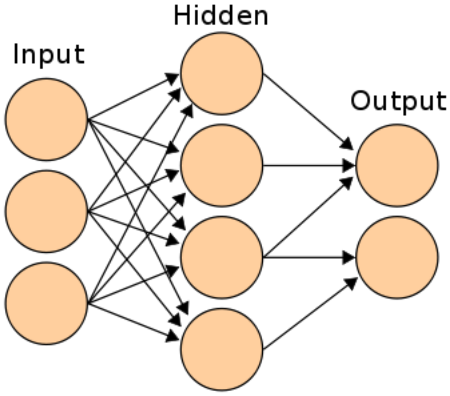

Figure 9: A Sample Neural Network, showing the nodes and connections

Neural network models are known to work well on large data sets. The nearly two thousand data items in the violent crime data set is a sizeable amount, so it should be enough to train an accurate neural network. R does not provide any step-wise searching function for neural networks, likely because running the neural network training just once takes a relatively long amount of time compared with a simple linear or logistic regression. Because of this, it did not make sense to create a custom R function for this purpose – it would simply take too long to run. Since there was no sense of step-wise, the obvious solution is to simply use all the variables as explanatory variables. The hope is that the training process of the neural network would provide low weights to the insignificant variables, preventing them from harming the fitness of the model.

As described above, the general procedure is to split the data into a training set and testing set. This was done by separating out 100 random points for testing and training with the remaining 1894 rows. After running R's neural network training on this data set, a model was produced. The model was used to predict whether the 100 test communities would exhibit high violent crime or not. It turns out that the model was 86% correct in its predictions. This is very close to the 90% achieved by the logistic regression.

A considerable attempt was made to produce a visualization of the resulting neural network. There has been a large amount of research by other students to this end, as well as the non-profit Data Mining Group [6] who pioneers a standard for representing neural network models with its PMML file format. There actually is a library for R which outputs neural networks to PMML format, but there are no publicly available visualizing tools which take PMML as input. Unfortunately, it was not possible to produce a visualization of the neural network.

All things considered, the logistic model should probably be used over the neural network model. It is slightly more accurate, and it is a much simpler model. By the concept of Ockam's razor, there is no point in involving the complex world of neural networks which mixes the latest biological and statistic concepts. Logistic regression does a satisfactory job at using the explanatory variables found to predict whether a community will have high crime or not.

4 Conclusion

The preliminary model found through linear regression and subsequently improved with polynomial and logarithmic transformations fit the data very well, as evidenced by the high adjusted-R^2 value. In addition, the residual plots showed desirable images. However, the Q-Q plot was sub-optimal, forming a weak "S"-shape. A thorough analysis of the outliers showed that removing the outliers did not improve the Q-Q plot.

[6] http://www.dmg.org/v4-0-1/NeuralNetwork.html

In an attempt to improve the model, three different experimental approaches were taken. The first approach still involved linear regression, but the data was split into high crime and low crime communities. The idea was that the communities with high crime would have different explanatory variables than the communities for low crime. Unfortunately, after using R to perform step-wise regression on the separated data sets (separated at two different thresholds), the resulting models were much weaker than the original linear model. This suggests that the high crime and low crime communities do not behave differently, and the completedness of the full data set allows R to find the best overall model. As a result, it is suggested that this data set not be split in finding the optimal model.

The other two experimental models took a different approach. Instead of trying to predict an exact value for the amount of violent crime per population in a community, they try to predict the simple binary result of whether it has high crime or not. Depending on the intended usage of the model, this approach may actually be more desirable. The first model implementing the binary response variable was constructed using logistic regression. R produced a model that successfully predicted a whopping 90% of the test communities' violent crime status. The second model was created using R's neural network training function. The resulting model was also very successful, achieving an 86% success rate. Both of these models show that the chosen explanatory variables can be used to predict macro trends very well.

In conclusion, the preliminary model discussed in section 2 is a strong model and can be used to predict an approximation of a community's violent crime per population. The dozen explanatory variables found are all statistically significant to the model, and so law makers, police, and anyone else curious about the causes of crime can look to these factors. With the understanding that this model provides, it would be possible to create a plan to reduce crime in a community by reducing one or more of the explanatory variables. Additionally, the very accurate logistic model found in section 3.2 can be used by those interested in the broader, macro causes of crime. Given the appropriate statistics on a community, the model will give a clear answer as to whether or not the community should be the subject of further criminal investigation. This could enable police to keep a keen eye on problem areas, catching crimes that otherwise would have gone unreported and created a false sense of safety.

The R file in the appendix is provided so that anyone interested in the models can use them.

List of Figures

1 Initial linear regression model. 4
2 Box-Cox plot for the initial model. 5
3 GAM plots for the initial model. 6
4 Linear model after polynomial and logarithmic transformations. 7
5 Cook's Distance Plot 9
6 Leverage-Residual (lr) Plot 10
7 Q-Q Plots of Transformed Model With and Without Outliers 11
8 Diagnostic Plots for Stepwise Logistic Model 18
9 A Sample Neural Network, showing the nodes and connections 19

List of Tables

1 Thirteen variables in all three models 3

R Code Listing

The following is the code listing for the R script file used to produce the models and analyses described in the report.

```
######################################
##                                  ##
##    HELPER FUNCTIONS              ##
##                                  ##
######################################

boxcoxplot = function (formula,data,p = seq(-2, 2, length = 20), ...)
{
# Draws a box-Cox plot for transforming to normality

  l <- length(p)
  boxcox <- seq(l)
  reg.lm<-lm(formula,data,x=TRUE,y=TRUE)
  y<-reg.lm$y
  X<-reg.lm$x
  n <- length(y)
  sumlog <- sum(log(y))
  for (i in seq(l))
  {
    y.p<- (y^p[i] -1)/p[i]
    trans.res<-residuals(lm(y.p~-1+X))
    ResSS.p<-sum(trans.res^2)
    boxcox[i] <- n * log(ResSS.p)/2 - (p[i] - 1) * sumlog
  }
  plot(p, boxcox, type = "l", ylab = "Profile␣likelihood",
      main = "Box-Cox␣plot", ...)
}


log_fix = function(data_)
{
  minval = data_[1]
  for (i in seq(1:length(data_))) {
    if (data_[i] < minval && data_[i] != 0)
      minval = data_[i]
  }
  #print(minval)
  minval = minval / 2

  for (i in seq(1:length(data_))) {
    if (data_[i] == 0)
      data_[i] = minval
  }
  return(data_)
}


#
```

```
# Perform a log transformation on a data and replace 0 with a log of half
# of the smallest non-zero value.
#
log_trans_n_fix = function(data_)
{
  minval = data_[1]
  for (i in seq(1:length(data_))) {
    if (data_[i] < minval && data_[i] != 0)
      minval = data_[i]
  }
  #print(minval)
  minval = minval / 2

  for (i in seq(1:length(data_))) {
    if (data_[i] == 0)
      data_[i] = log(minval)
    else
      data_[i] = log(data_[i])
  }
  return(data_)
}

# perform stepwise logistic regression on data to find best model
# fitness indicator used is AIC
glm_step = function(data_, response_index_, ignore_indices_ = c(), epsilon_ = 10)
{
  library(Design)
  resp = colnames(data_)[response_index_]
  subset = data_[, -ignore_indices_]
  model = paste(resp, "~ 0")
  model.AIC = AIC(glm(as.formula(model), data=subset))
  best.model.AIC = 1e99
  model.vars = ""
  # initialize skip.is to skip the response variable
  # (skip.is are the column indices which won't be tested in this round of stepwise)
  for(i in seq(1:length(colnames(subset))))
    if(colnames(subset)[i]==resp)
      skip.is = c(i)
  j = 0
  while (TRUE)
  {
    for(i in seq(1:length(colnames(subset))))
    {
      if (any(skip.is == i))
        next
      if (nchar(model.vars) > 0) {
        test.vars = paste(model.vars, paste("+", colnames(subset)[i]))
      } else
        test.vars = colnames(subset)[i]
      test.model = paste(resp, paste("~", test.vars))
      test.model.AIC = AIC(glm(as.formula(test.model), data=subset))
      if (test.model.AIC < best.model.AIC)
      {
        best.model.AIC = test.model.AIC
        best.model = test.model
        best.vars = test.vars
        best.i = i
      }
      #print(i)
    }
    if (abs(model.AIC - best.model.AIC) < epsilon_)
    {
      model = best.model
      print(paste("Final: ", model))
      break
    }
    model = best.model
    model.AIC = best.model.AIC
    model.vars = best.vars
    skip.is = c(skip.is, best.i)
    j = j + 1
    print(paste(j, paste(":", model)))
  }
}

# returns percentage of how many predicted (fitted) points are equal to real points
compare_to_real = function(fitted_, real_)
{
  correct = 0
  total = length(real_)
  for(i in 1:total)
  {
    if (round(fitted_)[i] == real_[i])
      correct = correct + 1
  }
  correct / total
}

######################################
##                                  ##
##    INITIALIZATION OF DATA        ##
##                                  ##
######################################

# read in file
original_com = read.csv("PROJECT/communities.csv", header=T)
com = read.csv("PROJECT/communities.csv", header=T)

# remove any non-numeric values from the data set -- leaves exactly 100 variables
rmcols <- rev(seq(1,ncol(com))[!as.logical(sapply(com, is.numeric))])
for (i in rmcols) com[[i]] <- NULL

# print out any non-numeric values (used to check previous)
for(i in 1:ncol(com)) {
  if(!is.numeric(com[,i])) {
    print(names(com)[i])
  }
}
```

```

# attach for easy use
attach(com)

######################################
##                                  ##
##    FINDING AN INITIAL MODEL      ##
##                                  ##
######################################

# make it easier to create full model with a starting point using all of them
model = "lm("
for(i in 1:ncol(com)) {

  model = paste(model, names(com)[i])
  if (i != ncol(com)) {
    model = paste(model, "+")
  }
}

# get a starting point using step—wise
null.model = lm( ViolentCrimesPerPop ~ 1)
full.model = lm( ViolentCrimesPerPop ~ ., data=com)
full.model.formula = ViolentCrimesPerPop ~ population + householdsize + racepctblack +
     racePctWhite + racePctAsian + racePctHisp + agePct12t21 + agePct12t29 +
    agePct16t24 + agePct65up + numbUrban + pctUrban + medIncome + pctWWage +
    pctWFarmSelf + pctWInvInc + pctWSocSec + pctWPubAsst + pctWRetire + medFamInc +
    perCapInc + whitePerCap + blackPerCap + indianPerCap + AsianPerCap + HispPerCap +
    NumUnderPov + PctPopUnderPov + PctLess9thGrade + PctNotHSGrad + PctBSorMore +
    PctUnemployed + PctEmploy + PctEmplManu + PctEmplProfServ + PctOccupManu +
    PctOccupMgmtProf + MalePctDivorce + MalePctNevMarr + FemalePctDiv + TotalPctDiv +
    PersPerFam + PctFam2Par + PctKids2Par + PctYoungKids2Par + PctTeen2Par +
    PctWorkMomYoungKids + PctWorkMom + NumIlleg + PctIlleg + NumImmig + PctImmigRecent
    + PctImmigRec5 + PctImmigRec8 + PctImmigRec10 + PctRecentImmig + PctRecImmig5 +
    PctRecImmig8 + PctRecImmig10 + PctSpeakEnglOnly + PctNotSpeakEnglWell +
    PctLargHouseFam + PctLargHouseOccup + PersPerOccupHous + PersPerOwnOccHous +
    PersPerRentOccHous + PctPersOwnOccup + PctPersDenseHous + PctHousLess3BR + MedNumBR
     + HousVacant + PctHousOccup + PctHousOwnOcc + PctVacantBoarded + PctVacMore6Mos +
    MedYrHousBuilt + PctHousNoPhone + PctWOFullPlumb + OwnOccLowQuart + OwnOccMedVal +
    OwnOccHiQuart + RentLowQ + RentMedian + RentHighQ + MedRent + MedRentPctHousInc +
    MedOwnCostPctInc + MedOwnCostPctIncNoMtg + NumInShelters + NumStreet +
    PctForeignBorn + PctBornSameState + PctSameHouse85 + PctSameCity85 + PctSameState85
     + LandArea + PopDens + PctUsePubTrans + LemasPctOfficDrugUn
step(null.model, full.model.formula, direction="forward") # gets model with AIC =
    −8027.3
step(full.model, full.model.formula, direction="backward") # gets model with AIC =
    −8043.5
step( null.model, full.model.formula) # gets same model as forward

# testing the regression returned from forward
test.reg = lm(formula = ViolentCrimesPerPop ~ PctKids2Par + racePctWhite +
    HousVacant + pctUrban + PctWorkMom + NumStreet + MalePctDivorce +     PctIlleg +
    numbUrban + PctPersDenseHous + racepctblack +     agePct12t29 +
    MedOwnCostPctIncNoMtg + PctPopUnderPov + pctWRetire +     MedRentPctHousInc +
    RentLowQ + MedRent + pctWWage + whitePerCap +     MalePctNevMarr + PctEmploy +
    PctEmplManu + TotalPctDiv +     perCapInc + pctWInvInc + LemasPctOfficDrugUn +
    MedOwnCostPctInc +     PctVacMore6Mos + PctVacantBoarded + HispPerCap +
    pctWFarmSelf +     indianPerCap + PctLess9thGrade + PctLargHouseFam + agePct12t21 +
         AsianPerCap)
summary(test.reg)
par(mfrow=c(2,2))
plot(test.reg)

# testing the regression returned from backward
test.reg2 = lm(formula = ViolentCrimesPerPop ~ racepctblack + racePctHisp +
    agePct12t29 + pctUrban + pctWWage + pctWFarmSelf + pctWInvInc +     pctWSocSec +
    pctWRetire + medFamInc + whitePerCap + indianPerCap +     HispPerCap +
    PctPopUnderPov + PctLess9thGrade + PctEmploy +     PctEmplManu + PctOccupManu +
    PctOccupMgmtProf + MalePctDivorce +     MalePctNevMarr + TotalPctDiv + PctKids2Par
    + PctWorkMom +     NumIlleg + PctIlleg + NumImmig + PctNotSpeakEnglWell +
    PctLargHouseOccup +     PersPerOccupHous + PersPerRentOccHous + PctPersOwnOccup +
        PctPersDenseHous + PctHousLess3BR + MedNumBR + HousVacant +     PctHousOccup +
    PctHousOwnOcc + PctVacantBoarded + PctVacMore6Mos +     OwnOccLowQuart +
    OwnOccMedVal + RentLowQ + RentHighQ + MedRent +     MedRentPctHousInc +
    MedOwnCostPctInc + MedOwnCostPctIncNoMtg +     NumInShelters + NumStreet +
    PctForeignBorn + PctUsePubTrans +     LemasPctOfficDrugUn)
summary(test.reg2)
par(mfrow=c(2,2))
plot(test.reg2)

######################################
##                                  ##
##    IMPROVING THE MODEL           ##
##                                  ##
######################################

#
# The backward regression is (very) slightly better with R^2 and AIC. The residual
    plots are nearly identical.
# Due to Ockam's Razor, there is no point to go with this much more complex model.
# We will try to use the best parts from both models in a simpler model below.
#
simpler.reg = lm(formula = ViolentCrimesPerPop ~ PctKids2Par + HousVacant + pctUrban +
     PctWorkMom + NumStreet + MalePctDivorce +     PctIlleg + numbUrban +
    PctPersDenseHous + racepctblack + MedOwnCostPctIncNoMtg + RentLowQ + MedRent )
summary(simpler.reg)
pdf("initial_model.pdf")
par(mfrow=c(2,2))
plot(simpler.reg)
dev.off()

# Start looking at transformations of these variables
new_data = subset(com, select=c(ViolentCrimesPerPop, PctKids2Par, HousVacant, pctUrban
    , PctWorkMom, NumStreet, MalePctDivorce, PctIlleg, numbUrban, PctPersDenseHous,
    racepctblack, MedOwnCostPctIncNoMtg, RentLowQ, MedRent))

#
# This one seems much better.
```

```
#
library(MASS)
pdf("boxcox.pdf")
boxcox(log_fix(ViolentCrimesPerPop) ~ PctKids2Par + HousVacant + pctUrban + PctWorkMom + NumStreet + MalePctDivorce + PctIlleg + numbUrban + PctPersDenseHous + racepctblack + MedOwnCostPctIncNoMtg + RentLowQ + MedRent, lambda=seq(0.25, 0.3, 1/40))
dev.off()

library(gam)
gam.stuff = gam(ViolentCrimesPerPop ~ s(PctKids2Par) + s(HousVacant) + s(pctUrban) + s(PctWorkMom) + s(NumStreet) + s(MalePctDivorce) + s(PctIlleg) + s(numbUrban) + s(PctPersDenseHous) + s(racepctblack) + s(MedOwnCostPctIncNoMtg) + s(RentLowQ) + s(MedRent))
#pdf("gam.pdf")
par(mfrow=c(4, 4))
plot(gam.stuff)
#dev.off()

# test polynomial transformations that GAM plots suggest
# need to use fix for 0 entries because log(0) causes issues
simpler.reg_trans = lm(formula = (ViolentCrimesPerPop^0.27-1)/0.27 ~ log_trans_n_fix(PctKids2Par) + log_trans_n_fix(HousVacant) + poly(pctUrban, 5) + log_trans_n_fix(PctWorkMom) + poly(NumStreet, 3) + log_trans_n_fix(MalePctDivorce) + poly(PctIlleg, 2) + poly(numbUrban, 2) + log_trans_n_fix(PctPersDenseHous) + poly(racepctblack, 2) + log_trans_n_fix(MedOwnCostPctIncNoMtg) + log_trans_n_fix(RentLowQ) + log_trans_n_fix(MedRent))
summary(simpler.reg_trans)
#pdf("transformed.pdf")
par(mfrow=c(2,2))
plot(simpler.reg_trans)
#dev.off()

# This "final" linear model has a much improved residual plot from the originals returned from step-wise, with an R^2 which is only 0.15 less.
# The model is also not too complex with only a dozen variables.

#####################################
##                                 ##
##    OUTLIER IDENTIFICATION       ##
##                                 ##
#####################################

# outliers from R's basic outlier test
library(car)
outlierTest(simpler.reg)
original_com[c(83, 376, 616, 401, 958, 150),]$communityname # Vernon, TX; La Canada Flintridge, CA; Glens Falls; Mansfield; West Hollywood; Plant City
# cook distance to find influential points
cutoff <- 4/((nrow(com)-length(simpler.reg$coefficients)-2))
plot(simpler.reg, which=4, cook.levels=cutoff)
original_com[c(683, 1699),]$communityname  # Philadelphia and Ft. Lauderdale
# lrplot to show influential outliers
source("http://www.stat.auckland.ac.nz/~lee/330/R330.txt")
```

```r
par(mfrow=c(1,1))
lrplot(simpler.reg)
original_com[c(1333, 1035),]$communityname # Ocean City, Gatesville are influential outliers not mentioned in above two analyses

# see how removing outliers affects regression
without.outliers = com[c(-83, -376, -616, -401, -958, -150, -683, -1699, -1333, -1035),]
without.outlier.reg = lm(formula = (ViolentCrimesPerPop^0.27-1)/0.27 ~ log_trans_n_fix(PctKids2Par) + log_trans_n_fix(HousVacant) + poly(pctUrban, 5) + log_trans_n_fix(PctWorkMom) + poly(NumStreet, 3) + log_trans_n_fix(MalePctDivorce) + poly(PctIlleg, 2) + poly(numbUrban, 2) + log_trans_n_fix(PctPersDenseHous) + poly(racepctblack, 2) + log_trans_n_fix(MedOwnCostPctIncNoMtg) + log_trans_n_fix(RentLowQ) + log_trans_n_fix(MedRent), data=without.outliers)
summary(without.outlier.reg)
par(mfrow=c(1,2))
plot(simpler.reg_trans, which=2, main="Transformed Model")
plot(without.outlier.reg, which=2, main="Without Outliers Model") # plot turns out pretty much the same; the Q-Q plot did not improve

# test recursively removing outliers, go again
outlierTest(without.outlier.reg)
without_outliers = com[c(-1231,-774,-343,-530,-1463),]
without.outlier.reg2 = lm(formula = (ViolentCrimesPerPop^0.27-1)/0.27 ~ log_trans_n_fix(PctKids2Par) + log_trans_n_fix(HousVacant) + poly(pctUrban, 5) + log_trans_n_fix(PctWorkMom) + poly(NumStreet, 3) + log_trans_n_fix(MalePctDivorce) + poly(PctIlleg, 2) + poly(numbUrban, 2) + log_trans_n_fix(PctPersDenseHous) + poly(racepctblack, 2) + log_trans_n_fix(MedOwnCostPctIncNoMtg) + log_trans_n_fix(RentLowQ) + log_trans_n_fix(MedRent), data=without.outliers)
summary(without.outlier.reg2)
par(mfrow=c(2,2))
plot(without.outlier.reg2) # doesn't do anything

#####################################
##                                 ##
##    PERFORMING THE DATA SPLIT    ##
##                                 ##
#####################################

com_low = data.frame(matrix(nrow = 484 + 420, ncol = 124))
names(com_low) <- as.vector(colnames(com))
com_high = data.frame(matrix(nrow = nrow(com) - 484 - 420, ncol = 124))
names(com_high) <- as.vector(colnames(com))
highct = 1
lowct = 1
for(i in 1:nrow(com))
{
  if(com$ViolentCrimesPerPop[i] < 0.133){ #This is approximately half of the observations
    com_low[lowct, ] = com[i, ]
    lowct = lowct + 1
  }
  else {
    com_high[highct, ] = com[i, ]
    highct = highct + 1
  }
}

### LOW CRIME ###
com_low.null.model = lm( com_low$ViolentCrimesPerPop ~ 1)
com_low.full.model = lm( com_low$ViolentCrimesPerPop ~ com_low$population + com_low$householdsize + com_low$racepctblack + com_low$racePctWhite + com_low$racePctAsian + com_low$racePctHisp + com_low$agePct12t21 + com_low$agePct12t29 + com_low$agePct16t24 + com_low$agePct65up + com_low$numbUrban + com_low$pctUrban + com_low$medIncome + com_low$pctWWage + com_low$pctWFarmSelf + com_low$pctWInvInc + com_low$pctWSocSec + com_low$pctWPubAsst + com_low$pctWRetire + com_low$medFamInc + com_low$perCapInc + com_low$whitePerCap + com_low$blackPerCap + com_low$indianPerCap + com_low$AsianPerCap + com_low$HispPerCap + com_low$NumUnderPov + com_low$PctPopUnderPov + com_low$PctLess9thGrade + com_low$PctNotHSGrad + com_low$PctBSorMore + com_low$PctUnemployed + com_low$PctEmploy + com_low$PctEmplManu + com_low$PctEmplProfServ + com_low$PctOccupManu + com_low$PctOccupMgmtProf + com_low$MalePctDivorce + com_low$MalePctNevMarr + com_low$FemalePctDiv + com_low$TotalPctDiv + com_low$PersPerFam + com_low$PctFam2Par + com_low$PctKids2Par + com_low$PctYoungKids2Par + com_low$PctTeen2Par + com_low$PctWorkMomYoungKids + com_low$PctWorkMom + com_low$NumIlleg + com_low$PctIlleg + com_low$NumImmig + com_low$PctImmigRecent + com_low$PctImmigRec5 + com_low$PctImmigRec8 + com_low$PctImmigRec10 + com_low$PctRecentImmig + com_low$PctRecImmig5 + com_low$PctRecImmig8 + com_low$PctRecImmig10 + com_low$PctSpeakEnglOnly + com_low$PctNotSpeakEnglWell + com_low$PctLargHouseFam + com_low$PctLargHouseOccup + com_low$PersPerOccupHous + com_low$PersPerOwnOccHous + com_low$PersPerRentOccHous + com_low$PctPersOwnOccup + com_low$PctPersDenseHous + com_low$PctHousLess3BR + com_low$MedNumBR + com_low$HousVacant + com_low$PctHousOccup + com_low$PctHousOwnOcc + com_low$PctVacantBoarded + com_low$PctVacMore6Mos + com_low$MedYrHousBuilt + com_low$PctHousNoPhone + com_low$PctWOFullPlumb + com_low$OwnOccLowQuart + com_low$OwnOccMedVal + com_low$OwnOccHiQuart + com_low$RentLowQ + com_low$RentMedian + com_low$RentHighQ + com_low$MedRent + com_low$MedRentPctHousInc + com_low$MedOwnCostPctInc + com_low$MedOwnCostPctIncNoMtg + com_low$NumInShelters + com_low$NumStreet + com_low$PctForeignBorn + com_low$PctBornSameState + com_low$PctSameHouse85 + com_low$PctSameCity85 + com_low$PctSameState85 + com_low$LandArea + com_low$PopDens + com_low$PctUsePubTrans + com_low$LemasPctOfficDrugUn)
com_low.full.model.formula = com_low$ViolentCrimesPerPop ~ com_low$population + com_low$householdsize + com_low$racepctblack + com_low$racePctWhite + com_low$racePctAsian + com_low$racePctHisp + com_low$agePct12t21 + com_low$agePct12t29 + com_low$agePct16t24 + com_low$agePct65up + com_low$numbUrban + com_low$pctUrban + com_low$medIncome + com_low$pctWWage + com_low$pctWFarmSelf + com_low$pctWInvInc + com_low$pctWSocSec + com_low$pctWPubAsst + com_low$pctWRetire + com_low$medFamInc + com_low$perCapInc + com_low$whitePerCap + com_low$blackPerCap + com_low$indianPerCap + com_low$AsianPerCap + com_low$HispPerCap + com_low$NumUnderPov + com_low$PctPopUnderPov + com_low$PctLess9thGrade + com_low$PctNotHSGrad + com_low$PctBSorMore + com_low$PctUnemployed + com_low$PctEmploy + com_low$PctEmplManu + com_low$PctEmplProfServ + com_low$PctOccupManu + com_low$PctOccupMgmtProf + com_low$MalePctDivorce + com_low$MalePctNevMarr + com_low$FemalePctDiv + com_low$TotalPctDiv + com_low$PersPerFam + com_low$PctFam2Par + com_low$PctKids2Par + com_low$PctYoungKids2Par + com_low$PctTeen2Par + com_low$PctWorkMomYoungKids + com_low$PctWorkMom + com_low$NumIlleg + com_low$PctIlleg + com_low$NumImmig + com_low$PctImmigRecent + com_low$PctImmigRec5 + com_low$PctImmigRec8 + com_low$PctImmigRec10 + com_low$PctRecentImmig + com_low$PctRecImmig5 + com_low$PctRecImmig8 + com_low$PctRecImmig10 + com_low$PctSpeakEnglOnly + com_low$
```

```r
PctNotSpeakEnglWell + com_low$PctLargHouseFam + com_low$PctLargHouseOccup + com_low$PersPerOccupHous + com_low$PersPerOwnOccHous + com_low$PersPerRentOccHous + com_low$PctPersOwnOccup + com_low$PctPersDenseHous + com_low$PctHousLess3BR + com_low$MedNumBR + com_low$HousVacant + com_low$PctHousOccup + com_low$PctHousOwnOcc + com_low$PctVacantBoarded + com_low$PctVacMore6Mos + com_low$MedYrHousBuilt + com_low$PctHousNoPhone + com_low$PctWOFullPlumb + com_low$OwnOccLowQuart + com_low$OwnOccMedVal + com_low$OwnOccHiQuart + com_low$RentLowQ + com_low$RentMedian + com_low$RentHighQ + com_low$MedRent + com_low$MedRentPctHousInc + com_low$MedOwnCostPctInc + com_low$MedOwnCostPctIncNoMtg + com_low$NumInShelters + com_low$NumStreet + com_low$PctForeignBorn + com_low$PctBornSameState + com_low$PctSameHouse85 + com_low$PctSameCity85 + com_low$PctSameState85 + com_low$LandArea + com_low$PopDens + com_low$PctUsePubTrans + com_low$LemasPctOfficDrugUn



step(com_low.null.model, com_low.full.model.formula, direction="forward")
com_low.for.reg = lm(formula = com_low$ViolentCrimesPerPop ~ com_low$FemalePctDiv +     com_low$racePctWhite + com_low$NumIlleg + com_low$PctBSorMore +     com_low$PctOccupManu + com_low$LandArea + com_low$HousVacant +     com_low$pctUrban + com_low$PctKids2Par + com_low$racepctblack +     com_low$PctWOFullPlumb + com_low$pctWInvInc + com_low$MalePctDivorce +     com_low$PctLess9thGrade + com_low$racePctHisp + com_low$agePct65up +     com_low$pctWRetire + com_low$PctWorkMom + com_low$PctEmplProfServ +     com_low$NumStreet + com_low$agePct12t21 + com_low$PctImmigRec5 +     com_low$TotalPctDiv)
# AIC = -6418.7
step(com_low.full.model, com_low.full.model.formula, direction="backward")
com_low.bac.reg = lm(formula = com_low$ViolentCrimesPerPop ~ com_low$population + com_low$racepctblack + com_low$racePctWhite + com_low$racePctHisp +     com_low$agePct12t21 + com_low$agePct65up + com_low$numbUrban +     com_low$pctUrban + com_low$medIncome + com_low$pctWInvInc +     com_low$pctWPubAsst + com_low$pctWRetire +  com_low$medFamInc +     com_low$PctPopUnderPov + com_low$PctLess9thGrade + com_low$PctEmploy +     com_low$PctEmplProfServ + com_low$PctOccupManu + com_low$MalePctDivorce +     com_low$FemalePctDiv + com_low$TotalPctDiv + com_low$PersPerFam +     com_low$PctKids2Par + com_low$PctImmigRec5 + com_low$PctImmigRec8 +     com_low$PctRecImmig5 + com_low$PctRecImmig10 + com_low$PctLargHouseFam + com_low$PctLargHouseOccup + com_low$PersPerOccupHous + com_low$PersPerOwnOccHous +     com_low$HousVacant + com_low$PctWOFullPlumb + com_low$OwnOccMedVal +     com_low$RentHighQ + com_low$MedRent + com_low$NumStreet +     com_low$PctBornSameState + com_low$PctSameState85 + com_low$LandArea)
# AIC = -6433.5

summary(com_low.for.reg)
summary(com_low.bac.reg)
par(mfrow=c(2,2))
plot(com_low.for.reg)
plot(com_low.bac.reg) # very good looking residuals, too bad R^2 is much lower than original model
##Adjustesd R-Squared 0.38 -- Result of backward regression.
##F-statistic 14.84
##PctPopUnderPov, FemalePctDiv, PctRecImmig5, PersPerOwnOccHous insignificant
#####BEST######
#####BECOMES WORSE FROM HERE ON#####

### HIGH CRIME ###
com_high.null.model = lm( com_high$ViolentCrimesPerPop ~ 1)
com_high.full.model = lm( com_high$ViolentCrimesPerPop ~ com_high$population + com_
    high$householdsize + com_high$racepctblack + com_high$racePctWhite + com_high$
    racePctAsian + com_high$racePctHisp + com_high$agePct12t21 + com_high$agePct12t29 +
     com_high$agePct16t24 + com_high$agePct65up + com_high$numbUrban + com_high$
    pctUrban + com_high$medIncome + com_high$pctWWage + com_high$pctWFarmSelf + com_
    high$pctWInvInc + com_high$pctWSocSec + com_high$pctWPubAsst + com_high$pctWRetire
    + com_high$medFamInc + com_high$perCapInc + com_high$whitePerCap + com_high$
    blackPerCap + com_high$indianPerCap + com_high$AsianPerCap + com_high$HispPerCap +
    com_high$NumUnderPov + com_high$PctPopUnderPov + com_high$PctLess9thGrade + com_
    high$PctNotHSGrad + com_high$PctBSorMore + com_high$PctUnemployed + com_high$
    PctEmploy + com_high$PctEmplManu + com_high$PctEmplProfServ + com_high$PctOccupManu
     + com_high$PctOccupMgmtProf + com_high$MalePctDivorce + com_high$MalePctNevMarr +
    com_high$FemalePctDiv + com_high$TotalPctDiv + com_high$PersPerFam + com_high$
    PctFam2Par + com_high$PctKids2Par + com_high$PctYoungKids2Par + com_high$
    PctTeen2Par + com_high$PctWorkMomYoungKids + com_high$PctWorkMom + com_high$
    NumIlleg + com_high$PctIlleg + com_high$NumImmig + com_high$PctImmigRecent + com_
    high$PctImmigRec5 + com_high$PctImmigRec8 + com_high$PctImmigRec10 + com_high$
    PctRecentImmig + com_high$PctRecImmig5 + com_high$PctRecImmig8 + com_high$
    PctRecImmig10 + com_high$PctSpeakEnglOnly + com_high$PctNotSpeakEnglWell + com_high
    $PctLargHouseFam + com_high$PctLargHouseOccup + com_high$PersPerOccupHous + com_
    high$PersPerOwnOccHous + com_high$PersPerRentOccHous + com_high$PctPersOwnOccup +
    com_high$PctPersDenseHous + com_high$PctHousLess3BR + com_high$MedNumBR + com_high$
    HousVacant + com_high$PctHousOccup + com_high$PctHousOwnOcc + com_high$
    PctVacantBoarded + com_high$PctVacMore6Mos + com_high$MedYrHousBuilt + com_high$
    PctHousNoPhone + com_high$PctWOFullPlumb + com_high$OwnOccLowQuart + com_high$
    OwnOccMedVal + com_high$OwnOccHiQuart + com_high$RentLowQ + com_high$RentMedian +
    com_high$RentHighQ + com_high$MedRent + com_high$MedRentPctHousInc + com_high$
    MedOwnCostPctInc + com_high$MedOwnCostPctIncNoMtg + com_high$NumInShelters + com_
    high$NumStreet + com_high$PctForeignBorn + com_high$PctBornSameState + com_high$
    PctSameHouse85 + com_high$PctSameCity85 + com_high$PctSameState85 + com_high$
    LandArea + com_high$PopDens + com_high$PctUsePubTrans + com_high$
    LemasPctOfficDrugUn)
com_high.full.model.formula = com_high$ViolentCrimesPerPop ~ com_high$population + com
    _high$householdsize + com_high$racepctblack + com_high$racePctWhite + com_high$
    racePctAsian + com_high$racePctHisp + com_high$agePct12t21 + com_high$agePct12t29 +
     com_high$agePct16t24 + com_high$agePct65up + com_high$numbUrban + com_high$
    pctUrban + com_high$medIncome + com_high$pctWWage + com_high$pctWFarmSelf + com_
    high$pctWInvInc + com_high$pctWSocSec + com_high$pctWPubAsst + com_high$pctWRetire
    + com_high$medFamInc + com_high$perCapInc + com_high$whitePerCap + com_high$
    blackPerCap + com_high$indianPerCap + com_high$AsianPerCap + com_high$HispPerCap +
    com_high$NumUnderPov + com_high$PctPopUnderPov + com_high$PctLess9thGrade + com_
    high$PctNotHSGrad + com_high$PctBSorMore + com_high$PctUnemployed + com_high$
    PctEmploy + com_high$PctEmplManu + com_high$PctEmplProfServ + com_high$PctOccupManu
     + com_high$PctOccupMgmtProf + com_high$MalePctDivorce + com_high$MalePctNevMarr +
    com_high$FemalePctDiv + com_high$TotalPctDiv + com_high$PersPerFam + com_high$
    PctFam2Par + com_high$PctKids2Par + com_high$PctYoungKids2Par + com_high$
    PctTeen2Par + com_high$PctWorkMomYoungKids + com_high$PctWorkMom + com_high$
    NumIlleg + com_high$PctIlleg + com_high$NumImmig + com_high$PctImmigRecent + com_
    high$PctImmigRec5 + com_high$PctImmigRec8 + com_high$PctImmigRec10 + com_high$
    PctRecentImmig + com_high$PctRecImmig5 + com_high$PctRecImmig8 + com_high$
    PctRecImmig10 + com_high$PctSpeakEnglOnly + com_high$PctNotSpeakEnglWell + com_high
    $PctLargHouseFam + com_high$PctLargHouseOccup + com_high$PersPerOccupHous + com_
    high$PersPerOwnOccHous + com_high$PersPerRentOccHous + com_high$PctPersOwnOccup +
    com_high$PctPersDenseHous + com_high$PctHousLess3BR + com_high$MedNumBR + com_high$
    HousVacant + com_high$PctHousOccup + com_high$PctHousOwnOcc + com_high$
    PctVacantBoarded + com_high$PctVacMore6Mos + com_high$MedYrHousBuilt + com_high$
    PctHousNoPhone + com_high$PctWOFullPlumb + com_high$OwnOccLowQuart + com_high$
    OwnOccMedVal + com_high$OwnOccHiQuart + com_high$RentLowQ + com_high$RentMedian +
    com_high$RentHighQ + com_high$MedRent + com_high$MedRentPctHousInc + com_high$
    MedOwnCostPctInc + com_high$MedOwnCostPctIncNoMtg + com_high$NumInShelters + com_
    high$NumStreet + com_high$PctForeignBorn + com_high$PctBornSameState + com_high$
    PctSameHouse85 + com_high$PctSameCity85 + com_high$PctSameState85 + com_high$
    LandArea + com_high$PopDens + com_high$PctUsePubTrans + com_high$
    LemasPctOfficDrugUn

step(com_high.null.model, com_high.full.model.formula, direction="forward")
com_high.for.reg = lm(formula = com_high$ViolentCrimesPerPop ~ com_high$PctIlleg +
        com_high$HousVacant + com_high$PctKids2Par + com_high$PctPersDenseHous +
    com_high$HispPerCap + com_high$MalePctDivorce + com_high$racepctblack +     com_
    high$PctWorkMom + com_high$NumStreet + com_high$population +     com_high$
    PctPopUnderPov + com_high$whitePerCap + com_high$AsianPerCap +     com_high$
    MedOwnCostPctIncNoMtg + com_high$MedYrHousBuilt +     com_high$PctHousOccup + com_
    high$pctWRetire + com_high$PctEmplManu +     com_high$pctWSocSec + com_high$
    numbUrban + com_high$MedNumBR +     com_high$MalePctNevMarr + com_high$pctWFarmSelf
     + com_high$PctForeignBorn +     com_high$RentLowQ + com_high$MedRent + com_high$
    RentHighQ +     com_high$PctLess9thGrade + com_high$PctOccupManu + com_high$
    PctVacMore6Mos +     com_high$PctVacantBoarded + com_high$PctEmploy + com_high$
    agePct12t29)
# AIC = —3985.0 much worse than low model
step(com_high.full.model, com_high.full.model.formula, direction="backward")
com_high.bac.reg = lm(formula = com_high$ViolentCrimesPerPop ~ com_high$racepctblack +
         com_high$agePct12t21 + com_high$agePct16t24 + com_high$pctUrban +     com_high
    $pctWFarmSelf + com_high$pctWSocSec + com_high$pctWRetire +     com_high$
    whitePerCap + com_high$AsianPerCap + com_high$HispPerCap +     com_high$
    PctPopUnderPov + com_high$PctLess9thGrade + com_high$PctEmploy +     com_high$
    PctEmplManu + com_high$PctOccupManu + com_high$MalePctDivorce +     com_high$
    MalePctNevMarr + com_high$PctKids2Par + com_high$PctWorkMomYoungKids +     com_high
    $PctWorkMom + com_high$PctIlleg + com_high$NumImmig +     com_high$PctRecImmig5 +
    com_high$PctRecImmig8 + com_high$PctSpeakEnglOnly +     com_high$PersPerOccupHous +
     com_high$PersPerRentOccHous +     com_high$PctPersOwnOccup + com_high$
    PctHousLess3BR + com_high$MedNumBR +     com_high$HousVacant + com_high$
    PctHousOccup + com_high$PctHousOwnOcc +     com_high$OwnOccLowQuart + com_high$
    OwnOccHiQuart + com_high$RentLowQ +     com_high$RentHighQ + com_high$MedRent + com
    _high$MedRentPctHousInc +     com_high$MedOwnCostPctInc + com_high$
    MedOwnCostPctIncNoMtg +     com_high$NumInShelters + com_high$NumStreet + com_high$
    PctUsePubTrans)
# AIC = —3988.9

summary(com_high.for.reg)
summary(com_high.bac.reg)
par(mfrow=c(2,2))
plot(com_high.bac.reg) # not looking as good as low model
# again backward stepwise is better
##Adjusted R—Squared 0.534
##F—statistic 29.34
##agePct12t21, agePct16t24, pctWFarmSElf, AsianPerCap, SpeakEnglOnly, PersPerOccupHous
    , PersPerRentOccHous, PersPerOwnOccup, PctHousLess3BR, MedNumBR, OwnOccLowQuart,
    RentHighQ, MedRentPctHousInc, PctUsePubTrans DEFINITELY not significant
```

```
## pctWSocSec, HispPerCap, PctRecImmig8, barely significant

com_high.reg2 = lm(com_high$ViolentCrimesPerPop ~ com_high$racepctblack + com_high$pctUrban + com_high$pctWRetire + com_high$whitePerCap + com_high$PctPopUnderPov + com_high$PctLess9thGrade + com_high$PctEmploy + com_high$PctEmplManu + com_high$PctOccupManu + com_high$MalePctDivorce + com_high$MalePctNevMarr + com_high$PctKids2Par + com_high$PctWorkMomYoungKids + com_high$PctWorkMom + com_high$PctIlleg + com_high$NumImmig + com_high$PctRecImmig5 + com_high$MedNumBR + com_high$HousVacant + com_high$PctHousOccup + com_high$PctHousOwnOcc + com_high$OwnOccHiQuart + com_high$RentLowQ + com_high$MedRent + com_high$MedOwnCostPctInc + com_high$MedOwnCostPctIncNoMtg + com_high$NumInShelters + com_high$NumStreet)
summary(com_high.reg2)
anova(com_high.reg2)
##Adjusted R-Squared 0.508
##F-statistic 41.13
####BECOMES WORSE#####

#### SPLIT AT 0.5 #####
new_com  <- na.omit(com)
lowct = 1
highct = 1
com_low = data.frame(matrix(nrow = 1705, ncol = 124))
names(com_low) <-  as.vector(colnames(com))
com_high = data.frame(matrix(nrow = nrow(com) - 1705, ncol = 124))
names(com_high) <-  as.vector(colnames(com))
for(i in 1:nrow(new_com))
{
  if(com$ViolentCrimesPerPop[i] < 0.5)
  {
    com_low[lowct, ] = new_com[i, ]
    lowct = lowct + 1
  }
  else
  {
    com_high[highct, ] = new_com[i, ]
    highct = highct + 1
  }
}

### LOW CRIME ###
com_low.null.model = lm( com_low$ViolentCrimesPerPop ~ 1)
com_low.full.model = lm( com_low$ViolentCrimesPerPop ~ com_low$population + com_low$householdsize + com_low$racepctblack + com_low$racePctWhite + com_low$racePctAsian + com_low$racePctHisp + com_low$agePct12t21 + com_low$agePct12t29 + com_low$agePct16t24 + com_low$agePct65up + com_low$numbUrban + com_low$pctUrban + com_low$medIncome + com_low$pctWWage + com_low$pctWFarmSelf + com_low$pctWInvInc + com_low$pctWSocSec + com_low$pctWPubAsst + com_low$pctWRetire + com_low$medFamInc + com_low$perCapInc + com_low$whitePerCap + com_low$blackPerCap + com_low$indianPerCap + com_low$AsianPerCap + com_low$HispPerCap + com_low$NumUnderPov + com_low$PctPopUnderPov + com_low$PctLess9thGrade + com_low$PctNotHSGrad + com_low$PctBSorMore + com_low$PctUnemployed + com_low$PctEmploy + com_low$PctEmplManu + com_low$PctEmplProfServ + com_low$PctOccupManu + com_low$PctOccupMgmtProf + com_low$MalePctDivorce + com_low$MalePctNevMarr + com_low$FemalePctDiv + com_low$TotalPctDiv + com_low$PersPerFam + com_low$PctFam2Par + com_low$PctKids2Par + com_
    low$PctYoungKids2Par + com_low$PctTeen2Par + com_low$PctWorkMomYoungKids + com_low$
    PctWorkMom + com_low$NumIlleg + com_low$PctIlleg + com_low$NumImmig + com_low$
    PctImmigRecent + com_low$PctImmigRec5 + com_low$PctImmigRec8 + com_low$
    PctImmigRec10 + com_low$PctRecentImmig + com_low$PctRecImmig5 + com_low$
    PctRecImmig8 + com_low$PctRecImmig10 + com_low$PctSpeakEnglOnly + com_low$
    PctNotSpeakEnglWell + com_low$PctLargHouseFam + com_low$PctLargHouseOccup + com_low
    $PersPerOccupHous + com_low$PersPerOwnOccHous + com_low$PersPerRentOccHous + com_
    low$PctPersOwnOccup + com_low$PctPersDenseHous + com_low$PctHousLess3BR + com_low$
    MedNumBR + com_low$HousVacant + com_low$PctHousOccup + com_low$PctHousOwnOcc + com_
    low$PctVacantBoarded + com_low$PctVacMore6Mos + com_low$MedYrHousBuilt + com_low$
    PctHousNoPhone + com_low$PctWOFullPlumb + com_low$OwnOccLowQuart + com_low$
    OwnOccMedVal + com_low$OwnOccHiQuart + com_low$RentLowQ + com_low$RentMedian + com_
    low$RentHighQ + com_low$MedRent + com_low$MedRentPctHousInc + com_low$
    MedOwnCostPctInc + com_low$MedOwnCostPctIncNoMtg + com_low$NumInShelters + com_low$
    NumStreet + com_low$PctForeignBorn + com_low$PctBornSameState + com_low$
    PctSameHouse85 + com_low$PctSameCity85 + com_low$PctSameState85 + com_low$LandArea
    + com_low$PopDens + com_low$PctUsePubTrans + com_low$LemasPctOfficDrugUn)
com_low.full.model.formula = com_low$ViolentCrimesPerPop ~ com_low$population + com_
    low$householdsize + com_low$racepctblack + com_low$racePctWhite + com_low$
    racePctAsian + com_low$racePctHisp + com_low$agePct12t21 + com_low$agePct12t29 +
    com_low$agePct16t24 + com_low$agePct65up + com_low$numbUrban + com_low$pctUrban +
    com_low$medIncome + com_low$pctWWage + com_low$pctWFarmSelf + com_low$pctWInvInc +
    com_low$pctWSocSec + com_low$pctWPubAsst + com_low$pctWRetire + com_low$medFamInc +
     com_low$perCapInc + com_low$whitePerCap + com_low$blackPerCap + com_low$
    indianPerCap + com_low$AsianPerCap + com_low$HispPerCap + com_low$NumUnderPov + com
    _low$PctPopUnderPov + com_low$PctLess9thGrade + com_low$PctNotHSGrad + com_low$
    PctBSorMore + com_low$PctUnemployed + com_low$PctEmploy + com_low$PctEmplManu + com
    _low$PctEmplProfServ + com_low$PctOccupManu + com_low$PctOccupMgmtProf + com_low$
    MalePctDivorce + com_low$MalePctNevMarr + com_low$FemalePctDiv + com_low$
    TotalPctDiv + com_low$PersPerFam + com_low$PctFam2Par + com_low$PctKids2Par + com_
    low$PctYoungKids2Par + com_low$PctTeen2Par + com_low$PctWorkMomYoungKids + com_low$
    PctWorkMom + com_low$NumIlleg + com_low$PctIlleg + com_low$NumImmig + com_low$
    PctImmigRecent + com_low$PctImmigRec5 + com_low$PctImmigRec8 + com_low$
    PctImmigRec10 + com_low$PctRecentImmig + com_low$PctRecImmig5 + com_low$
    PctRecImmig8 + com_low$PctRecImmig10 + com_low$PctSpeakEnglOnly + com_low$
    PctNotSpeakEnglWell + com_low$PctLargHouseFam + com_low$PctLargHouseOccup + com_low
    $PersPerOccupHous + com_low$PersPerOwnOccHous + com_low$PersPerRentOccHous + com_
    low$PctPersOwnOccup + com_low$PctPersDenseHous + com_low$PctHousLess3BR + com_low$
    MedNumBR + com_low$HousVacant + com_low$PctHousOccup + com_low$PctHousOwnOcc + com_
    low$PctVacantBoarded + com_low$PctVacMore6Mos + com_low$MedYrHousBuilt + com_low$
    PctHousNoPhone + com_low$PctWOFullPlumb + com_low$OwnOccLowQuart + com_low$
    OwnOccMedVal + com_low$OwnOccHiQuart + com_low$RentLowQ + com_low$RentMedian + com_
    low$RentHighQ + com_low$MedRent + com_low$MedRentPctHousInc + com_low$
    MedOwnCostPctInc + com_low$MedOwnCostPctIncNoMtg + com_low$NumInShelters + com_low$
    NumStreet + com_low$PctForeignBorn + com_low$PctBornSameState + com_low$
    PctSameHouse85 + com_low$PctSameCity85 + com_low$PctSameState85 + com_low$LandArea
    + com_low$PopDens + com_low$PctUsePubTrans + com_low$LemasPctOfficDrugUn



step(com_low.null.model, com_low.full.model.formula, direction="forward")
com_low.for.reg = lm(formula = com_low$ViolentCrimesPerPop ~ com_low$PctKids2Par +
        com_low$racePctWhite + com_low$HousVacant + com_low$racePctHisp +     com_low$
    MalePctDivorce + com_low$pctUrban + com_low$PersPerRentOccHous +     com_low$
    PctWorkMom + com_low$PctIlleg + com_low$MedOwnCostPctIncNoMtg +     com_low$
    agePct65up + com_low$pctWInvInc + com_low$RentHighQ +     com_low$
    LemasPctOfficDrugUn + com_low$PctNotSpeakEnglWell +     com_low$PctEmploy + com_low
    $MedRentPctHousInc + com_low$RentLowQ +     com_low$PctForeignBorn + com_low$
    PctOccupMgmtProf + com_low$NumStreet +     com_low$OwnOccLowQuart + com_low$
    RentMedian + com_low$blackPerCap +     com_low$pctWWage + com_low$PctTeen2Par + com
    _low$PctVacMore6Mos +     com_low$PctHousOccup + com_low$MedYrHousBuilt + com_low$
    PctWOFullPlumb +     com_low$pctWFarmSelf + com_low$PctPopUnderPov + com_low$
    PctLargHouseFam +     com_low$PctPersDenseHous + com_low$PctBornSameState + com_low
    $TotalPctDiv +     com_low$PctSpeakEnglOnly + com_low$PctRecImmig10 + com_low$
    PctRecImmig8 +     com_low$pctWRetire)
# AIC = -8649.2
step(com_low.full.model, com_low.full.model.formula, direction="backward")
com_low.bac.reg = lm(formula = com_low$ViolentCrimesPerPop ~ com_low$population +
    com_low$racePctWhite + com_low$racePctHisp + com_low$agePct65up +     com_low$
    numbUrban + com_low$pctUrban + com_low$medIncome +     com_low$pctWWage + com_low$
    pctWFarmSelf + com_low$pctWInvInc +     com_low$pctWRetire + com_low$medFamInc +
    com_low$NumUnderPov +     com_low$PctPopUnderPov + com_low$PctEmploy + com_low$
    PctOccupMgmtProf +     com_low$MalePctDivorce + com_low$MalePctNevMarr + com_low$
    PersPerFam +     com_low$PctKids2Par + com_low$PctTeen2Par + com_low$PctWorkMom +
        com_low$PctIlleg + com_low$PctRecImmig8 + com_low$PctRecImmig10 +     com_low$
    PctSpeakEnglOnly + com_low$PctNotSpeakEnglWell +     com_low$PctLargHouseOccup +
    com_low$PersPerOccupHous + com_low$PctPersOwnOccup +     com_low$PctHousLess3BR +
    com_low$PctHousOccup + com_low$PctHousOwnOcc +     com_low$PctVacMore6Mos + com_low
    $MedYrHousBuilt + com_low$PctWOFullPlumb +     com_low$OwnOccLowQuart + com_low$
    RentLowQ + com_low$RentMedian +     com_low$MedRentPctHousInc + com_low$
    MedOwnCostPctIncNoMtg +     com_low$NumStreet + com_low$PctForeignBorn + com_low$
    PctBornSameState +     com_low$LemasPctOfficDrugUn)
# AIC = -8678.9

summary(com_low.for.reg)
summary(com_low.bac.reg)
par(mfrow=c(2,2))
plot(com_low.for.reg)
plot(com_low.bac.reg) # very good looking residuals, too bad R^2 is much lower than
    original model
##Adjustesd R-Squared 0.5925
##result of backwards

### HIGH CRIME ###
com_high.null.model = lm( com_high$ViolentCrimesPerPop ~ 1)
com_high.full.model = lm( com_high$ViolentCrimesPerPop ~ com_high$population + com_
    high$householdsize + com_high$racepctblack + com_high$racePctWhite + com_high$
    racePctAsian + com_high$racePctHisp + com_high$agePct12t21 + com_high$agePct12t29 +
     com_high$agePct16t24 + com_high$agePct65up + com_high$numbUrban + com_high$
    pctUrban + com_high$medIncome + com_high$pctWWage + com_high$pctWFarmSelf + com_
    high$pctWInvInc + com_high$pctWSocSec + com_high$pctWPubAsst + com_high$pctWRetire
    + com_high$medFamInc + com_high$perCapInc + com_high$whitePerCap + com_high$
    blackPerCap + com_high$indianPerCap + com_high$AsianPerCap + com_high$HispPerCap +
    com_high$NumUnderPov + com_high$PctPopUnderPov + com_high$PctLess9thGrade + com_
    high$PctNotHSGrad + com_high$PctBSorMore + com_high$PctUnemployed + com_high$
    PctEmploy + com_high$PctEmplManu + com_high$PctEmplProfServ + com_high$PctOccupManu
     + com_high$PctOccupMgmtProf + com_high$MalePctDivorce + com_high$MalePctNevMarr +
    com_high$FemalePctDiv + com_high$TotalPctDiv + com_high$PersPerFam + com_high$
    PctFam2Par + com_high$PctKids2Par + com_high$PctYoungKids2Par + com_high$
    PctTeen2Par + com_high$PctWorkMomYoungKids + com_high$PctWorkMom + com_high$
    NumIlleg + com_high$PctIlleg + com_high$NumImmig + com_high$PctImmigRecent + com_
    high$PctImmigRec5 + com_high$PctImmigRec8 + com_high$PctImmigRec10 + com_high$
    PctRecentImmig + com_high$PctRecImmig5 + com_high$PctRecImmig8 + com_high$
    PctRecImmig10 + com_high$PctSpeakEnglOnly + com_high$PctNotSpeakEnglWell + com_high
    $PctLargHouseFam + com_high$PctLargHouseOccup + com_high$PersPerOccupHous + com_
    high$PersPerOwnOccHous + com_high$PersPerRentOccHous + com_high$PctPersOwnOccup +
    com_high$PctPersDenseHous + com_high$PctHousLess3BR + com_high$MedNumBR + com_high$
    HousVacant + com_high$PctHousOccup + com_high$PctHousOwnOcc + com_high$
    PctVacantBoarded + com_high$PctVacMore6Mos + com_high$MedYrHousBuilt + com_high$
    PctHousNoPhone + com_high$PctWOFullPlumb + com_high$OwnOccLowQuart + com_high$
    OwnOccMedVal + com_high$OwnOccHiQuart + com_high$RentLowQ + com_high$RentMedian +
    com_high$RentHighQ + com_high$MedRent + com_high$MedRentPctHousInc + com_high$
    MedOwnCostPctInc + com_high$MedOwnCostPctIncNoMtg + com_high$NumInShelters + com_
    high$NumStreet + com_high$PctForeignBorn + com_high$PctBornSameState + com_high$
    PctSameHouse85 + com_high$PctSameCity85 + com_high$PctSameState85 + com_high$
    LandArea + com_high$PopDens + com_high$PctUsePubTrans + com_high$
    LemasPctOfficDrugUn)
com_high.full.model.formula = com_high$ViolentCrimesPerPop ~ com_high$population + com
    _high$householdsize + com_high$racepctblack + com_high$racePctWhite + com_high$
    racePctAsian + com_high$racePctHisp + com_high$agePct12t21 + com_high$agePct12t29 +
     com_high$agePct16t24 + com_high$agePct65up + com_high$numbUrban + com_high$
    pctUrban + com_high$medIncome + com_high$pctWWage + com_high$pctWFarmSelf + com_
    high$pctWInvInc + com_high$pctWSocSec + com_high$pctWPubAsst + com_high$pctWRetire
    + com_high$medFamInc + com_high$perCapInc + com_high$whitePerCap + com_high$
    blackPerCap + com_high$indianPerCap + com_high$AsianPerCap + com_high$HispPerCap +
    com_high$NumUnderPov + com_high$PctPopUnderPov + com_high$PctLess9thGrade + com_
    high$PctNotHSGrad + com_high$PctBSorMore + com_high$PctUnemployed + com_high$
    PctEmploy + com_high$PctEmplManu + com_high$PctEmplProfServ + com_high$PctOccupManu
     + com_high$PctOccupMgmtProf + com_high$MalePctDivorce + com_high$MalePctNevMarr +
    com_high$FemalePctDiv + com_high$TotalPctDiv + com_high$PersPerFam + com_high$
    PctFam2Par + com_high$PctKids2Par + com_high$PctYoungKids2Par + com_high$
    PctTeen2Par + com_high$PctWorkMomYoungKids + com_high$PctWorkMom + com_high$
    NumIlleg + com_high$PctIlleg + com_high$NumImmig + com_high$PctImmigRecent + com_
    high$PctImmigRec5 + com_high$PctImmigRec8 + com_high$PctImmigRec10 + com_high$
    PctRecentImmig + com_high$PctRecImmig5 + com_high$PctRecImmig8 + com_high$
    PctRecImmig10 + com_high$PctSpeakEnglOnly + com_high$PctNotSpeakEnglWell + com_high
    $PctLargHouseFam + com_high$PctLargHouseOccup + com_high$PersPerOccupHous + com_
    high$PersPerOwnOccHous + com_high$PersPerRentOccHous + com_high$PctPersOwnOccup +
    com_high$PctPersDenseHous + com_high$PctHousLess3BR + com_high$MedNumBR + com_high$
    HousVacant + com_high$PctHousOccup + com_high$PctHousOwnOcc + com_high$
    PctVacantBoarded + com_high$PctVacMore6Mos + com_high$MedYrHousBuilt + com_high$
    PctHousNoPhone + com_high$PctWOFullPlumb + com_high$OwnOccLowQuart + com_high$
    OwnOccMedVal + com_high$OwnOccHiQuart + com_high$RentLowQ + com_high$RentMedian +
    com_high$RentHighQ + com_high$MedRent + com_high$MedRentPctHousInc + com_high$
    MedOwnCostPctInc + com_high$MedOwnCostPctIncNoMtg + com_high$NumInShelters + com_
    high$NumStreet + com_high$PctForeignBorn + com_high$PctBornSameState + com_high$
    PctSameHouse85 + com_high$PctSameCity85 + com_high$PctSameState85 + com_high$
    LandArea + com_high$PopDens + com_high$PctUsePubTrans + com_high$
    LemasPctOfficDrugUn

step(com_high.null.model, com_high.full.model.formula, direction="forward")
com_high.for.reg = lm(formula = com_high$ViolentCrimesPerPop ~ com_high$PctKids2Par +
        com_high$blackPerCap + com_high$MalePctDivorce + com_high$PctPersDenseHous +
        com_high$TotalPctDiv + com_high$racepctblack + com_high$MalePctNevMarr +
    com_high$pctWWage + com_high$MedOwnCostPctInc + com_high$racePctAsian +     com_
    high$pctWSocSec + com_high$OwnOccLowQuart + com_high$LemasPctOfficDrugUn +     com_
    high$LandArea)
# AIC = -1123.4 MUCH MUCH MUCH WORSE --- wow this is ugly
step(com_high.full.model, com_high.full.model.formula, direction="backward")
com_high.bac.reg = lm(formula = com_high$ViolentCrimesPerPop ~ com_high$population +
        com_high$householdsize + com_high$racepctblack + com_high$racePctAsian +
    com_high$agePct12t21 + com_high$agePct12t29 + com_high$agePct16t24 +     com_high$
    numbUrban + com_high$pctWWage + com_high$pctWPubAsst +     com_high$perCapInc + com
    _high$blackPerCap + com_high$AsianPerCap +     com_high$PctNotHSGrad + com_high$
    PctBSorMore + com_high$PctUnemployed +     com_high$PctOccupManu + com_high$
    MalePctDivorce + com_high$FemalePctDiv +     com_high$TotalPctDiv + com_high$
    PctKids2Par + com_high$PctYoungKids2Par +     com_high$PctImmigRec5 + com_high$
    PctImmigRec8 + com_high$PctRecImmig5 +     com_high$PctRecImmig8 + com_high$
    PersPerOccupHous + com_high$PersPerRentOccHous +     com_high$PctPersOwnOccup + com
    _high$PctPersDenseHous + com_high$PctHousOwnOcc +     com_high$MedYrHousBuilt + com
    _high$PctWOFullPlumb + com_high$RentMedian +     com_high$MedRent + com_high$
    MedRentPctHousInc + com_high$MedOwnCostPctIncNoMtg +     com_high$NumInShelters +
    com_high$LandArea + com_high$LemasPctOfficDrugUn)
# AIC = -1122.3 also terrible

summary(com_high.for.reg)
summary(com_high.bac.reg)

# R^2 best backwards = 0.371
# R^2 for both vastly different -- and not even good, highest is .5276
par(mfrow=c(2,2))
plot(com_high.bac.reg)


#####################################
##                                 ##
##    LOGISTIC REGRESSION          ##
##                                 ##
#####################################

log_com = com

# create binary column for High Crime, with cut-off of 0.5
for(i in 1:nrow(log_com))
{
  if(log_com$ViolentCrimesPerPop[i] < 0.5)
    log_com$HighCrime[i] = 0
  else
    log_com$HighCrime[i] = 1
}

# see how our polynomial model from before works with logistic regression
library(Design)
log.reg = glm(HighCrime ~ log_trans_n_fix(PctKids2Par) + log_trans_n_fix(HousVacant) +
     poly(pctUrban, 5) + log_trans_n_fix(PctWorkMom) + poly(NumStreet, 3) + log_trans_n
    _fix(MalePctDivorce) + poly(PctIlleg, 2) + poly(numbUrban, 2) + log_trans_n_fix(
    PctPersDenseHous) + poly(racepctblack, 2) + log_trans_n_fix(MedOwnCostPctIncNoMtg)
```

```r
+ log_trans_n_fix(RentLowQ) + log_trans_n_fix(MedRent), data=log_com)
log.reg.lrm = lrm(HighCrime ~ log_trans_n_fix(PctKids2Par) + log_trans_n_fix(HousVacant) + poly(pctUrban, 5) + log_trans_n_fix(PctWorkMom) + poly(NumStreet, 3) + log_trans_n_fix(MalePctDivorce) + poly(PctIlleg, 2) + poly(numbUrban, 2) + log_trans_n_fix(PctPersDenseHous) + poly(racepctblack, 2) + log_trans_n_fix(MedOwnCostPctIncNoMtg) + log_trans_n_fix(RentLowQ) + log_trans_n_fix(MedRent), data=log_com)
summary(log.reg)
log.reg.lrm
par(mfrow=c(3,3))
plot(log.reg)
glm.diag.plots(log.reg)

# use custom stepwise function to find best AIC combination
glm_step(log_com, 101, c(100), 10)
# returns HighCrime ~ PctIlleg + NumStreet + agePct12t29 + racePctWhite + PctHousOccup + pctUrban + PctKids2Par
step.reg = glm(HighCrime ~ PctIlleg + NumStreet + agePct12t29 + racePctWhite + PctHousOccup + pctUrban + PctKids2Par, data=log_com)
step.reg.lrm = lrm(HighCrime ~ PctIlleg + NumStreet + agePct12t29 + racePctWhite + PctHousOccup + pctUrban + PctKids2Par, data=log_com)
summary(step.reg)
step.reg.lrm
par(mfrow=c(2,2))
plot(step.reg)
glm.diag.plots(step.reg)

# result from stepwise has better AIC by 4 (not a lot), but not as good pseudo-R-squared (.53 vs. .56)
# residual vs. fitted is better slightly, but Q-Q plot is identical and bad in both cases


######################################
##                                  ##
##    TESTING THE LOGISTIC MODEL    ##
##                                  ##
######################################

# loop thru fitted points for model and see if they are correct
compare_to_real(fitted(log.reg), log_com$HighCrime)
compare_to_real(fitted(step.reg), log_com$HighCrime)


############### Testing the logistic model. ##################

# generate 100 distinct random points which refer to observations
random_point_count = 100
random_points = c()
while (random_point_count > 0)
{
  rand = ceiling(runif(1,1,1994))
  if (any(random_points == rand))
    next
  random_points = c(random_points, rand)
  random_point_count = random_point_count − 1
}

# split data into two sets —— one for training model and one for testing
j = 1
k = 1
test.com = data.frame(matrix(nrow = 100, ncol = 101)) # 100 observations for test, 101 variables in log_com model
names(test.com) <−  as.vector(colnames(log_com))
train.com = data.frame(matrix(nrow = 1994 − 100, ncol = 101))
names(train.com) <−  as.vector(colnames(log_com))
for(i in 1:nrow(log_com)) {
  if (any(random_points == i)) {
    test.com[j, ] = log_com[i, ]
    j = j + 1
  } else {
    train.com[k, ] = log_com[i, ]
    k = k + 1
  }
}

glm_step(train.com, 101, c(100), 10)
step.reg = glm(HighCrime ~ PctIlleg + NumStreet + racePctWhite + agePct12t29 + pctUrban + PctHousOccup, data=log_com)
step.reg.lrm = lrm(HighCrime ~ PctIlleg + NumStreet + racePctWhite + agePct12t29 + pctUrban + PctHousOccup, data=log_com)
summary(step.reg)
step.reg.lrm
compare_to_real(round(predict(step.reg,test.com,type="response")), test.com$HighCrime)

######################################
##                                  ##
##    NEURAL NETWORK TESTING        ##
##                                  ##
######################################

library(nnet)
train.com_result = train.com[,101]
train.com_explain = train.com[,c(−100,−101)]
com.nnet = nnet(train.com_explain, train.com_result, size=10,  linout=T, decay=0.01, maxit=1000, MaxNWts=2000)
plot(train.com_result,predict(com.nnet,train.com_explain), main="Neural_Net_Results", xlab="True",ylab="NN_predictions",xlim=c(−0.01,0.01))
test.com_result = test.com[,101]
test.com_explain = test.com[,c(−100,−101)]
test.com_result == c(round(predict(com.nnet,test.com_explain)))
compare_to_real(test.com_result,round(predict(com.nnet,test.com_explain))) # 86% are correct

library(pmml)
pmml(com.nnet, model.name="Violent_Crimes_Per_Population_Model", app.name="Com/PMML")
```

R code