Communities and Crime Data – Linear Modelling

by 1710770, 1826598, 1831807, …

1 Findings

We are going to explain how we developed our suggested models, go through their limitations and give you our interpretation of the data. In Section 2, you can find an in depth analysis of why we made each of our modelling decisions.

1.1 Data Cleaning

Based on our findings from the exploratory data analysis, we made a number of adjustments to the original data set. This included: removing the 73 missing values in both ***medIncome*** and ***pctEmploy***; Merging Pacific into West within the ***region*** variable; Changing ***pctUrban*** into a factor variable, dividing it into two intervals <50% and >50%.

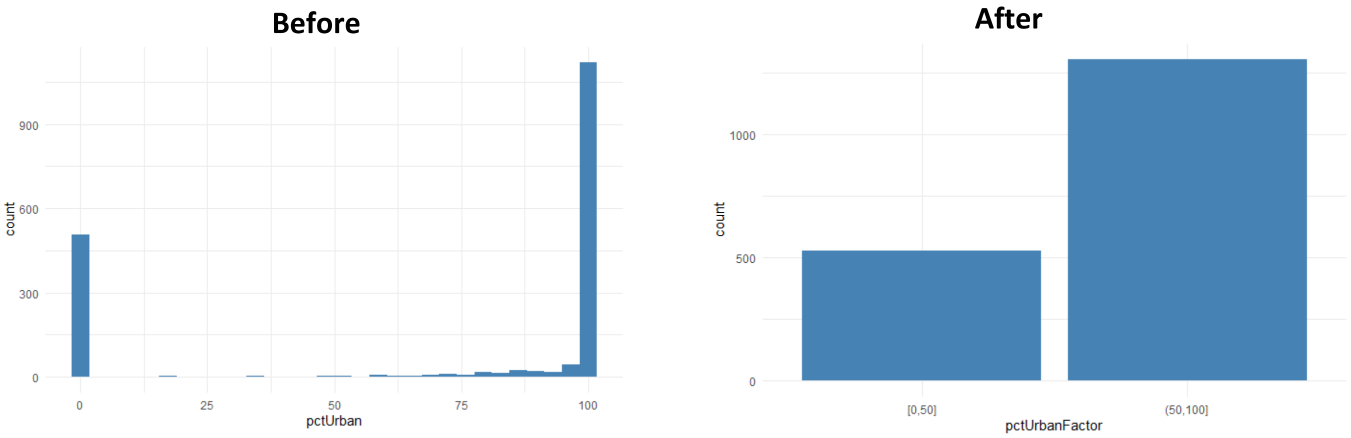


Figure 1: Transformation of pctUrban

By observing scatter plots of the predictor variables against ***violentPerPop*** and ***nonViolPerPop*** we identified that the variable ***pctVacant6up*** had no relationship with either, hence we made the decision not to consider this variable in the rest of the analysis.

1.2 Transformations

When building a model there are certain assumptions that must be fulfilled. We had to ensure the predictor variables had a linear relationship with the outcome variables (i.e. if one of them increases by a unit the other increases by a multiple of that unit). We also had to ensure the relationship was homoscedastic (constant variance in the errors).

We ended up transforming both of the outcome variables using a $\log_2$ transformation and majority of the predictor variables with either log or power transformations. This ensured that both linearity and homoscedasticity assumptions were fulfilled so that we could build a linear model. The transformations for the variables can be seen in figure 10.

1.3 Influential observations

An observation having high leverage means it is made up of an unusual combination of predictor variables, if these observations are also large outliers they are known as influence points. Observations with high influence have a significant effect on the regression analysis. This can be problematic as if these observations are not valid, this could negatively affect our model.

Group	Influence Points	Reason for Removal
Alaska	816, 1128, 1758	Specifically, observation 1128 shows high influence. Also, we decided that Alaska isn't representative of the rest of the US.
Zero values for ownHousQrange	40, 493, 520, 565, 1737	Specifically, observations 40 and 1737 are both outliers and have high leverage. Also, the zero value for ownHousQrange seems very unlikely and could be an error.
High influence points	322, 1329	Both observations have very high influence, so we removed these to avoid them significantly affecting the data. (For 1329 this is only the case for nonViolPerPop hence only removed for that model).

Figure 2: Removed Influence points

1.4 Multicollinearity

We identified 8 groups of variables with collinearity between them, this means these variables can to some degree predict the other variables in their group.

		Collinear Variables
Least Significant	**(1)**	***pctKids2Par pctKidsBornNevrMarr***
	(2)	***pctCollGrad pctKids2Par pctKidsBornNevrMarr***
	(3)	***ownHousMed rentMed***
	(4)	***pctWdiv pctCollGrad popDensity***
	(5)	***pctLowEdu pctNotHSgrad popDensity***
	(6)	***medIncome pctEmploy pctHousOwnerOccup***
	(7)	***pctLowEdu pctNotHSgrad pctCollGrad pctWdiv***
Most Significant	**(8)**	***ownHousMed ownHousQrange***

Figure 3: Groups of multicollinear variables

As you can see some of these groups are also fairly intuitive and can broadly be categorised into topics, for example group 3 relates to house prices.

Multicollinearity can be an issue when building a model, as it makes it hard to interpret and understand the effect of your variables, hence we proceed with variable selection.

1.5 Variable Selection

In order to find the best model which balanced both predictive and explanatory power we tested multiple variable selection techniques. The final models we decided on are below as well as a table to provide a better understanding of these models and to account for the transformations performed:

Violent Model: Chosen using the BIC stepwise regression variable selection method

$$\begin{aligned}\log_2(\mathbf{violentPerPop}) = 8.9 &- (2.0\times10^{-1})\mathbf{NorthEast} + (4.6\times10^{-1})\mathbf{West} + (3.7\times10^{-1})\mathbf{South}\\ &+ (3.8\times10^{-1})\mathbf{pctUrban(50,100]} + (6.6\times10^{-3})\mathbf{medIncome}^{0.5} - (1.8\times10^{-2})\mathbf{pctWdiv}\\ &- (4.6\times10^{-6})\mathbf{pctKids2Par}^{3} + (3.4\times10^{-1})\mathbf{pctKidsBornNevrMarr}^{0.5}\\ &- (8.0\times10^{-21})\mathbf{pctHousOccup}^{10} + (8.9\times10^{-2})\mathbf{pctVacantBoarded}^{0.5}\\ &+ (1.2\times10^{-1})\mathbf{pctForeignBorn}^{0.5}\end{aligned}$$

Note: To find the effective percentage change in the outcome variables from a 1% increase in the predictor, we used the median value of each of the predictor variables in order to gain a rough estimate. This would change as the value of the predictor variable changes.

Selected Variables	Violent Crime - Model Interpretation
Intercept	If all other variables are zero, then violentPerPop is 478
NorthEast	If the region is NorthEast, violentPerPop decreases by ≈ 13%
West	If the region is West, violentPerPop increases by ≈ 38%
South	If region is South, violentPerPop increases by ≈ 29%
pctUrban(50,100]	If pctUrban is between (50,100], violentPerPop increases by ≈ 30%
medIncome	For every 1% increase in medIncome, violentPerPop increases by ≈ 0.4%
pctWdiv	For every 1% increase in pctWdiv, violentPerPop decreases by ≈ 0.5%
pctKids2Par	For every 1% increase in pctKids2Par, violentPerPop decreases by ≈ 3.5%
pctKidsBornNevrMarr	For every 1% increase in pctKidsBornNevrMarr, violentPerPop increases by ≈ 0.2%
pctHousOccup	For every 1% increase in pctHousOccup, violentPerPop decreases by ≈ 3.1%
pctVacantBoarded	For every 1% increase in pctVacantBoarded, violentPerPop increases by ≈ 0.04%
pctForeignBorn	For every 1% increase in pctForeignBorn, violentPerPop increases by ≈ 0.09%

Figure 4: Interpretation of Violent Model

Non-violent Model: Chosen using the LASSO variable selection method

$$\log_2(\mathbf{nonViolPerPop}) = 13.1 - (9.5 \times 10^{-2})\mathbf{MidWest} - (4.3 \times 10^{-1})\mathbf{NorthEast} + (8.8 \times 10^{-2})\mathbf{West} + (4.9 \times 10^{-2})\mathbf{South} + (1.6 \times 10^{-1})\mathbf{pctUrban(50, 100]} - (2.5 \times 10^{-6})\mathbf{pctKids2Par}^3 - (1.2 \times 10^{-21})\mathbf{pctHousOccup}^{10} - (3.2 \times 10^{-6})\mathbf{pctHousOwnerOccup}^2$$

Selected Variables	Non-violent Crime - Model Interpretation
Intercept	If all other variables are zero, then nonViolPerPop is 8192
MidWest	If the region is MidWest, nonViolPerPop decreases by ≈ 6.4%
NorthEast	If the region is NorthEast, nonViolPerPop decreases by ≈ 26%
West	If the region is West, nonViolPerPop increases by ≈ 6.3%
South	If region is South, nonViolPerPop increases by ≈ 3.5%
pctUrban(50,100]	If pctUrban is between (50,100], nonViolPerPop increases by ≈ 12%
pctKids2Par	For every 1% increase in pctKids2Par, nonViolPerPop decreases by ≈ 2%
pctHousOccup	For every 1% increase in pctHousOccup, nonViolPerPop decreases by ≈ 0.5%
pctHousOwnerOccup	For every 1% increase in pctHousOwnerOccup, nonViolPerPop decreases by ≈ 0.02%

Figure 5: Interpretation of Non-Violent Model

1.6 Limitations

Before we get on to answering some of the questions posed, we believe it is first of all important to discuss the limitations of our models and the data. The first thing we would like to elaborate on is that this model is not a "black-box" it is informative about which areas have high and low ***violentPerPop*** and ***nonViolPerPop*** however it won`t necessarily tell you the exact crime rate in each community, rather, a prediction. This issue is particularly salient when we look at areas of the US that don`t conform to the general pattern in section 1.9.

Along with this, there are predictor variables which are highly correlated with variables that have been kept in the final model. These have not been included in our last model due to their redundancy. We let the BIC and LASSO algorithms decide which variables should stay in the final model. However, LASSO chooses variables based on its performance in the particular data sample, and so if two variables are highly correlated, then it may drop one of them arbitrarily leaving out a variable which in a different dataset would prove to be more important. On the other hand, the BIC method of variable selection chooses one "True" candidate model out of many candidates. The issue with BIC is that it assumes that reality is instantiated in one of the models.

Moving onto the data, an issue is the imbalanced distribution of state observations. There were a lot of states which had little to no observations. This limitation led to an inability to make reliable inferences about the States, which based on intuition could have been very useful information. Also a dataset much larger to train our model on would have been useful. Considering the size of the US a dataset with 1902 observations is not particularly large.

1.7 What are the major determinants of high crime rates?

Based on the ***violentPerPop*** model the major determinants of violent crime rates in the US are, ***region***, ***pctUrban***, ***medIncome***, ***pctWdiv***, ***pctKids2Par***, ***pctKidsBornNevrMarr***, ***pctHousOccup***, ***pctVacantBoarded*** and ***pctForeignBorn***.

However, it is very important to note that there is a large amount of multi-collinearity in this dataset. Just because some variables are not included in the model it does not necessarily mean they are not important. For example, ***pctWdiv*** is in the model but ***pctLowEdu***, ***pctNotHSgrad***, ***pctCollGrad*** are not. However, there is a high degree of collinearity between these variables so they could also be important determinants of high crime rates. Along with this, as mentioned in the limitations maybe in a different set of data one of the other colinear variables would be the one included in the model. Hence it is important to refer to figure 3.

Figure 4 illustrates the effect changes in these predictor variables have on ***violentPerPop***. These should only be used as rough guides and when comparing variables with similar levels of effect shouldn`t be used to rank importance as the variables do not have the same underlying distribution which can cause issues.

However we can make some inferences, for example geographical location is an important indicator of ***violentPerPop***, “West” tends to have the highest violent crime rates out of the regions. Similarly urban areas tend to have higher violent crime rates than non-urban areas. Beyond that the strongest determinants of high violent crime are ***pctKids2Par*** and ***pctHousOccup***, with the higher they are, the lower the crime rates.

Based on the ***nonViolPerPop*** model the major determinants of high non-violent crime rates in the US are, ***region***, ***pctUrban***, ***pctKids2Par***, ***pctHousOccup***, ***pctHousOwnerOccup***. As before collinearity needs to be taken into account.

Looking at figure 5 there are some inferences to be made. “West” is the ***region*** that tends to have the highest non-violent crime rates. Also areas that are more Urban tend to have higher non-violent crime rates. Beyond this ***pctKids2Par*** is the strongest indicator of non-violent crime rates, with crime rates decreasing as it increases.

1.8 Are the causes of violent and non-violent crime similar, or are there important differences?

Overall, the causes of violent and non-violent crime are very similar. The key indicators are ***region***, ***pctUrban***, ***pctKids2Par*** and ***pctHousOccup***. This is to be expected as ***violentPerPop*** and ***nonViolPerPop*** are highly correlated.

There are however some important differences. First of all, ***pctUrban*** is a stronger indicator of ***violentPerPop*** than it is of ***nonViolPerPop***. There are also some differences in the indicators even when accounting for multicollinearity. The variables ***pctWdiv***, ***pctVacantBoarded*** and ***pctForeignBorn*** are indicators of ***violentPerPop*** but none of them or variables they are colinear to are indicators of ***nonViolPerPop***.

1.9 Are there areas of the US with unusually low or high crime rates that do not conform to the general pattern?

Looking at the upper and lower 2.5% of the data for both ***violentPerPop*** and ***nonViolPerPop*** that Florida stood out as an area with unusually high violent and especially non-violent crime. It accounted for 1/5 of the high violent crime and almost 2/5 of the high non-violent crime representing 11% and 20% of Florida’s data points respectively. On the other hand, Massachusetts stood out as an area with unusually low crime. It accounted for roughly 1/5 of the low violent and non-violent crime representing 7% and 10% of its data points respectively.

There are also some extreme outliers in this dataset. For example, observations 1011, 988, 1794 and 968 (in states RI, OH, IN and ME respectively) appear to be outliers that our model does not predict too well due to their low violentPerPop values. Observation 776 from California is an outlier for both violentPerPop and nonViolPerPop due to its high crime rates so that is an area to note. However, with these observations being spread across different states there doesn’t appear to be any large areas posing a significant difference in crime rates other than those mentioned above.

2 Statistical Methodology

Full information about the packages used, as well as a comprehensive set of code used in this section, can be found in the appendix.

2.1 Data Cleaning

Before beginning the model fitting process we implemented some data cleaning. We found 21 NA values in ***pctEmploy*** and 52 NA values in ***medIncome***. Boxplots of the data with and without the missing values showed no pattern so that data was Missing Completely at Random and being just 0.038% the data, we decided to remove the rows containing these values. In ***region*** we merged the category "Pacific" into "West" as the density of Pacific is extremely low (just 3 observations) compared to the other regions. Finally, as ***pctUrban*** was mainly bimodal we factorized it into 2 categories, [0,50] and (50,100] renaming this ***pctUrbanFactor***. Finally, we noticed that the row labels did not match up with the data, so we renamed the rows as that they correspond to the rows in the data frame.

2.2 Multiple Linear Regression

We followed the flow chart found in [1]. We have also put this in the references as that our methodology is easy to follow – headings in this section will correspond to the flow chart.

2.2.1 Fit a model based on subject matter expertise and/or observation of the scatter plots

By inspecting the scatter plots it was clear that the variable ***pctVacant6up*** had no relationship with either of the outcome variables. However, for some variables the decision was not as clear, we decided to investigate further by fitting a model of these variables against the outcome variables and seeing if the relationship was significant.

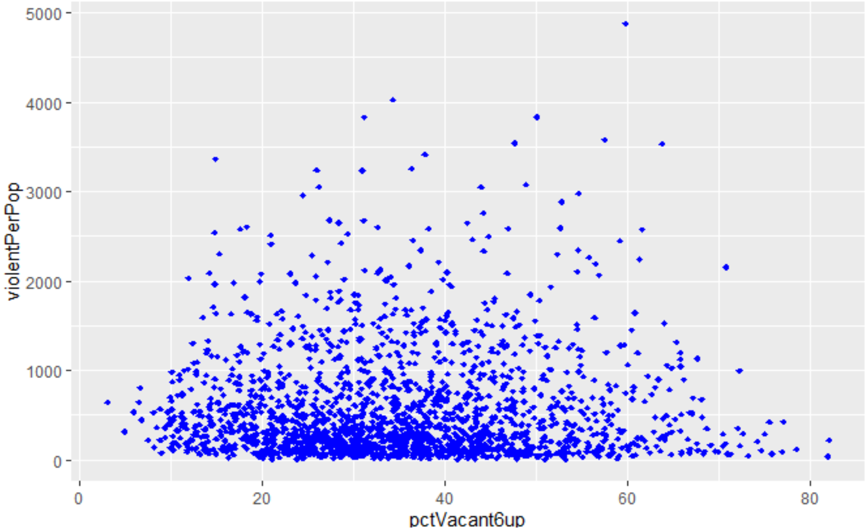


Figure 6: Scatterplot of pctVacant6up

For ***violentPerPop*** we investigated ***pctUrbanFactor*** and found there was a significant relationship so it should be put in the model. For ***nonViolPerPop*** we investigated ***pctVacantBoared***, ***pctForeignborn*** and ***pctUrbanFactor***, we found the first two had a significant relationship however ***pctUrbanFactor*** did not. Here we made a decision to deviate slightly from the flow chart and decided to put ***pctUrbanFactor*** in the model for ***nonViolPerPop***. We did this as sometimes variable selection will choose to keep a categorical variable as it is useful, even if it appears there is no relationship in prior testing due to correlations.

We fitted two models and at this point both the ***violentPerPop*** and ***nonViolPerPop*** model contained all the explanatory variables other than ***pctVacant6up***.

2.2.2 Assess the Adequacy of the model

Following fitting the models, we immediately spotted an issue, the parameter estimates for some of the categories of ***State*** were NA. Calling the **allias()** function on the models it was clear ***State*** and ***region*** were linearly dependent variables and so were perfectly colinear. This can also be seen in figure 7.

One of the variables needed to be removed, we decided to remove ***State*** as:

1) ***State*** has a low number of observations in lots of categories versus ***region*** that has lots of observations per category, so inferences made about regions are more likely to be accurate than those made about states.
2) There are less categories in ***region*** than in ***State*** so ***region*** is easier to use and explain in a model.

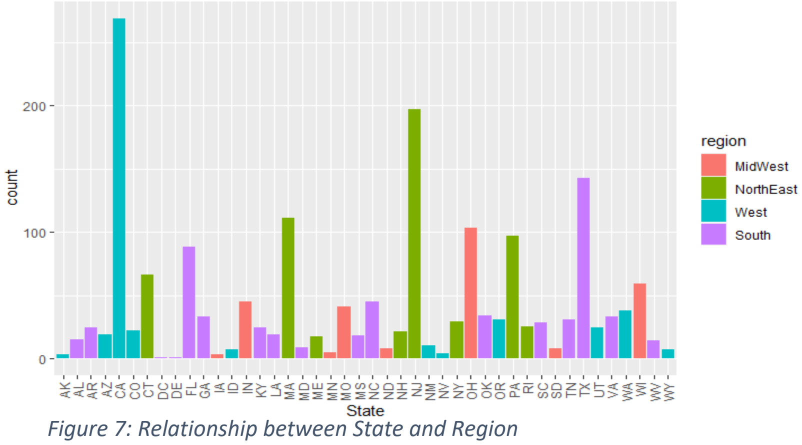


Figure 7: Relationship between State and Region

In order to assess the adequacy of the models we produced the plots seen in figure []. At this stage the residual vs fitted value plots, scale-location plots and partial residual plots were of most interest. It was clear that the variance of the errors were non constant and the predictor variables were not linear with the outcome variables. These are two of the key assumptions required when building a linear model so this needed to be fixed. Looking closer at each individual predictor variable modelled against the outcome variables we also noticed that the majority of the variables were not linear and had a lot of heteroscedasticity.

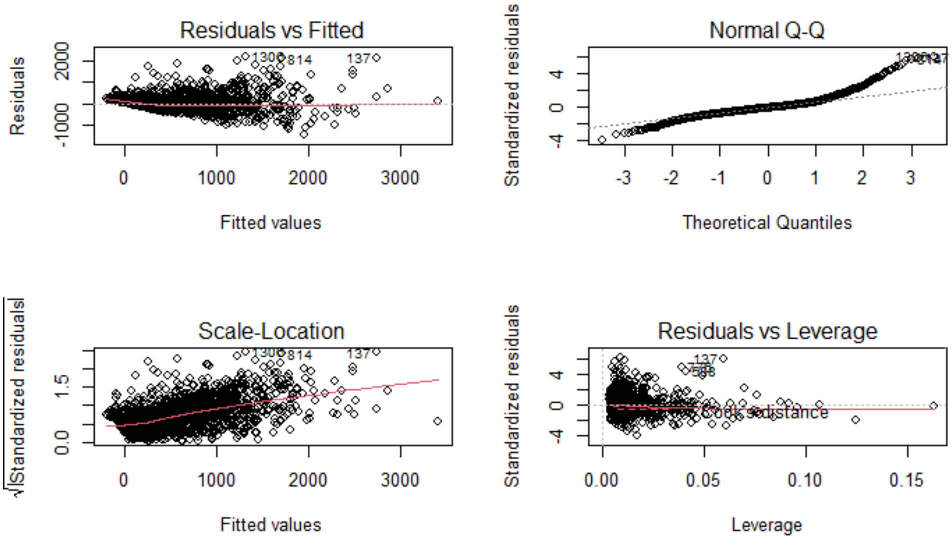


Figure 8: Violent residual plots before transformation

2.2.3 Add new terms to the model and/or transform x and/or Y

The heteroscedasticity in most of the variables was quickly fixed by applying a $\log_2$ transformation to both the outcome variables. We picked a $\log_2$ transformation over a natural log transformations since this makes it easier to explain and understand the relationship with the predictor variables when it comes to interpreting the final model. However, following these transformations the problem of linearity still existed.

We looked through graphs of the outcome variables against each of the predictor variables. We used a combination of techniques such as Box-Cox and trial and error (using intuition from Mosteley's and Tukey's bulging rule) to find the best transformations for each variable in order to satisfy linearity. We then looked back at the now updated full linear models to confirm that linearity and homoscedasticity is observed.

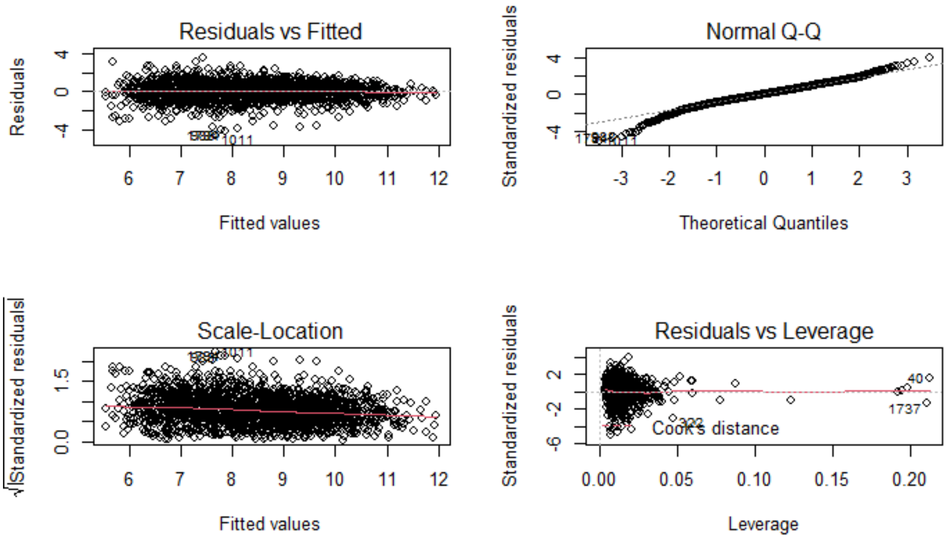


Figure 9: Violent Residual plots after transformation

Variables	Transformations	Variables	Transformations
medIncome	f(x) = $x^{0.5}$	pctHousOwnerOccup	f(x) = $x^{0.5}$
pctWdiv	No transformation	pctVacantBoarded	f(x) = $x^{0.5}$
pctLowEdu	f(x) = $\log_2(x)$	ownHousMed	f(x) = $x^{-0.5}$
pctNotHSgrad	f(x) = $x^{0.5}$	ownHousQrange	f(x) = log(x + 0.1)
pctCollGrad	f(x) = $x^{0.5}$	rentMed	No transformation
pctUnemploy	f(x) = $\log_2(x)$	rentQrange	No transformation
pctEmploy	f(x) = x^2	popDensity	f(x) = $\log_2(x)$
pctKids2Par	f(x) = x^3	pctForeignBorn	f(x) = $x^{0.5}$
pctKidsBornNevrMarr	f(x) = $x^{0.5}$	violentPerPop	f(x) = $\log_2(x)$
pctHousOccup	f(x) = x^{10}	nonViolPerPop	f(x) = $\log_2(x)$

Figure 10: Transformations for all the variables

2.2.4 Do outliers and/or leverage points exist?

We began our analysis of the outliers and leverage points by using the **outlierTest()** function, this alerted our attention to the extreme outliers in the data set. However, at this stage we weren`t just interested in outliers, but rather “bad” outliers, outliers that also have high leverage and so high influence. Observations that have high influence could affect our model.

In order to locate high influence observations we used the **ols_plot_dfbetas()** function which measures the parameter estimate of the explanatory variables with and without each observation. If an observation has a high DFBETA value it has high influence. We didn`t choose a cut-off level of difference between parameter values but rather picked out points that made an especially large difference compared to other observations. For ***violentPerPop*** these were observations, 1236, 367, 1341, 968, 40, 1737, 776, 322 and for ***nonViolPerPop*** these were, 40, 322, 1128, 1261, 1235, 1565, 367, 1737, 482, 1329 and 430.

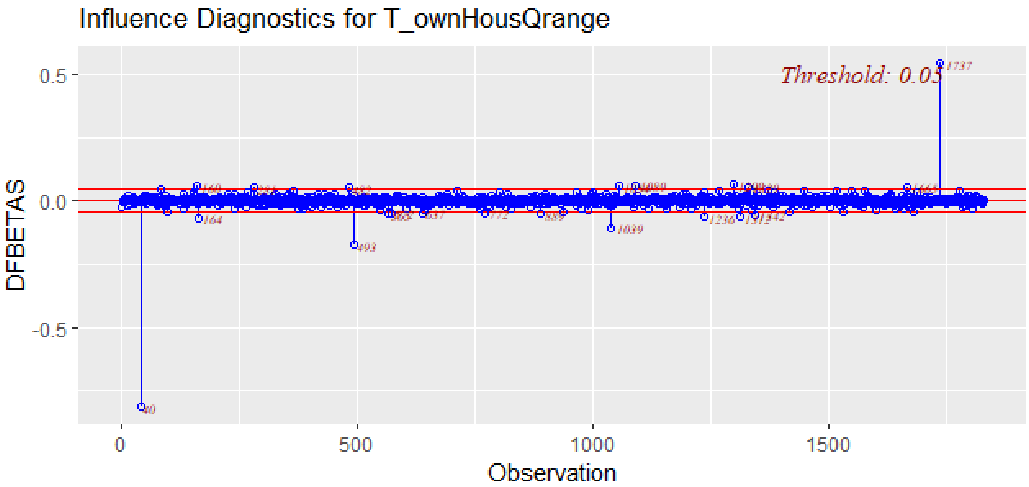


Figure 11: DFBETA plot for ownHousQrange in the Violent model

Following this we used the **ols_plot_resid_lev()** and **avPlots()** functions, the first allowed us to work out if these high influence points were large outliers or had high leverage and the latter allowed us to identify any clusters of influential points.

2.2.5 Are the outliers and leverage points valid?

When investigating observations 40 and 1737, we noticed they had values that seemed nonsensical. In particular ***rentQrange*** and ***ownHousQrange*** had minimum values of 0. Along with this, observations 493, 520 and 565 also had these 0 values as well as high leverage as can be seen on the right hand side of figure 12. We decided to remove this group of observations as they affect the models and seem like errors.

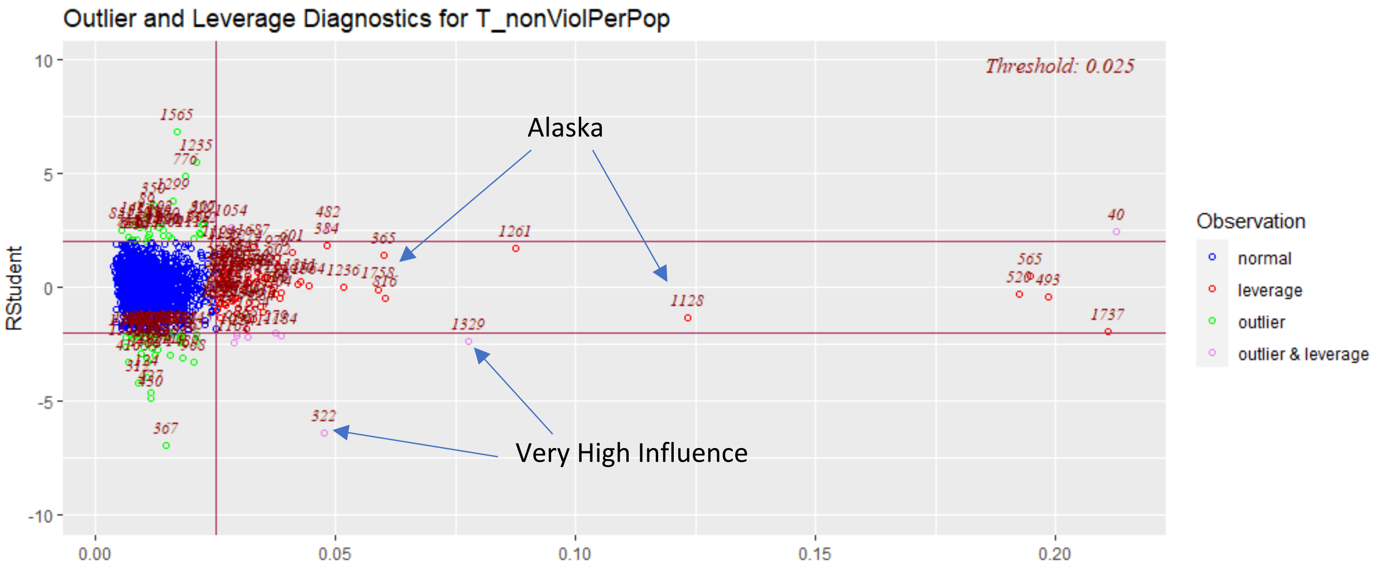


Figure 12: Plot showing high influence points for Non-Violent

Looking at observation 1128, we realised that it was from Alaska and looking at figure [] there were also two other observations, 816 and 1758, from Alaska that despite not being high influence did have high leverage. Based on our intuition Alaska is a very different place than the rest of the United States and any trends seen there wouldn`t necessarily inform us about the rest of the US population. This combined with the fact that there were only 3 observations and any inferences we made about Alaska could very well be incorrect we decided to remove these observations from the models.

Most of the other observations we identified as having high influence were valid points and with their influence not being especially large there was no convincing argument to simply remove these observations as they could hold useful information. There were however two observation 322 and 1329 that despite seeming to be valid points had especially large influence. Observation 322 had very high influence on both models and 1329 had a very high influence only on the ***nonViolPerPop*** model hence we made the decision to remove these observations from the respective models.

2.2.6 Is the Sample Size large?

Following removal of the NA values there are 1829 observations which is a large sample size. Normality is less important for large data sets and so we don`t need to address the lack of normality in the variables or the errors.

2.2.7 Is there a significant relationship association between Y and any of the x`s?

Looking at the F-statistic it was clear there was a significant relationship between ***violentPerPop*** and ***nonViolPerPop*** and the predictor variables. So we established there was a model nested within these variables, the task now was to find the optimum model that is both predictive but also explanatory.

2.2.8 Is there a great deal of redundancy in the full model?

Looking at the model output it was clear that there was redundancy in the model. A number of variables that we expected to have positive relationships with the outcome variables (through initial scatter plot analysis) produced negative parameter estimates in these full models and vice versa. For example, ***pctUnemploy*** had a negative coefficient in both models, but the scatter plots suggest that it has a positive relationship with the outcome variables. This suggested there was some multicollinearity.

Wanting to investigate further we looked at the output of the **vif()** function, this indicates how easily a predictor variable it is predicted using the other predictors, and so the degree of multicollinearity. With variables ***pctWdiv***, ***medIncome, pctLowEdu***, ***pctNotHSgrad***, ***pctKids2Par***, ***ownHousMed***, ***rentMed*** all having VIF scores greater than 10, this indicated that there was lots of multicollinearity in our model.

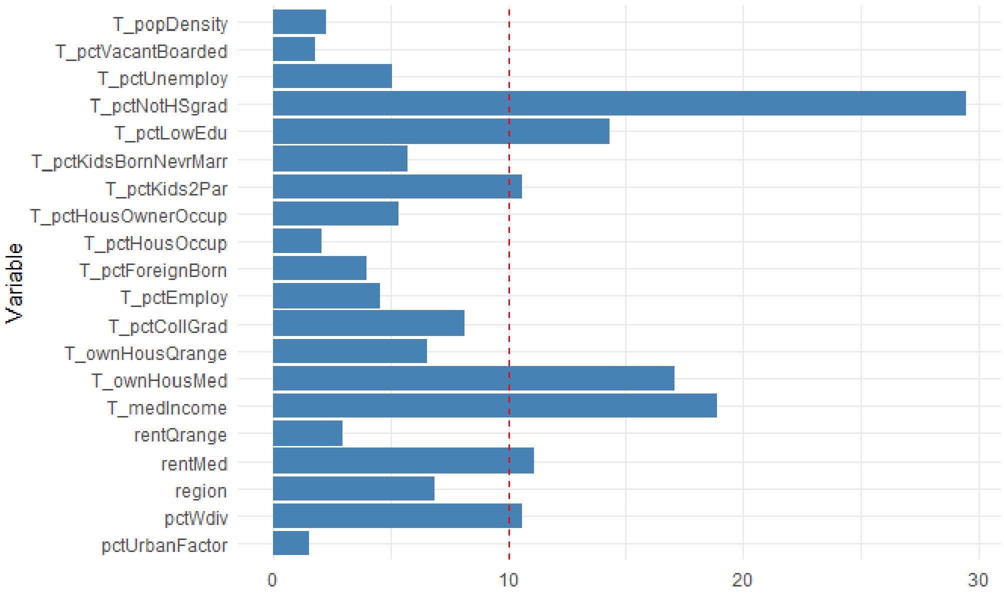


Figure 13: Illustration of VIF for Violent

In order to identify which groups of variables had high collinearity between them we called the **ols_eigen_cindex()** function on the two models. We decided to investigate rows with condition indices of 30 or greater and used a cut-off of greater than 0.2 at which point we considered variables to be colinear. The groups we identified can be seen in the findings in section 1.4.

2.2.9 Use Variable Selection to obtain a final model

As there was multicollinearity and hence redundancy in our model, we went on to perform variable selection in order to decide on a final model.

The difficulty in selecting a model was balancing the predictive and explanatory components. To attempt to come to a good decision we decided to try AIC, BIC, Ridge Regression and LASSO, before comparing the predictive and explanatory power of the resulting models.

In order to measure the predictive power of the models used two different measures [3]:

- Error measure in the estimation period: RMSE (Root Mean Squared Error)
- Error measure in the validation period: R^2 calculated via Cross-Validation

So after running each algorithm we calculated the R^2 via Cross-Validation and the RMSE of the models.

We used the hybrid versions of AIC and BIC and ran each of them 4 times from different starting points to ensure they came to the same result. Ridge Regression ran as expected. With LASSO we encountered some problems, it removed some categories of ***region*** but not others.

The default LASSO algorithm does not group the categories of factor variables, i.e. it may remove some categories of a factor variable but leave others, however we want either all or none of the categories when dealing with factor variables. A possible solution here was to use a package with the Group LASSO, however we found the inbuilt cross validation capabilities in the Non-Grouped LASSO very useful so opted not to do this. Instead we used LASSO on our data set when ***region*** was forced to be in the model (using a penalty factor) and compared this to running LASSO with region removed from the model. As you can see from figure 14, the model with ***region*** had marginally better predictive power with both better R^2 and RMSE. Hence, we used the model with ***region*** included when comparing the LASSO models to the other models.

	Region			No Region		
Variable	p	R^2	RMSE	p	R^2	RMSE
violentPerPop	12	0.6501593	0.9313836	10	0.6331284	0.9513222
nonViolPerPop	9	0.5311418	0.5692164	9	0.5080735	0.5836575

Figure 14: Table of LASSO Region and No Region

As you can see from figure 15 there was very little difference in terms of the models predictive power. We made the decision that trading off model complexity and explainability for a small percentage improvement in RMSE and R^2 was not worth it.

	violentPerPop			nonViolPerPop		
Method	p	R^2	RMSE	p	R^2	RMSE
OLS	23	0.6624136	0.9003539	23	0.5551260	0.5470958
AIC	15	0.6648898	0.9014250	14	0.5587500	0.5481817
BIC	12	0.6642046	0.9039431	12	0.5539244	0.5518254
Ridge Regression	23	0.6677315	0.9141737	23	0.5599642	0.5578307
LASSO	12	0.6501593	0.9313836	9	0.5311418	0.5692164

Figure 15: Table to compare the different Variable Selection methods

For ***nonViolPerPop*** we decided to go with the 1se LASSO model, this is the LASSO model that was minimum in complexity but no more than 1 standard error from the minimum model. Its R^2 and RMSE compared to the model with the best R^2 , the Ridge Regression model, were just 5.1% worse and 2.0% worse respectively. As well as when compared to the model with the best RMSE, the OLS model, were just 4.3% and 3.9% worse respectively. Whilst being much more explainable with 9 parameters versus 23 parameters in both cases.

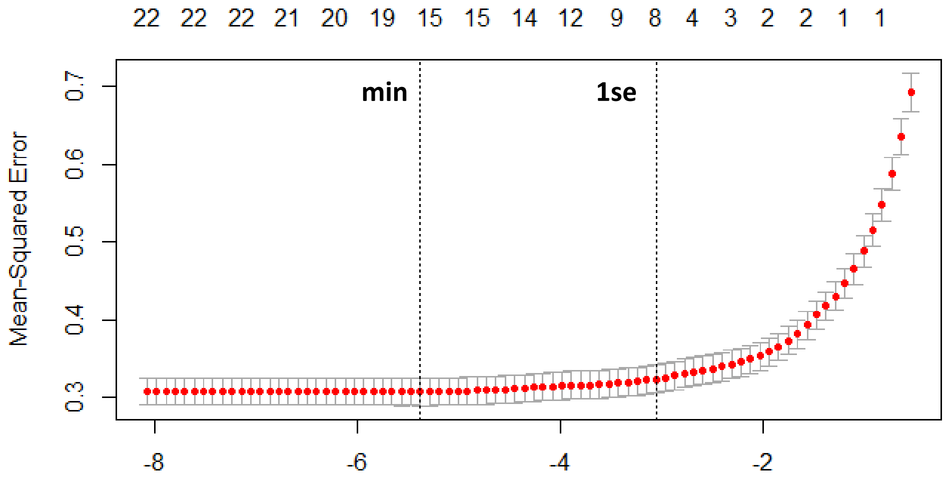


Figure 16: Graph illustrating the different between the 1se and min model for Non-Violent

For ***violentPerPop*** we decided to go with the BIC model as its R^2 and RMSE compared to the model with the best R^2, the Ridge Regression model, were just 0.5% worse and 1.0% better respectively. As well as when compared to the model with the best RMSE, the OLS model, were just 0.3% better and 0.1% worse respectively. Whilst being much more explainable with 12 parameters versus 23 parameters in both cases. We were cautious about choosing a step-wise regression model, as they are in certain situations more prone to finding flawed models than LASSO or Ridge regression. However, in this case p << n and the data set isn`t sparse, along with this we used the hybrid version of BIC so it took the correlation between the variables into account.

3 Authors’ Contributions

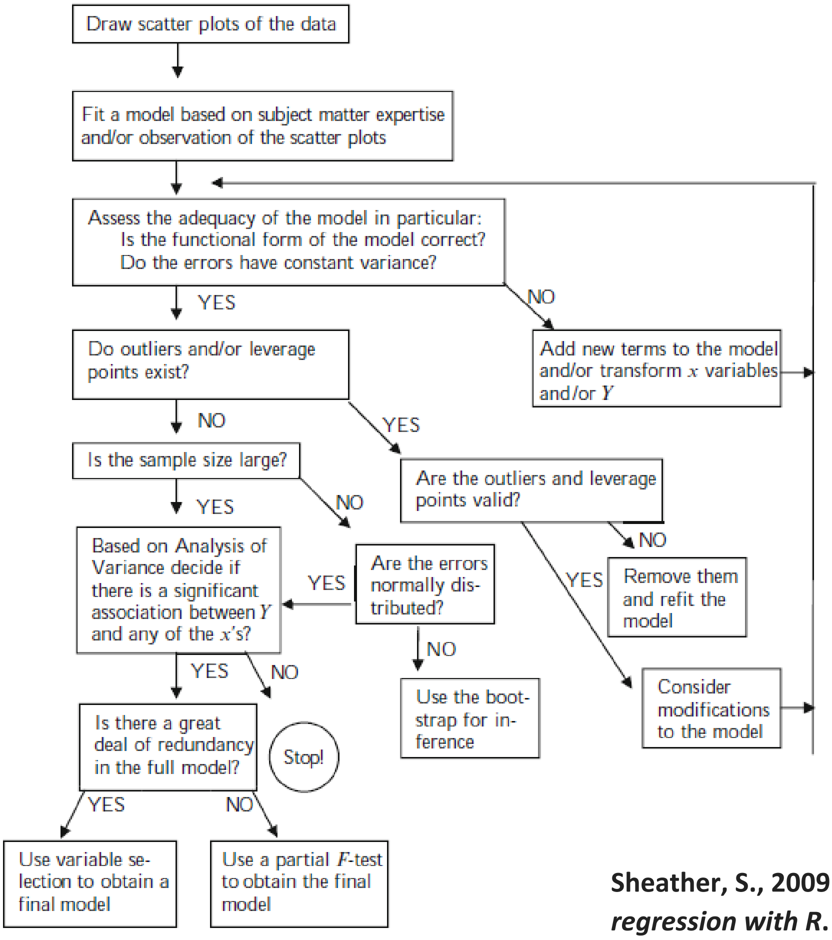
Authors	Weighting	Contribution
Aviv Silver	100%	Transformations; Multicollinearity; Variable Selection
Andre Yiasoumi	100%	Outlier Analysis; Multicollinearity; Variable Selection
Jacob Sushenok	100%	Findings Questions; Outlier Analysis; Multicollinearity
Jun Yi Lee	100%	Data Cleaning; Multicollinearity; Post Model Analysis

4 References

[1] Sheather, S., 2009. *A modern approach to regression with R*. Springer Science & Business Media, pp.252

[2] People.duke.edu. 2021. *How to compare regression models*. [online] Available at: <https://people.duke.edu/~rnau/compare.htm> ,pp.20

[3] People.duke.edu. 2021. *How to compare regression models*. [online] Available at: <https://people.duke.edu/~rnau/compare.htm>



Draw scatter plots of the data
Fit a model based on subject matter expertise and/or observation of the scatter plots
Assess the adequacy of the model in particular: Is the functional form of the model correct? Do the errors have constant variance?
YES
NO
Add new terms to the model and/or transform x variables and/or Y
Do outliers and/or leverage points exist?
NO
YES
Is the sample size large?
Are the outliers and leverage points valid?
YES
NO
NO
Based on Analysis of Variance decide if there is a significant association between Y and any of the x's?
YES
Are the errors normally distributed?
YES
Remove them and refit the model
YES
NO
NO
Use the bootstrap for inference
Consider modifications to the model
Is there a great deal of redundancy in the full model?
Stop!
YES
NO
Use variable selection to obtain a final model
Use a partial F-test to obtain the final model

Sheather, S., 2009. *A modern approach to regression with R*. Springer Science & Business Media, pp.252

5 Appendix

Wherever the code (minus the variable names) is very similar for both ***violentPerPop*** and ***nonViolPerPop*** we have shown only the code for ***violentPerPop*** as to limit the length of the appendix.

5.1 Packages and Dataset

```
install.packages("ggplot2")
library(ggplot2)
install.packages("car")
library(car)
install.packages("olsrr")
library(olsrr)
install.packages("dplyr")
library(dplyr)
install.packages("MASS")
library(MASS)
install.packages("glmnet")
library(glmnet)
load("USACrime.Rda")
```

5.2 Statistical Methodology Sections

Data Cleaning

```
#Renaming Rows and removing NA values
rownames(USACrime) <- 1:nrow(USACrime)
c.USACrime <- na.omit(USACrime)
rownames(c.USACrime) <-
1:nrow(c.USACrime)

#Merging Regions and changing pctUrban
to a Factor Variable
c.USACrime$region <-
factor(c.USACrime$region,
levels=c("MidWest", "NorthEast",
"Pacific", "South", "West"),
labels=c("MidWest", "NorthEast", "West",
"South", "West"))

c.USACrime$pctUrbanFactor <-
as.factor(cut_interval(c.USACrime$pctUrb
an,n=2))
```

Fit a model based on Subject matter expertise and/or observation of the scatterplots

```
#Checking if certain variables had a
relationship with the outcome variables
summary(lm(violentPerPop ~
pctUrbanFactor, data = c.USACrime))
summary(lm(nonViolPerPop ~
pctForeignBorn, data = c.USACrime))
summary(lm(nonViolPerPop ~
pctVacantBoarded, data = c.USACrime))

#Defining a dataset for each outcome
variable
viol.c.USACrime <- subset(c.USACrime,
select = -c(3,16,24))
nonviol.c.USACrime <- subset(c.USACrime,
select = -c(16,3,23))
```

Assess the Adequacy of the Model

```
#Fitting a model for each outcome
variable
ViolentModel <- lm(violentPerPop ~ .,
data = viol.c.USACrime)

#Investigating States
alias(ViolentModel)

#Refitting model without States
viol.c.USACrime <- subset(c.USACrime,
select = -c(1,3,16,24))
nonviol.c.USACrime <- subset(c.USACrime,
select = -c(1,16,3,23))
ViolentModel <- lm(violentPerPop ~ .,
data = viol.c.USACrime)

#Assessing model assumptions
par(mfrow = c(2,2))
plot(ViolentModel)
```

Add new terms and/or transform X and/or Y

```
#Boxcox transformation function
boxcox <- function(x, p) {
  if (p == 0) {
    log(x)
  }
  else {
    (x**p - 1)/p
  }
}

#Defining possible transformations
p <- c(-2, -1, -0.5, 0, 0.5, 1, 2, 3,
10)

#Finding the best transformation for
each variable
for (i in 1:9) {
  if (min(x) > 0 | i > 4) {
    par(mfrow = c(1,4))

plot(lm(c.USACrime$transformedviolentPer
Pop ~ boxcox(x, p[i])), main = p[i])
  }
}
```

Do Outliers and/or Leverage points exist?

```
#Identifying extreme outliers
outlierTest(transformed.ViolentModel)

#Locating and investigating influence
points
ols_plot_dfbetas(transformed.ViolentMode
l)
ols_plot_resid_lev(transformed.ViolentMo
del)
avplots(transformed.ViolentModel)

#Identifying Alaska and 0 values
observations
```

```
which(c.USACrime$ownHousQrange == 0 |
c.USACrime$rentQrange == 0)
which(c.USACrime$State == "AK")
```

Are the outliers and leverage points valid?

```
#Removing the points we decided on
out.transformed.viol.c.USACrime <-
transformed.viol.c.USACrime[-c(40, 493,
520, 565, 1737, 1128, 816, 1758, 322),]
out.transformed.nonviol.c.USACrime <-
transformed.nonviol.c.USACrime[-c(40,
493, 520, 565, 1737, 1128, 816, 1758,
322,1329),]

#Fitting a new model with these
observations removed
out.transformed.ViolentModel <-
lm(T_violentPerPop ~ ., data=
out.transformed.viol.c.USACrime)
```

Is there a significant relationship between Y and any of the X`s?

```
#Looked at the F-test result
summary(out.transformed.ViolentModel)
```

Is there a great deal of redundancy in the model?

```
#Early indication of multicollinearity
summary(out.transformed.ViolentModel)

#Identifying variables with collinearity
vif(out.transformed.ViolentModel)

#Identifying groups of variables with
multicollinearity
ols_eigen_cindex(out.transformed.Violent
Model)
```

Use Variable Selection to obtain a final model

```
#LOOCV function, note k=nrow(xviol)
kfoldCV <- function(y, x, K=nrow(xviol))
{

  N <- nrow(x)
  df <- data.frame("y"=y, x)

  ## Create random subsets
  subset <- rep_len(1:K, N)
  subset <- sample(subset)

  yhat <- numeric(N)
  for (k in 1:K) {
    kfold.fit <- lm(y ~ .,
data=df[subset!=k,])
    yhat[subset==k] <-
predict(kfold.fit,
newdata=df[subset==k,])
  }
  list(yhat=yhat, ssr=sum((y - yhat)^2))
}

#MSE Function
mse <- function(Model)
  mean(Model$residuals^2)

#Defining variables for LOOCV of OLS
xviol <-
out.transformed.viol.c.USACrime[, -
c(21)]
yviol <-
out.transformed.viol.c.USACrime$T_violen
tPerPop

#R^2 Violent OLS
kfoldcv.out.viol <- kfoldCV(yviol,
xviol)
kfoldcv.out.viol$ssr
(cor(kfoldcv.out.viol$yhat, yviol))^2

#MSE of Violent OLS
mse(out.transformed.ViolentModel)

#We defined a number of different
starting points. Only 1 has been shown
here.
a.min = lm(T_violentPerPop ~ 1, data =
out.transformed.viol.c.USACrime)

a.test = lm(T_violentPerPop ~ 1 +
T_pctForeignBorn + T_pctHousOccup, data
= out.transformed.viol.c.USACrime)

#Performing AIC for Violent, repeated 3
more times but shown once here
step.both.viol <- step(a.test,
direction="both",

scope=list("lower"=a.min,
"upper"=out.transformed.ViolentModel),
k=2)

#Calculating Violent AIC R^2
xviolAICboth <-
out.transformed.viol.c.USACrime[,
c("T_pctKids2Par", "region",
"T_pctKidsBornNevrMarr",
"T_pctForeignBorn", "T_pctHousOccup",
"T_medIncome", "pctWdiv",
"T_pctVacantBoarded", "pctUrbanFactor",
"rentQrange", "T_pctUnemploy",
"T_pctCollGrad")]
yviolAICboth <-
out.transformed.viol.c.USACrime$T_violen
tPerPop
kfoldcv.out.violAICboth <-
kfoldCV(yviolAICboth, xviolAICboth)
kfoldcv.out.violAICboth$ssr
(cor(kfoldcv.out.violAICboth$yhat,
yviolAICboth))^2

#Calculating Violent AIC MSE
a <- cbind(yviolAICboth,xviolAICboth)
mse(lm(yviolAICboth ~ ., data = a))

#Performing BIC for Violent, repeated 3
more times but shown once here
step.both.viol <- step(a.test,
direction="both",
scope=list("lower"=a.min,
"upper"=out.transformed.ViolentModel),
k=log(nrow(out.transformed.viol.c.USACri
me)))
```

```
#Calculating Violent BIC R^2
xviolBICboth <-
out.transformed.viol.c.USACrime[,
c("region", "T_medIncome", "pctWdiv",
"T_pctKids2Par",
"T_pctKidsBornNevrMarr",
"T_pctHousOccup", "pctUrbanFactor",
"T_pctVacantBoarded",
"T_pctForeignBorn")]
yviolBICboth <-
out.transformed.viol.c.USACrime$T_violen
tPerPop
kfoldcv.out.violBICboth <-
kfoldCV(yviolBICboth, xviolBICboth)
kfoldcv.out.violBICboth$ssr
(cor(kfoldcv.out.violBICboth$yhat,
yviolBICboth))^2

#Calculating Violent BIC MSE
a <- cbind(yviolBICboth,xviolBICboth)
mse(lm(yviolBICboth ~ ., data = a))

#Ridge Regression Violent
lm.ridgeviol <- lm(T_violentPerPop ~ 0 +
.,data=out.transformed.viol.c.USACrime)
xridge.viol <-
model.matrix(lm.ridgeviol)
yridge.viol <-
out.transformed.viol.c.USACrime$T_violen
tPerPop

#Finding the Optimum Lambda value and
defining it
ridgecv.fit.viol <-
cv.glmnet(xridge.viol,yridge.viol,alpha
= 0,nfolds =
nrow(out.transformed.viol.c.USACrime))
opt.lambda.ridgeviol <-
ridgecv.fit.viol$lambda.min
y_predicted_ridgeviol <-
predict(ridge.fit.viol,s=opt.lambda.ridg
eviol,newx=xridge.viol)

#Calculating MSE of Ridge Regression
Violent
pr.ridge =
cv.glmnet(xridge.viol,yridge.viol,type.m
easure='mse', keep=TRUE, alpha=0,nfolds
= nrow(out.transformed.viol.c.USACrime))
lambda.ridge = pr.ridge$lambda.min
mse.2 = pr.ridge$cvm[pr.ridge$lambda ==
lambda.ridge]
mse.2

#Calculating R^2 value of Ridge
Regression Violent and seeing what the
coefficients are
ridge.viol.sst <- sum((yridge.viol-
mean(yridge.viol))^2)
ridge.viol.sse <-
sum((y_predicted_ridgeviol -
yridge.viol)^2)
ridge.violrsq <- 1-
ridge.viol.sse/ridge.viol.sst
ridge.violrsq
coef(ridgecv.fit.viol,opt.lambda.ridgevi
ol)
```

```
#LASSO Violent Default
lm.lassoviol <- lm(T_violentPerPop ~ 0 +
.,data=out.transformed.viol.c.USACrime)
xlassoviol = model.matrix(lm.lassoviol)
ylassoviol =
out.transformed.viol.c.USACrime$T_violen
tPerPop

#Finding the Optimum Lambda value
lassofit.viol2 <-
cv.glmnet(xlassoviol,ylassoviol,
alpha=1, nfolds =
nrow(out.transformed.viol.c.USACrime))
plot(lassofit.viol2)
lassofit.viol2

#Calculating R^2 value of LASSO Violent
Default
lassofit.viol2$index
cvmvalue <-lassofit.viol2$cvm[31]
rsq <- 1 - cvmvalue/var(ylassoviol)
rsq

#Calculating MSE of LASSO Violent
Default
pr.lasso =
cv.glmnet(xlassoviol,ylassoviol,type.mea
sure='mse', keep=TRUE, alpha=1,nfolds =
nrow(out.transformed.viol.c.USACrime))
lambda.lasso = pr.lasso$lambda.1se
mse.2 = pr.lasso$cvm[pr.lasso$lambda ==
lambda.lasso]
mse.2

#LASSO Violent Default Coefficients
coef(lassofit.viol2,lassofit.viol2$lambd
a.1se)

#LASSO Violent Region Forced in Model
lm.lassoviol <- lm(T_violentPerPop ~ 0 +
.,data=out.transformed.viol.c.USACrime)
xlassoviol = model.matrix(lm.lassoviol)
ylassoviol =
out.transformed.viol.c.USACrime$T_violen
tPerPop

#Finding the Optimum Lambda value
lassofit.viol2 <-
cv.glmnet(xlassoviol,ylassoviol,
alpha=1, nfolds =
nrow(out.transformed.viol.c.USACrime),pe
nalty.factor =
c(0,0,0,0,0,rep(1,(ncol(xlassoviol)-
4))))
plot(lassofit.viol2)
lassofit.viol2

#Calculating R^2 value of LASSO Violent
Region Forced in Model
lassofit.viol2$index
cvmvalue <-lassofit.viol2$cvm[27]
rsq <- 1 - cvmvalue/var(ylassoviol)
rsq

#Calculating MSE of LASSO Violent Region
Forced in Model
```

```
pr.lasso =
cv.glmnet(xlassoviol,ylassoviol,type.mea
sure='mse', keep=TRUE, alpha=1,nfolds =
nrow(out.transformed.viol.c.USACrime))
lambda.lasso = pr.lasso$lambda.1se
mse.2 = pr.lasso$cvm[pr.lasso$lambda ==
lambda.lasso]

#LASSO Violent Region Forced in model
Coefficients
coef(lassofit.viol2,lassofit.viol2$lambd
a.1se)

#LASSO Violent Region Removed
lm.lassoviol <- lm(T_violentPerPop ~ 0 +
.,data=out.transformed.viol.c.USACrime[c
(-1)])
xlassoviol = model.matrix(lm.lassoviol)
ylassoviol =
out.transformed.viol.c.USACrime$T_violen
tPerPop

#Finding the Optimum Lambda value
lassofit.viol2 <-
cv.glmnet(xlassoviol,ylassoviol,
alpha=1, nfolds =
nrow(out.transformed.viol.c.USACrime))
plot(lassofit.viol2)
lassofit.viol2

#Calculating R^2 value of LASSO Violent
Region Removed
lassofit.viol2$index
cvmvalue <-lassofit.viol2$cvm[33]
rsq <- 1 - cvmvalue/var(ylassoviol)
rsq

#Calculating MSE of LASSO Violent Region
Removed
pr.lasso =
cv.glmnet(xlassoviol,ylassoviol,type.mea
sure='mse', keep=TRUE, alpha=1,nfolds =
nrow(out.transformed.viol.c.USACrime))
lambda.lasso = pr.lasso$lambda.1se
mse.2 = pr.lasso$cvm[pr.lasso$lambda ==
lambda.lasso]
mse.2

#LASSO Violent Region Removed
Coefficients
coef(lassofit.viol2,lassofit.viol2$lambd
a.1se)
```

5.3 Figures

Any variables not defined in this section are fully defined in section 5.2.

Figure 1

```
ggplot(c.USACrime, aes(x=pctUrban)) +
  geom_bar(stat="bin", fill="steelblue")
+
  theme_minimal()

ggplot(c.USACrime,
aes(x=pctUrbanFactor)) +
  geom_bar(stat="count",
fill="steelblue") +
  theme_minimal()
```

Figure 6

```
ggplot(c.USACrime, aes(x=pctVacant6up,
y=violentPerPop)) +
  geom_point(shape=18, color="blue")
```

Figure 7

```
ggplot(c.USACrime,aes(x=c.USACrime$State
,fill=c.USACrime$region)) + geom_bar() +
  theme(axis.text.x = element_text(angle
= 90, vjust = 0.5, hjust=1)) +
  scale_fill_discrete(name = "region") +
  scale_x_discrete(name = "State")
```

Figure 8

```
plot(ViolentModel)
```

Figure 9

```
plot(transformed.ViolentModel)
```

Figure 11

```
ols_plot_dfbetas(transformed.ViolentMode
l)
```

Figure 12

```
ols_plot_resid_lev(transformed.NonViolMo
del)
```

Figure 13

```
xt <-
data.frame(vif(out.transformed.ViolentMo
del))
ggplot(xt, aes(y = row.names(xt),
x=xt[,1])) +
  geom_bar(stat="identity",
fill="steelblue") +
  theme_minimal() +
  geom_vline(xintercept=10,  col =
"red",lty=2) +
  xlab("VIF") +
  ylab("Variable")
```

Figure 16

```
lm.lassononviol <- lm(T_nonViolPerPop ~
0 +
.,data=out.transformed.nonviol.c.USACrim
e)
xlassononviol =
model.matrix(lm.lassononviol)
ylassononviol =
out.transformed.nonviol.c.USACrime$T_non
ViolPerPop
lassofit.nonviol2 <-
cv.glmnet(xlassononviol,ylassononviol,al
pha=1,nfolds =
nrow(out.transformed.nonviol.c.USACrime)
, penalty.factor =
c(0,0,0,0,0,rep(1,(ncol(xlassononviol)-
4))), standardize = TRUE)
plot(lassofit.nonviol2)
```