Assignment 2 ST404

Zhicong Hu, Oliver Robinson, Liam Briggs, Tom Tanner

March 2021

Contents

Findings

Recommendations from Exploratory Data Analysis

A key part of the model-making process is applying the results of Exploratory Data Analysis to ensure that our data is "clean" (i.e., it contains no faulty or incorrect values) by transforming variables to address their skew/relationship with other variables and potentially exclude variables from the model to make it simpler. The recommended adjustments to the USACrime data are:

- Remove entries in the data set with missing values (in **medIncome** and **pctEmploy**) as well as entries that appear to be missing values, since we found this data to be missing completely at random. Additionally, leave the outliers in the data set before modelling.
- Ignore states that have no data entries and combine the "Pacific" region with the "West" region, this reduces the number of levels for this variable to 4.
- Apply several transformations to the outcome and predictor variables. These are listed in the Statistical Methodology.
- Omit certain variables which represent methods of encoding similar information: we choose **medIncome** for income/wealth, **pctKidsBornNeverMarr** for family, and **pctLowEdu** for education.

2.1

Modelling

After making changes to our data, we go through the process of developing our linear regression models. Of the various penalized likelihood strategies, we choose **LASSO regression** as it is both a penalized likelihood strategy and a form of variable selection, which will reduce model complexity.

After going through the process of LASSO regression, we end up with a model for predicting violent crime with the corresponding variables and coefficients. It is important to note that these values, which are from one run of the LASSO regression, may not be the same as the ones calculated in the Statistical Methodology. This has to do with how LASSO finds the optimal λ, which can calculate a different value every time the process is run. However, these coefficients are similar enough to a typical LASSO regression to draw conclusions from.

2.2

Table 1: Violent crime and non-violent crime model variables and coefficients

Variable	Coefficient	Variable	Coefficient
Intercept	10.6298	Intercept	18.1484
pctUrban	0.1619	pctUrban	0.0612
medIncome	-0.1697	medIncome	-0.3618
pctLowEdu	0.0127	pctLowEdu	·
pctUnemploy	0.0958	pctUnemploy	·
pctKidsBornNevrMarr	0.9568	pctKidsBornNevrMarr	0.3057
pctHousOccup	-1.677777e-06	pctHousOccup	-8.189186e-07
pctHousOwnerOccup	-0.0012	pctHousOwnerOccup	-0.0043
pctVacantBoarded	0.0832	pctVacantBoarded	0.0291
pctVacant6up	-0.0831	pctVacant6up	-0.0672
pctForeignBorn	0.0096	pctForeignBorn	·

2.3

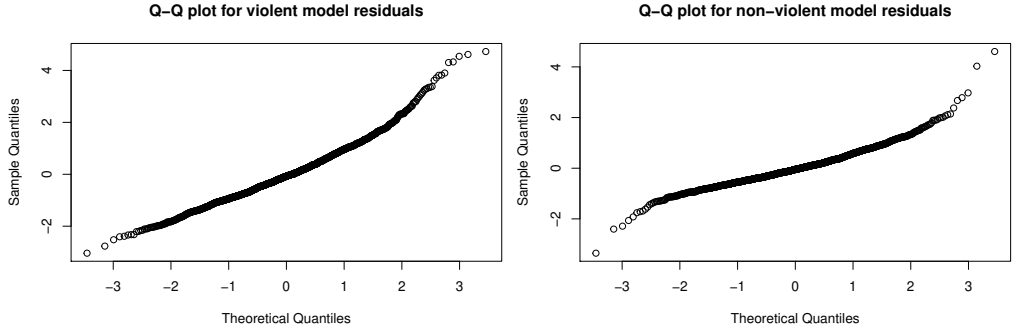
By using this strategy, the variables **pctEmploy** and **popDensity** have their coefficients reduced to zero and are excluded from the model. Therefore, we can conclude that the employment rate and population density are not important factors in predicting levels of violent crime. Conversely, the major takeaway here is that **pctUrban**, **medIncome**, and **pctKidsBornNevrMarr**, with coefficients of 0.1619, -0.1869, and 0.9500 respectively are the most notable determinants for violent crime in a given county. The coefficient of 0.9500 for **pctKidsBornNevrMarr** suggests that a 10% decrease in the percentage of children born to parents who have never married would reduce violent crime by 9.53%. Similarly, the coefficient for **medIncome** means that a 10% increase in Median Income in a county would reduce violent crime by 1.75%.

Looking at our model for non-violent crime, we notice some interesting results. As for the violent crime model, the LASSO process discards **pctEmploy** and **popDensity**. But unlike for violent crime, the non-violent crime model also discards **pctLowEdu**, **pctUnemploy**, and **pctForeign-Born**. This means the model for non-violent crime is simpler than that for violent crime, with fewer variables included. Additionally, the major determinants of non-violent crime, namely **medIncome** and **pctKidsBornNevrMarr**, with coefficients -0.3618 and 0.3057 respectively, are also two of the most important variables used to predict violent crime. From this, we gather that while many of the variables associated with higher violent and non-violent crime are the same, the major difference is that **pctLowEdu**, **pctUnemploy**, and **pctForeignBorn** are deemed to be significantly stronger determinants of violent crime than non-violent crime.

We also note that while **pctKidsBornNevrMarr** is again a significant predictor for non-violent crime, it is less important than the violent crime model, with a 10% decrease in the percentage of children born to parents who have never married reducing non-violent crime by 3.17%, which is much less than the associated decrease of 9.53% for violent crime. However, the effect of an increase in Median Income results in a larger decrease in non-violent crime compared to violent crime. A 10% increase in Median Income results in a 3.4% decrease in non-violent crime, which is approximately double that of the associated decrease for violent crime. It should be noted that being an urban area has a more significant effect on levels of violent crime than non-violent crime. An area being classified as Urban increases violent crime by 11.7%, compared to non-violent crime which increases by just 4.2%. These are important differences between causes of the two types of crime for policymakers to consider.

Model Diagnostics

In model diagnostics, we will check the assumptions of normality of the errors in linear regression. We will plot a Q-Q plot for the residuals of both the models.



From the Q-Q plots, we can deduce that:

- Residuals of the model demonstrate normality, this means that there is a linear relationship between our predictor and outcome variables and linear regression is a valid model to use.
- The extreme values for both models are not following the same linear relationship as the rest of the data, this means that there are some communities/states/regions where our model has trouble predicting. We will further discuss this in the limitations of our model.
- Between the Q-Q plot of the two models, the line in our violent model is a lot steeper than the line in our non-violent model. This means that the residuals in our non-violent model are smaller than the ones in our violent model, this corresponds to the lower Mean-Squared Error in our non-violent model compared with the violent model.

Residual Analysis

Once the model has been created, we can examine how well our model fits the data by creating residual plots. The main findings from doing this are:

- The shape of the plot (i.e. the distribution) demonstrate that both of our models are generally a good fit for the data.
- The non-violent crime model has better predictive power than the violent crime model. We can gather this from the shape of both the Q-Q and residual plots. This is also seen by calculating the Mean-Squared Error, where the lower value for the non-violent crime model also supports these findings.
- Generally speaking, it appears the NorthEast region of the US contains the most counties that have abnormally high and low levels of crime; the South also has some areas with unusually high crime rates. We will look at these in more detail.
- The violent crime model has more areas with abnormally high/low levels compared to the non-violent crime model, these tend to be in the Northeast, Midwest, and South.

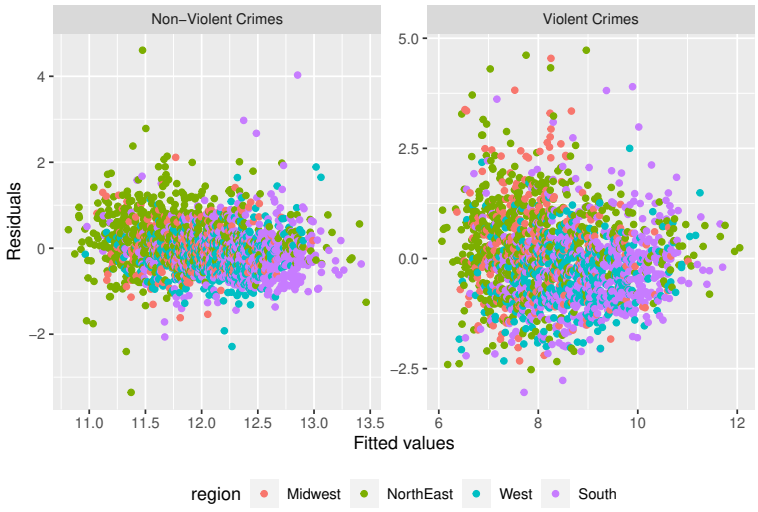


Figure 1: Residual plot for Non-Violent Crime and Violent Crime model

Limitations of our Model

Though we apply a number of techniques to create as robust a model as possible, various trade-offs between effectiveness and simplicity should be noted. These include:

- For the various transformations applied to the data, we often choose ones easier to explain to non-statisticians or ones that may be sub-optimal but not needlessly complex. This could impact larger data values and cause the model to lose some accuracy.
- LASSO, the method we use to select variables, has some drawbacks. For example, the number of variables it selects is limited by the number of data observations, which could cause problems with smaller data sets. This may make our models less reproducible.
- While we did perform analysis to remove as many variables as possible which displayed evidence of collinearity, we cannot guarantee that the variables removed from our model did not impact the crime rates, as well as this, there could be causation instead of correlation.
- Another drawback of LASSO comes from the penalty function. As LASSO adds a penalty function to the model, our model will have a higher Mean-Squared Error compared to a genuine linear model. However, this is an example of a bias-variance trade-off. By increasing the bias, we can prevent overfitting of our model to the training data. This makes sure that our model can work as effectively when predicting future data, which may not be the case without the presence of a penalty function.
- There are regions and states where our non-violent model performs poorly. These are regions that have unusually high or low crime rates which do not conform to the general pattern. From the figure below, we can see that the majority of the communities that the model did not work well for come from the NorthEast region. Looking deeper into individual states, MA which stands for Massachusetts stood out in this regard. Therefore, we need to be more aware when using the non-violent model for prediction purposes in the NorthEast region, especially Massachusetts.

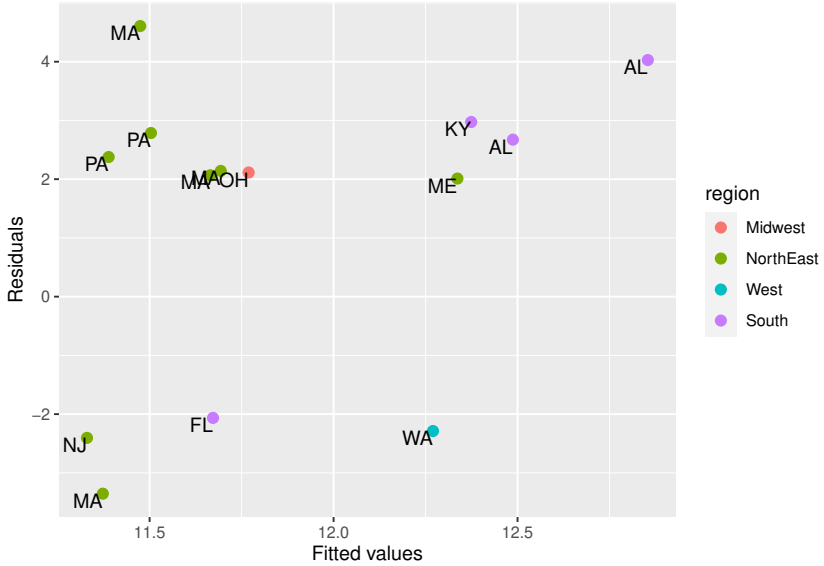


Figure 2: Residual plot of residuals greater than 2 for our Non-Violent Crime model

Statistical Methodology

Data cleaning

Before creating and iterating our final models, we must take the preliminary step of applying the results of the EDA to clean the USACrime data.

Firstly, we omit the observations that contain the 52 and 21 missing values of **medIncome** and **pctEmploy** respectively, as these have been determined to be MCAR (Missing Completely at Random) during the EDA. Furthermore, there are some entries in the variables **ownHousMed**, **rentMed**, **ownHousQrange**, and **rentQrange** which appear to be entered in place of a missing value. We omit these also, leading to a total of 92 data points being removed. In addition, the results of the EDA suggest that the outliers within the data can be left in before we fit the models.

Additionally, the variable **State** has some empty levels and these are dropped at the beginning to clean up the variable. The variable **region** has a level **Pacific** which only contains three observations, which are all from Alaska. To fix this, the **Pacific** level was merged with **West** as this made the most sense geographically. In the end, **region** has four levels: Midwest, Northeast, West and South.

Finally, from the EDA, several variable transformations were recommended to address the skew of both the outcome and predictor variables. These will improve the robustness of the model, with the predictors becoming much more symmetrical after being transformed. They are:

- Log2 transformation: **violentPerPop**, **nonViolPerPop** (outcome variables), **medIncome**, **pctLowEdu**, **pctUnemploy**
- Log2+1 transformation: **pctKidsBornNevrMarr**, **pctVacantBoarded**
- Power transformation: **pctEmploy** (^2), **pctHousOccup** (^3)
- Root transformation: **pctVacant6up**, **popDensity**
- Categorical transformation: Since **pctUrban** mostly has values which either 0 or 100, **pctUrban** becomes a categorical variable assigning 1 to entries with pctUrban>85 and 0 otherwise

Note that we have chosen to use log base 2 as it is easier to interpret for a non-statistician.

Variable choice

The first variables we will leave out of the model are the **State** and **region** variables. We suggest dropping the **State** variable, as it has a very large number of levels and may be difficult to interpret. There are also several states with a small number of observations, so the estimates may not be very reliable. For the **region** variable, we determine that regional differences in our data are already captured in the other predictor variables; that is, it would be better to fit the model without **region** and then sort the fitted values by **region** to see how well the model works by region.

In the EDA stage, we found that the variables **medIncome**, **ownHousMed**, **ownHousQrange**, **rentMed** and **rentQrange**, which are related to measuring income/wealth, exhibit a strong positive linear correlation with each other (correlation coefficients above 0.6). To further assess the problem of multicollinearity, we can calculate the Variance Inflation Factor (VIF) values for each variable to see which variables are causing the issue. VIF values over 5-10 indicate the given variable may be highly correlated to at least one of the other variables.

Table 2: VIF before and after variable selection

Variable	Coefficient
pctUrban	1.5956
medIncome	16.6049
pctWdiv	7.2124
pctLowEdu	8.2135
pctNotHSgrad	11.9291
pctCollGrad	4.5886
pctUnemploy	4.3961
pctEmploy	5.0872
pctKids2Par	10.0282
pctKidsBornNevrMarr	7.1099
pctHousOccup	1.6566
pctHousOwnerOccup	4.379
pctVacantBoarded	1.9866
pctVacant6up	1.8895
ownHousMed	12.6412
ownHousQrange	6.3884
rentMed	9.7279
rentQrange	2.8701
popDensity	2.5185
pctForeignBorn	3.7914

Variable	Coefficient
pctUrban	1.5042
medIncome	6.3634
pctLowEdu	2.8152
pctUnemploy	3.3160
pctEmploy	2.8375
pctKidsBornNevrMarr	2.9202
pctHousOccup	1.3652
pctHousOwnerOccup	3.2498
pctVacantBoarded	1.8900
pctVacant6up	1.7220
popDensity	2.3233
pctForeignBorn	2.5826

7.1

To see which variables are correlated with each other, we then compute the associated variance proportions. There appears to be certain groups of variables that are highly correlated:

- Income: **medIncome**, **ownHousMed**, **ownHousQrange**, **rentMed** and **rentQrange**, **pctWdiv**
- Family: **pctKidsNevrMarr**, **pctKids2Par**
- Education: **pctLowEdu**, **pctNotHSgrad**, **pctCollGrad**

7.2

Upon further examination of these variables, it becomes clear that we only need to include one variable from each category in the model. This helps the model become more parsimonious and reduces the problem of multicollinearity. This is logical, as these variables encode similar information about each county. To decide which variable to include from each group, we have considered which variable is most highly correlated with the outcome variables. Hence we have kept **medIncome**, **pctKidsBornNevrMarr** and **pctLowEdu** in the model and removed the others.

Bear in mind that VIF is not a method of variable selection, with our penalized likelihood strategy being more appropriate. We are merely using VIF to diagnose multicollinearity in our model and gather appropriate conclusions. After removing the variables described above, we can see **medIncome** is the only variable with a VIF between 5 and 10 from this process, with all other variables below 5, which suggests that the issue of multicollinearity has been improved. This means that the variables selected by LASSO and their associated coefficients will be much more robust and can be accurately interpreted for their effect on crime.

Modelling

With the problem of collinearity solved, we go on to building our linear regression models. We have chosen to use **LASSO regression** for model building for several reasons:

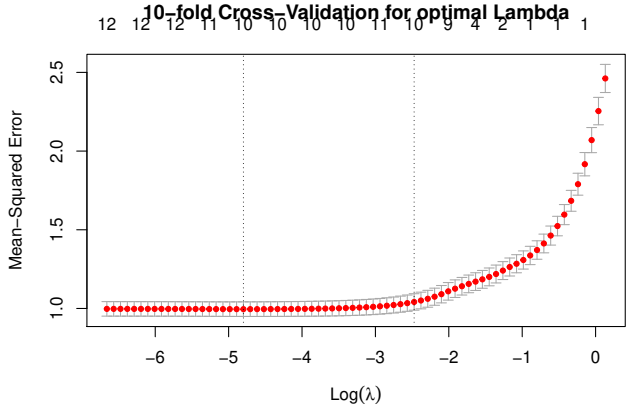
1) LASSO regression is able to perform **variable selection**, meaning that only variables which help predict our outcome variables will be used in our model and the rest will be removed from the model (coefficient equals zero). This helps to reduce the number of predictors needed in the model, thereby making it more parsimonious and allowing stronger explanatory power.

2) LASSO regression uses a **penalized likelihood strategy**. This means that our model is not overfitted to our training data, allowing it to perform strongly when analysing future data. However, this will also result in our model having higher MSE for our training data compared to a normal linear model, but we believe that it is a worthy trade-off between bias and variance in order to improve the performance of the model with future data.

When considering other approaches to model building, such as Ridge regression or stepwise regression with AIC/BIC we find that they lack some of the benefits LASSO can provide. For example, Ridge regression does not perform variable selection, so does not help to reduce the complexity of the model. Meanwhile, stepwise regression is often guilty of overfitting the model to the training data and as such may not fit future data as well as a model generated using LASSO. Hence we have concluded we will use LASSO regression to build our model.

LASSO regression model for violent Crimes

While building the LASSO regression model, we used 10-fold Cross-Validation to find the optimal λ. The graph below shows $log(\lambda)$ against the Mean-Squared Error of the model with the labels on top showing the number of variables in the model. To maintain our model's explanatory power, we will restrict the number of variables in our model to a maximum of 10.



From our graph, we can see that the optimal min value of λ is `0.0082425`, which is shown by the first vertical line. The second vertical line shows the λ value with 1 standard error. This is equal to `0.0843648` and is the value of λ that we will be using.

With this value of λ, we have our LASSO regression model. The coefficients for each of the variables are:

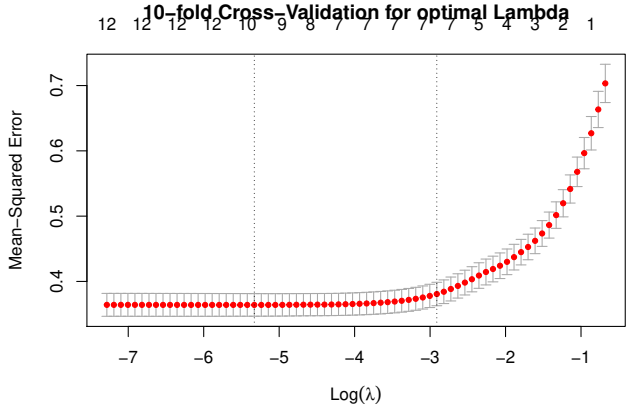
Variable	Coefficient
Intercept	10.629771
pctUrban	0.161918
medIncome	-0.169687
pctLowEdu	0.012729
pctUnemploy	0.09584
pctEmploy	·
pctKidsBornNevrMarr	0.956776
pctHousOccup	-2e-06
pctHousOwnerOccup	-0.001203
pctVacantBoarded	0.083182
pctVacant6up	-0.083068
popDensity	·
pctForeignBorn	0.009602

9.1

By far the largest coefficient here is **pctKidsBornNevrMarr**, which suggests it is the most important determinant of violent crime, even once the transformations of the variables are taken into account. **medIncome** and **pctUrban** also stand out as having a strong influence on violent crime. The variables **pctEmploy** and **popDensity** are omitted which implies that they have little influence on violent crime rates. The Mean-Squared Error for our model is `0.9808623`.

LASSO regression model for non-violent crimes

In a similar fashion to the LASSO model used for predicting violent crime, we also used 10-fold Cross-Validation to find the optimal λ for our non-violent crime model. The graph below shows $log(\lambda)$ versus Mean-Squared Error of the non-violent crime model with the top labels showing the number of variables in the model.



As before, we can use this graph to find our optimal value of λ. We will again use the value of λ with 1 standard error, as this allows us to make the model more parsimonious while still retaining accuracy in the model. This value of λ is given by `0.0544333`.

With this value of λ, we have our LASSO regression model. The coefficients for each of the variables are:

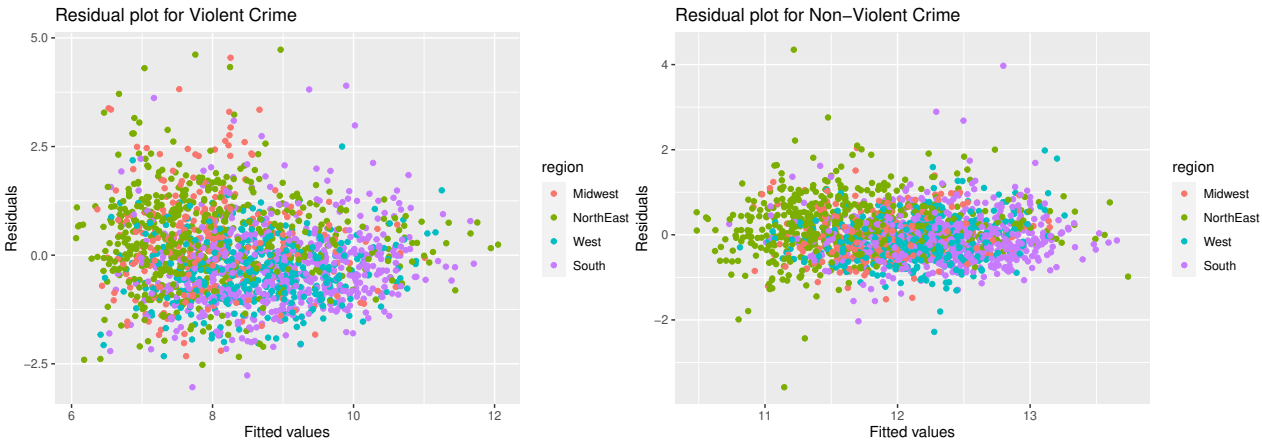
9.2

Variable	Coefficient
Intercept	18.148361
pctUrban	0.061155
medIncome	-0.361809
pctLowEdu	·
pctUnemploy	·
pctEmploy	·
pctKidsBornNevrMarr	0.305732
pctHousOccup	-1e-06
pctHousOwnerOccup	-0.004321
pctVacantBoarded	0.029102
pctVacant6up	-0.067164
popDensity	·
pctForeignBorn	·

The variables with the highest coefficients in the non-violent crime model are **medIncome** and **pctKidsBornNevrMarr**, though the coefficient for the latter variable shows that **pctKidsBornNevrMarr** has a relatively lower impact on non-violent crime compared to violent crime. Additionally, the LASSO process removes the **pctLowEdu**, **pctUnemploy**, and **pctForeignBorn** variables in addition to **pctEmploy** and **popDensity**. We can conclude that these are not crucial factors in predicting non-violent crime; **pctLowEdu**, **pctUnemploy**, and **pctForeignBorn** appear to be much more important for violent crime compared to non-violent crime.

The Mean-Squared Error for our model is `0.3586807`, which is significantly lower than the Mean-Squared Error in the violent crime model. This suggests that our model for non-violent crime is stronger than our model for violent crime.

Residual Analysis



The above plots show the residuals plotted against the fitted values for the two models. We can see that the assumption of linearity in both our models hold, as shown by the fact that the residuals have approximately mean 0 for all fitted values in both plots.

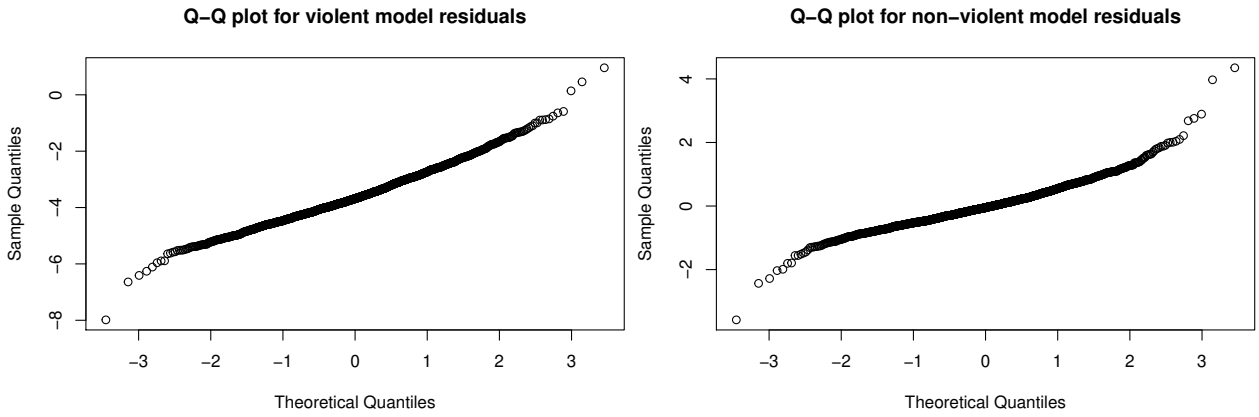
In the case of the violent crime model, we can see some minor heteroscedasticity in the residual

plot, with the residuals appearing more spread out on the left of the plot, with lower fitted values. There are, however, more observations for the lower fitted values and so this may help to explain this phenomenon. The majority of the residuals do appear homoscedastic, which means that our model is still reliable.

11.1

For the non-violent model, the residuals are clearly homoscedastic, with the plot exhibiting points that are evenly spread out above and below the x-axis. This demonstrates the robustness of our model and means that the coefficients for our model are accurate and can be interpreted. In these plots, we can also see the presence of some outliers, which will be talked about in a later section.

In addition to calculating the Mean-Squared Error, we can analyse the residuals for each of our models. We will do this by using Q-Q plots, beginning with the violent crime model.



There are some things that we picked up from the Q-Q plot of the models' residuals:

- **The LASSO model for both violent and non-violent crime is generally a good fit.** From the Q-Q plot, we can deduce that the residuals of the model are normally distributed, this means that there is a linear relationship between our explanatory variables and our outcome variables, violent/non-violent crimes per 100k population.
- Points at the extreme values are not following the same linear relationship as the rest of the data. These points represent extreme values of residuals; these are also points that we investigated earlier where we identify places that have an unusual amount of violent crime.
- However, there is one key difference between our violent and non-violent model. The line in our violent model is a lot steeper than the line in our non-violent model, which means that the residuals in our non-violent model are smaller than the ones in our violent model.

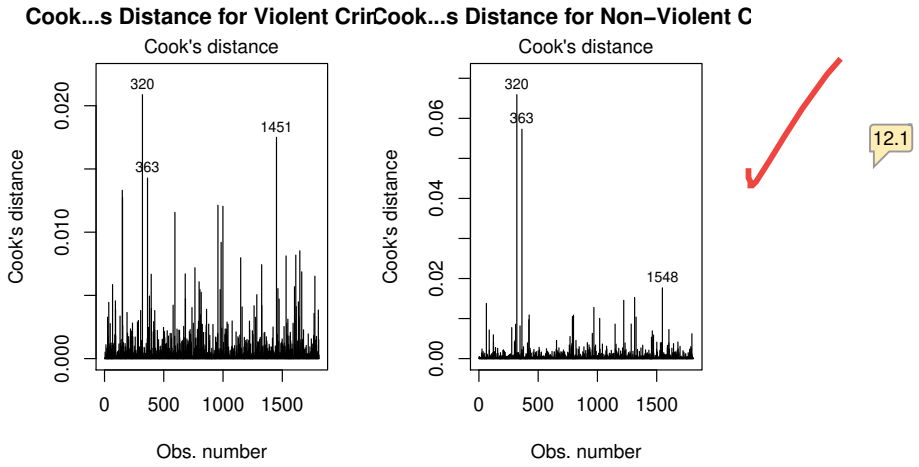
11.2

We can also conduct some investigation of our model to look for outliers. As seen on the residual plots for the non-violent crime model, there appears to be two points with significantly larger residuals than the other observations. To assess whether these points can be considered as outliers, we can look at the Cook's distance of each point in the plot below. For the non-violent model, the values of the Cook's distance for the observations with index 320 and 363 are over three times that of any other observation and so can be considered outliers in the model.

When considering the violent crime model, we can see that although there are some values with larger Cook's distance than the rest of the points, these are nowhere near as extreme as the values seen in the non-violent crime model. It is important to remember that with a large number of observations, it is likely that some will have larger residual values, which does not necessarily make

them outliers. Hence we do not consider any of the observations from the violent crime model to be outliers.

To deal with the outliers described in the non-violent crime model, we can fit the model with and without these two points, to see how sensitive the model is to them. After removing these two points and refitting the model, the values of the coefficients hardly change. This is logical, as although these points are influential, they are just 2 of 1810 observations in the entire model. Hence we can proceed to fit the model with these two outliers still included.



12.1

Limitations of our Model

One of the first limitations of our model is in the transformations applied to the predictor variables. We are in a trade-off between accuracy and interpretability. The more intuitive transformations used do make the model easier to interpret, however, it does mean that we could lose some accuracy in our model, with the model potentially being less suited to the data compared with some more complex transformations. This was also the case if a fractional power was suggested, for most of the transformations, we rounded the suggested power transform, which will impact the larger values of the data. This may result in the densities of some of the variables not being totally symmetrical, as may be possible with a specific power, but it does have the benefit of being easier to interpret and implement for policymakers.

12.2

Additionally, at the start, we excluded some variables based on findings from the EDA. While we showed that this helped with the multicollinearity of the variables, we cannot guarantee doing this did not adversely affect the model.

As well, we used the LASSO method for variable selection. One limitation of this method is that if it has to choose between two correlated variables, it makes the choice arbitrarily. Also, the number of variables chosen in this method is limited by the number of data observations. While this did not impact our model, if someone was to recreate our model with a smaller data set, they would get a limited model. This could impact the reproducibility of our model.

Appendix

```r
#Loads in data, adjust this for own use
load("USACrime.rda")

#Loads appropriate packages, which will be useful when creating plots.
library(knitr)
library(car)
library(dplyr)
library(glmnet)
library(ggplot2)

#Data cleaning
USACrime$ownHousMed[USACrime$ownHousMed==500001]<-NA
USACrime$rentMed[USACrime$rentMed==1001]<-NA
USACrime$ownHousQrange[USACrime$ownHousQrange==0]<-NA
USACrime$rentQrange[USACrime$rentQrange==0]<-NA
USACrime <- na.omit(USACrime)

#Create a copy of the original data for future use
USACrimeCopy <- USACrime

#Fixing State/region variable
USACrime$State<-droplevels(USACrime$State)
levels(USACrime$region) <- c("Midwest", "NorthEast", "West", "South", "West")

#Removing State and region variables for investigation
USACrime = subset(USACrime, select = -c(State,region))

#Transformation of the outcome variables
USACrime$violentPerPop <- log(USACrime$violentPerPop,2)
USACrime$nonViolPerPop <- log(USACrime$nonViolPerPop,2)

#Categorical Transformation
USACrime = USACrime %>% mutate(pctUrban = ifelse(pctUrban >= 85,1,0))

#Log2 Transformations
USACrime$medIncome <- log(USACrime$medIncome,2)
USACrime$pctLowEdu <- log(USACrime$pctLowEdu,2)
USACrime$pctUnemploy <- log(USACrime$pctUnemploy,2)

#Log2+1 Transformations
USACrime$pctKidsBornNevrMarr <- log(USACrime$pctKidsBornNevrMarr+1,2)
USACrime$pctVacantBoarded <- log(USACrime$pctVacantBoarded+1,2)

#Power Transformations
USACrime$pctEmploy <- (USACrime$pctEmploy)^2
```

```r
USACrime$pctHousOccup <- (USACrime$pctHousOccup)^3

#Root Transformations
USACrime$pctVacant6up <- sqrt(USACrime$pctVacant6up)
USACrime$popDensity <- sqrt(USACrime$popDensity)

#Calculating VIF
model1 <- lm(violentPerPop ~ pctUrban + medIncome + pctLowEdu +
               pctUnemploy + pctEmploy
+ pctKidsBornNevrMarr + pctHousOccup +
               pctHousOwnerOccup + pctVacantBoarded +
pctVacant6up + popDensity + pctForeignBorn, data = USACrime)
vif(model1)

model1 <- lm(violentPerPop ~ pctUrban + medIncome + pctLowEdu + pctUnemploy + pctEmploy +
pctKidsBornNevrMarr + pctHousOccup + pctHousOwnerOccup + pctVacantBoarded + pctVacant6up +
popDensity + pctForeignBorn, data = USACrime)
vif(model1)


#LASSO Regression (violentPerPop)
USACrimeLasso.viol = subset(USACrime, select = c(violentPerPop, pctUrban, medIncome,
pctLowEdu, pctUnemploy, pctEmploy, pctKidsBornNevrMarr, pctHousOccup, pctHousOwnerOccup,
pctVacantBoarded, pctVacant6up, popDensity, pctForeignBorn))

exp.variables.viol <- data.matrix(USACrimeLasso.viol[,2:13])
out.variables.viol <- USACrimeLasso.viol$violentPerPop

#Fit and testing the lasso model
lassoviol.fit <- cv.glmnet(exp.variables.viol, out.variables.viol,
type.measure = "mse", alpha = 1, family = "gaussian")
lassoviol.predicted <- predict(lassoviol.fit, s = lassoviol.fit$lambda.1se,
newx = exp.variables.viol)

plot(lassoviol.fit, main = "10-fold Cross-Validation for optimal Lambda") #Graph for Lambda

mean((out.variables.viol -lassoviol.predicted)^2) #Finding Mean-Squared Error

#Look at the coefficients
coef(lassoviol.fit)

#LASSO Regression (nonViolPerPop)
USACrimeLasso.nonviol = subset(USACrime, select = c(nonViolPerPop, pctUrban,
medIncome, pctLowEdu, pctUnemploy, pctEmploy, pctKidsBornNevrMarr, pctHousOccup,
pctHousOwnerOccup, pctVacantBoarded, pctVacant6up, popDensity, pctForeignBorn))

exp.variables.nonviol <- data.matrix(USACrimeLasso.nonviol[,2:13])
```

```
out.variables.nonviol <- USACrimeLasso.nonviol$nonViolPerPop

#Fit and testing the lasso model
lassononviol.fit <- cv.glmnet(exp.variables.nonviol, out.variables.nonviol,
type.measure = "mse", alpha = 1, family = "gaussian")
lassononviol.predicted <- predict(lassononviol.fit, s = lassononviol.fit$lambda.1se,
newx = exp.variables.nonviol)

plot(lassononviol.fit, main = "10-fold Cross-Validation for optimal Lambda") #Graph for Lambda

mean((out.variables.nonviol - lassononviol.predicted)^2) #Finding Mean-Squared Error

#Look at the coefficients
coef(lassononviol.fit)

#Graph Plotting
lassoviol.predict.test <- data.frame(lassoviol.predicted, out.variables.viol)
lassoviol.predict.test$State <- USACrimeCopy$State #Add State and region variables back
#for analysis
lassoviol.predict.test$region <- USACrimeCopy$region

lassononviol.predict.test <- data.frame(lassononviol.predicted, out.variables.nonviol)
lassononviol.predict.test$State <- USACrimeCopy$State #Add State and region variables
#back for analysis
lassononviol.predict.test$region <- USACrimeCopy$region

#Residual Plots
ggplot(data = lassoviol.predict.test , aes(X1,(X1-out.variables.viol))) +
  geom_point(aes(color = region)) + labs(title="Residual plot for Violent Crime")

ggplot(data = lassononviol.predict.test , aes(X1,(X1-out.variables.nonviol))) +
  geom_point(aes(color = region)) + labs(title="Residual plot for Non-Violent Crime")

#Q-Q plot analysis for both violent and non-violent model
lassoviol.predict.test$residual <- (lassoviol.predict.test$X1 -
lassoviol.predict.test$out.variables.viol)
qqnorm(lassoviol.predict.test$residual, main = "Q-Q plot for violent model residuals")

lassononviol.predict.test$residual <- (lassononviol.predict.test$X1 -
lassononviol.predict.test$out.variables.nonviol)
qqnorm(lassononviol.predict.test$residual, main = "Q-Q plot for non-violent model residuals")

#Facet Residual Plot
lassoviol.predict.test$viol <- "Violent Crimes"
lassononviol.predict.test$viol <- "Non-Violent Crimes"
viol.results <- rename(lassoviol.predict.test, out.variables = out.variables.viol)
nonviol.results <- rename(lassononviol.predict.test, out.variables = out.variables.nonviol)
```

```r
all.results = merge(viol.results,nonviol.results,all = TRUE)
ggplot(data = all.results , aes(X1,residual)) + geom_point(aes(color = region)) +
  labs(title="Residual plot for Non-Violent Crime and Violent Crime model",
 x = "Fitted values", y = "Residuals") +
facet_wrap(~viol, scales = "free") + theme(legend.position="bottom")

#Cook's Distance for both models
lm1<-lm(violentPerPop~pctUrban+medIncome+pctLowEdu+pctUnemploy+ pctKidsBornNevrMarr
       +pctHousOccup+pctHousOwnerOccup+pctVacantBoarded+pctVacant6up+pctForeignBorn)
lm2<-lm(nonViolPerPop~pctUrban+medIncome+pctKidsBornNevrMarr+pctHousOccup
      +pctHousOwnerOccup+pctVacantBoarded+pctVacant6up)
par(mfrow=c(1,2))
plot(lm1,4, main="Cook's Distance for Violent Crime")
plot(lm2,4, main="Cook's Distance for Non-Violent Crime")

#Residual Plot with high residuals
ggplot(data = lassononviol.predict.test[which(
abs(lassononviol.predict.test$X1-lassononviol.predict.test$out.variables.nonviol) > 2),] ,
aes(X1,(X1-out.variables.nonviol))) +
  geom_point(aes(color = region), size = 2.5) +
  labs(y="Residuals", x="Fitted values") +
  geom_text(aes(label=State),hjust=1, vjust=1)
```

Author's Contributions

Contributor	Contributed Work	Proposed mark weighting (%)
Liam Briggs	**Findings:** Data Cleaning, Variable Choice **Statistical Methodology:** Data Cleaning, Variable Choice General writing/editing/proofreading	100
Oliver Robinson	**Findings:** Modelling **Statistical Methodology:** Data Cleaning, Variable Choice, Residual Analysis Proofreading	100
Tom Tanner	**Findings:** Limitations of the model **Statistical Methodology:** Data Cleaning, Variable Choice, Limitations of the model	100
Zhicong Hu	**Findings:** Model Diagnostics, Limitations of the model **Statistical Methodology:** Modelling, Residual Analysis	100

Index of comments

2.1 I think this is poorly phrased. These are the variables you kept, not the ones you omitted.

2.2 This is true, but in practice you should set the seed for the random number generator so that your results are reproducible.

2.3 It is important to have a summary of the estimates for the model. However, this could be made more accessible by describing the effect of changes in each variable on the crime rate (e.g. % increase in crime rate for a given increase in predictor variable) taking into account the transformations of the predictor variables, which are not specified here. In addition, standardization of coefficients can help to draw attention to the major determinants of differences in crime rates.

3.1 You need the standardized coefficients to draw this conclusion.

5.1 It should have lower mean squared error for a new data set.

6.1 Good.

6.2 I am not sure this is true. We found substantial differences between regions that were not accounted for by the other variables.

7.1 You do not need this level of precision

7.2 Good. It is always important to think about what the variables mean. This can guide when interpreting strong correlations.

8.1 You are repeating yourself here. It is a good idea when editing your submission to make sure that you do not make the same point twice.

9.1 It is true that this variable is one of the most important predictors. However, the argument based on the size of the coefficient is only valid if you use standardized coefficients.

9.2 Some more repetition. You do not need to explain the process again.

11.1 Linearity is more important than homoscedasticity

11.2 You do not need a QQ plot for this. The summary() function will give you directly the residual standard error.

12.1 Good.

12.2 Good.