Assignment 2

Group A - 1806987, 1800673, 1802212, 1824442

Findings

Our models found several major determinants of high crime rates. Income, region, the proportion of population born in foreign countries, house occupancy, urbanisation, education, house/rental prices and the environment in which children are raised are all important factors in models for both violent and non-violent crime. There are some differences in how these factors influence violent and non-violent crime. The models gave the following coefficients (rounded to three significant figures)

Variable	**Violent**	**Non-Violent**
Intercept	6.28	10.8
`region` - North East	-0.171	-0.161
`region` - West	0.206	0.113
`region` - South	0.182	0.0388
`isUrban` - Urban	0.151	0.0787
`medIncome`	-0.112	0
`pctWdiv`	-0.0113	-0.00298
`rentMed`	0.0688	0
`rentQrange`	0.109	-0.107
`ownHousMed`	0	-0.0650
`pctKids2Par`	-0.0182	-0.0171
`pctKidsBornNevrMarr`	0.702	0
`pctHousOccup`	-0.0000279	-0.0000198
`pctVacantBoarded`	0.131	0.0666
`pctVacant6up`	-0.00318	-0.00353
`pctLowEdu`	0.0900	0.0581
`pctCollGrad`	0	0.0686
`pctEmploy`	0	0.000505
`popDensity`	0.0702	0
`pctForeignBorn`	0.0634	0.0385

These coefficients are after transformation on both the outcome and several of the explanatory variables which can be seen in the formulae for the models here:

$$
\begin{gathered}
\log(\texttt{violentPerPop}) = 6.28 - 0.171\mathbb{1}_{\texttt{NorthEast}} + 0.206\mathbb{1}_{\texttt{West}} + 0.182\mathbb{1}_{\texttt{South}} \\
+0.151\mathbb{1}_{\texttt{isUrban}>85} - 0.0182\texttt{pctKids2Par} + 0.0634\log(\texttt{pctForeignBorn}) - 0.0113\texttt{pctWdiv} \\
-0.0000279e^{\frac{\texttt{pctHousOccup}}{10}} + 0.131\log(\texttt{pctVacantBoarded} + 1) + 0.0900\log(\texttt{pctLowEdu}) \\
+0.109\log(\texttt{rentQrange}) - 0.00318\texttt{pctVacant6up} + 0.702\texttt{pctKidsBornNevrMarr}^{0.25} \\
-0.112\log(\texttt{medIncome}) + 0.0688\log(\texttt{rentMed}) + 0.0702\log(\texttt{popDensity}) + \epsilon
\end{gathered}
$$

$$
\begin{gathered}
\log(\texttt{nonViolPerPop}) = 10.8 - 0.161\mathbb{1}_{\texttt{NorthEast}} + 0.113\mathbb{1}_{\texttt{West}} + 0.0338\mathbb{1}_{\texttt{South}} \\
+0.0787\mathbb{1}_{\texttt{isUrban}>85} - 0.0171\texttt{pctKids2Par} + 0.0385\log(\texttt{pctForeignBorn}) - 0.00298\texttt{pctWdiv} \\
-0.0000279e^{\frac{\texttt{pctHousOccup}}{10}} + 0.0666\log(\texttt{pctVacantBoarded} + 1) + 0.0581\log(\texttt{pctLowEdu}) \\
-0.107\log(\texttt{rentQrange}) - 0.00353\texttt{pctVacant6up} + 0.000505\texttt{pctEmploy} + \\
0.0686\log(\texttt{pctCollGrad}) - 0.0650\log(\texttt{ownHousMed}) + \epsilon
\end{gathered}
$$

From the table we can gather the differences in what affects violent and non-violent crime. `isUrban` has a larger coefficient in the violent crime model than in the non-violent model so we gather that urban areas have higher incidences of both violent and non-violent crime but the effect is much more dramatic for violent crime.

Income appears to be less important in predicting violent crime than non-violent crime since the non-violent model does not include `medIncome` and also has a smaller coefficient than the violent model for `pctWdiv`. However the model for non-violent crime does include `ownHousMed` which is closely related to income. Areas where income is higher or where there are more people receiving dividend or rent income have lower rates of violent crime.

When it comes to housing/rent prices, the median rent price is an important predictor for violent crime - higher rent predicts lower crime rates, while median house prices and rent interquartile range is important for predicting non-violent crime rates. Higher hose prices correspond to lower non-violent crime rates and higher disparity between rent prices — i.e. more income inequality in an area — leads to higher rates of non-violent crime.

Both models predict that a higher percentage of children being raised in two-parent homes leads to a lower rate of violent and non-violent crime, but a much larger effect can be seen in the model for violent crime with regards to the percentage of children born to unmarried parents. Having a high coefficient of 0.702, our model predicts that violent crime increases fairly dramatically with a rise in children being born to single parents.

The proportion of housing being occupied has a similarly small effect on crime rates in both models, more housing being left empty leads to a slight increase in both crime rates.
The percentage of vacant and boarded up housing (`pctVacantBoarded`) has a fairly large effect on crime rates — more so for violent crime — however the percentage of housing vacant for over 6 months has the opposite (albeit much smaller) effect on crime rates. As such we might expect a county which has a large increase in the percentage of boarded up housing to have an increase in crime rates which would then slowly decrease after about six months. This could, for example, be due to more development in areas where housing has been boarded up however this would be

impossible to confirm without looking at other variables which are not included in this data set such as city/county budgets etc.

Low education (defined as people 25 or over with less than 9th grade education) predicts higher rates of both violent and non-violent crime, with a stronger effect being seen for violent crime. On the other hand the percentage of the population who graduated from college predicts higher rates of non-violent crime but does not seem to have much effect on rates of violent crime, at least after the proportion of those with low education has been accounted for.

Higher employment rates lead to a fairly small increase in rates of non-violent crime and have not been included in our violent crime model after variable selection - presumably since accounting for income and education already can better explain crime rates than employment can.

A higher population density seems to increase violent crime rates but not non-violent crime, this is consistent with `isUrban` having a larger effect for violent crime than for non-violent crime since we expect urban areas to have more dense populations.

Finally our models found that the percentage of population born in a foreign country had an increasing effect on crime rates, more so for violent crime.

So we can see that the factors effecting violent crime are mostly similar. Some variables have different effects as seen by differences in magnitude of coefficients. The two models select different variables in a lot of cases but usually they both select at least one variable to explain a particular factor such as housing/rent prices. The exception here is employment rate, which only seems to affect non-violent crime.

To examine whether there were any areas in the US which did not fit the general pattern, we calculated the difference between actual and fitted values from our models for violent and non violent crime across states. We found that all states fit the pattern predicted by the model well in the case of non-violent crime, however when looking at errors in predicting violent crime there was one state — North Dakota — whose actual values for rates of violent crime were fairly consistently lower than our model predictions, so it appears that North Dakota does not fit the general pattern in crime rates across the US for violent crime.

In terms of problems in our data set, we would have liked to have seen data on community/local government budgets and police numbers/police spending. More police reducing crime is not something which has been completely proven but there is some evidence that increase in policing numbers can lead to reduction in crime, at least in the UK (Police and Crime Reduction Paper) and as such the data on number of police or police spending in each community may have proven useful in predicting crime rates, since we would expect this trend to be similar in other countries.

Statistical Methodology

Cleaning Data

After the exploratory data analysis completed in assignment 1 we decided to do the following to clean up the data:

- Remove the NAs in `medIncome` and `pctEmploy` and incorrect values in `ownHousMed`, `ownHousQrange`, `rentMed` and `rentQrange` (Appendix 1)
- Merge the Pacific level in region into West (Appendix 2)
- Use `pctUrban` values to create a Boolean variable splitting at 85% with 1 being Urban and 0 being not Urban (Appendix 3)

Transformations

To create a model of this data we must first transform some of the variables that are heavily skewed. We decided that a skewness value between -0.7 and 0.7 was acceptable and found the variables that were outside of this range. We will only use this method to transform the explanatory variables and transform the outcome variables to ensure homoscedasticity.

`medIncome`, `pctLowEdu`, `pctNotHSgrad`, `pctCollGrad`, `pctUnemploy`, `pctKidsBornNevrMarr`, `pctVacantBoarded`, `ownHousMed`, `ownHousQrange`, `rentQrange`, `popDensity` and `pctForeignBorn` are all the positively skewed variables and `pctHousOccup` is the only negatively skewed variable. (Appendix 4)

For the positively skewed variables our aim was to get the skewness in the acceptable range so transformed using log and square root and used the least skewed transformation as we look forward to modeling. However variables `pctKidsBornNevrMarr` and `pctVacantBoarded` both have 0 values so we cannot use the log transformation so we will deal with this separately. (Appendix 5)

From `pctKidsBornNevrMarr`, `pctVacantBoarded` we try square root, 4th root and taking the log(x+1) and from this we see that `pctKidsBornNevrMarr` should be 4th rooted and `pctVacantBoarded` should have a log(x+1) transformation.

For the 1 negatively skewed variable, `pctHousOccup`, taking the exponential of (`pctHousOccup`/10) completed the transformation to an acceptable level.

The list of transformations for all variables (Appendix 6):

- `medIncome` - log
- `pctLowEdu` - log
- `pctNotHSgrad` - square root
- `pctCollGrad` - log
- `pctUnemploy` - log
- `ownHousMed` - log
- `ownHousQrange` - log
- `rentQrange` - log
- `popDensity` - log
- `pctForeignBorn` - log
- `pctKidsBornNevrMarr` - 4th root
- `pctVacantBoarded` - log(x + 1)
- `pctHousOccup` - e^(x/10)

Variables Selection

During our EDA we gave suggestions of the variables we would recommend for a linear model. We wanted variables that covered the whole data set with high correlation to the outcome variables however low correlation between each other so elected for the following: `region`, `isUrban`, `pctNotHSgrad`, `pctWdiv`, `medIncome`, `rentQrange`, (`pctUnemploy` or `pctEmploy`), (`pctKids2Par` or `pctKidsBornNevrMarr`), (`pctHousOccup` or `pctHousOwnerOccup`), (`ownHousMed` or `rentMed`).

To find the best variables of the ones above we chose to find the combination that minimized the RSS/AIC (as for a fixed number of explanatory variables the result will be the same). The resulting linear models gave an AIC of 3564.779 for `violentPerPop` and 1777.114 for `nonViolPerPop` (Appendix 7). We will use these models and values as a bench-mark for more rigorous model creation.

Stepwise regression with AIC

We tried applying bidirectional elimination step-wise regression with AIC, one starting with the null model and another starting from the maximal model. (Appendix 8)
Through this method, we found that both violent crime models have 14 parameters, and only 1 out of the 14 parameters differ (`pctHousOwnerOccup` in first model, `medIncome` in second model). Since step-wise regression lead to very similar final models, we can be confident that one of the models is the global optimal model. The coefficients shown below show how similar the models are.
Applying the same technique to model non-violent crime, a similar result is obtained (Appendix 8). Both models have 13 explanatory variables in common and differ only by two variables (first model has `pctNotHSgrad` and `ownHousQrange`, while second model has `pctCollGrad`). Since the two models overlap significantly, we are again confident that the global optimal model is the one with the lower AIC value. The coefficients shown below show how similar the models are. (Note: b/f is step-wise starting at maximal model, f/b is step-wise starting at minimal model)

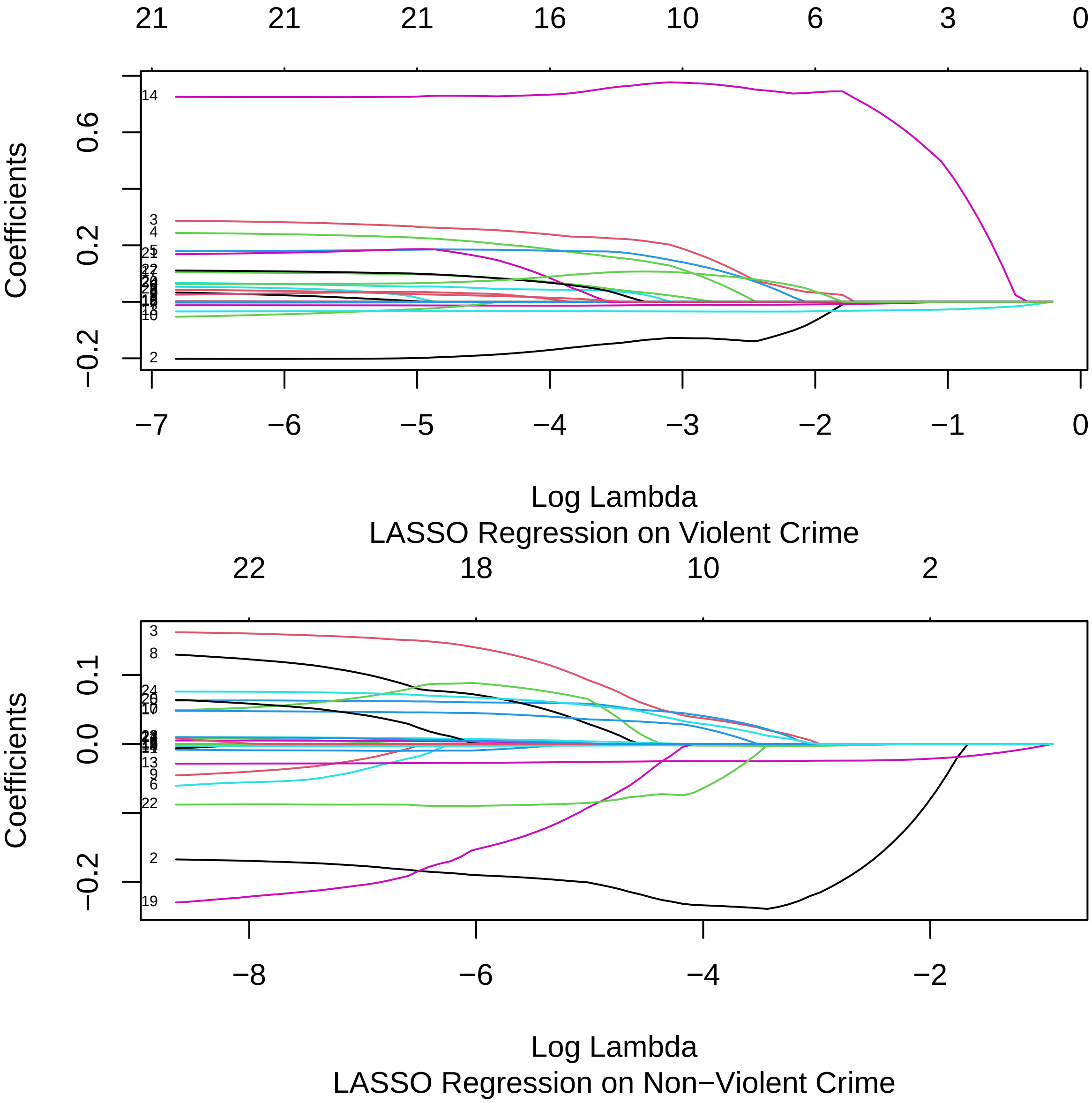
Variable	**Violent b/f**	**Violent f/b**	**Non-Violent b/f**	**Non-Violent f/b**
Intercept	4.93	3.71	12.4	11.9
`region` - North East	-0.203	-0.205	-0.162	-0.180
`region` - West	0.297	0.305	0.162	0.163
`region` - South	0.247	0.243	0.0149	0.00553
`isUrban` - Urban	0.182	0.180	0.0649	0.0581
`pctWdiv`	-0.0134	-0.0137	0.00526	0.00470
`pctNotHSgrad`	N/A	N/A	-0.0843	N/A
`pctCollGrad`	N/A	N/A	N/A	0.121
`pctLowEdu`	0.0761	0.0930	0.164	0.0884
`pctEmploy`	N/A	N/A	0.00929	0.00909
`pctKids2Par`	-0.0334	-0.0347	-0.0298	-0.0292
`pctKidsBornNevrMarr`	0.737	0.680	N/A	N/A
`pctHousOccup`	0.0000308	0.0000324	0.0000208	0.0000229
`pctHousOwnerOccup`	0.00246	N/A	N/A	N/A
`pctVacantBoarded`	0.110	0.108	0.0449	0.0401
`pctVacant6up`	-0.00274	-0.00239	-0.00299	-0.00276
`medIncome`	N/A	0.215	N/A	N/A
`rentMed`	0.290	0.186	N/A	N/A
`rentQrange`	0.123	0.114	-0.0921	-0.121
`ownHousMed`	N/A	N/A	-0.258	-0.171
`ownHousQrange`	N/A	N/A	0.0847	N/A
`popDensity`	0.0433	0.0401	N/A	N/A
`pctForeignBorn`	0.0569	0.0560	0.0811	0.0756

With both models having the same number of parameters (in the case of violent crime), we decided to choose the model with the lower AIC value, which corresponds to the model with the lower mean squared error, as such we choose the model we get when starting from the minimal model for both violent and non-violent crime (with AIC values of 3505.804 and 1525.758 respectively). Both of these models produced better AIC values than our EDA produced benchmark.

LASSO regression

We also wanted to consider penalised regression methods for finding an optimal linear model. Here, we use LASSO regression which will perform both variable selection and coefficient penalisation. We started with all of the possible explanatory variables (excluding `State`) and ran a LASSO regression algorithm which gave us the following plots (Appendix 9):

Figure A

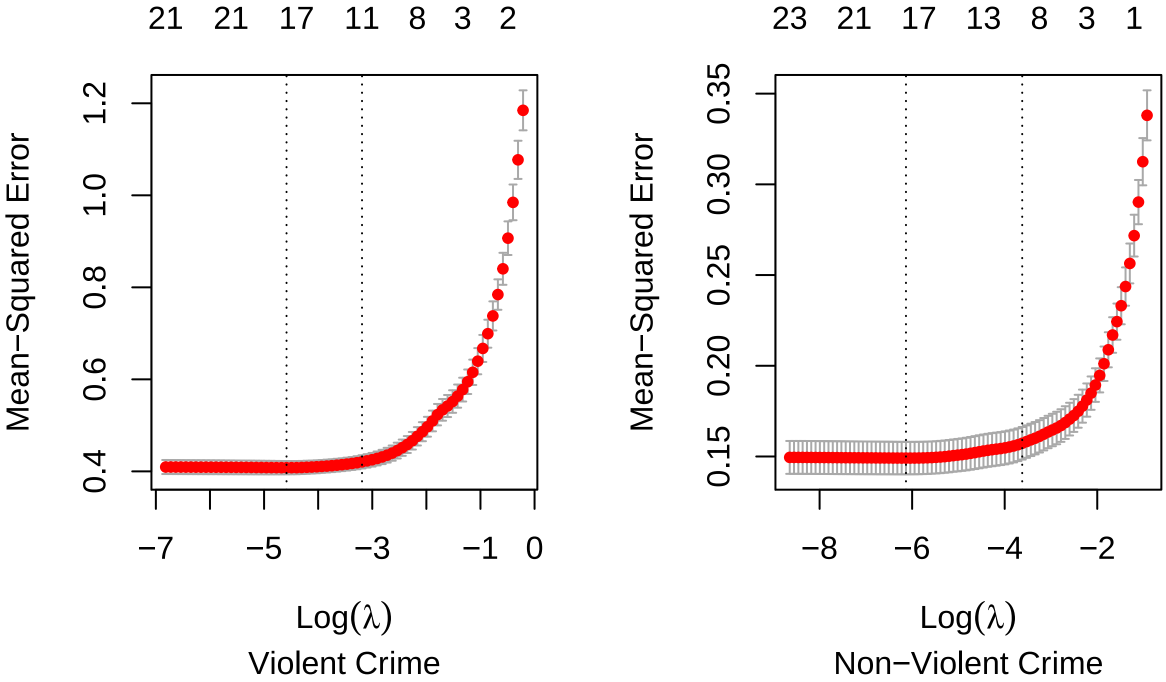


Note that the variables `region`, `isUrban`, `pctHousOccup` and `pctHousOwnerOccup` still have non-zero coefficients for higher values of lambda (our penalty factor) in the case of violent crime, and `region`, `isUrban`, `pctHousOccup` and `pctWdiv` have non-zero coefficients for higher values of lambda. This is fairly consistent with what we found when using step-wise regression for variable selection:

`region` and `isUrban` are identified as important by both step-wise and LASSO regression for both types of crime. We also see that `pctHousOccup` and `pctWdiv` are important explanatory variables for non-violent crime according to both step-wise and LASSO regression. Additionally `pctHousOccup` is chosen by step-wise regression for violent crime and is kept for most values of lambda in lasso regression. The only difference we see is that `pctHousOwnerOccup` is not chosen by step-wise regression when modelling violent crime rates but it is kept for most values of lambda by LASSO, although with a very small coefficient.

Here we find optimal values of lambda to minimise error in the model (Appendix 10):

Figure B



Violent Crime

Non−Violent Crime

We take the value of lambda as the maximum which is one standard error away from the value of lambda which minimises mean square error, which for violent crime is $\lambda = 0.05445$ and for non-violent crime is $\lambda = 0.03881$. We then get the following model coefficients (rounded to 3 significant figures):

Variable	**Violent**	**Non-Violent**
Intercept	7.49	10.6
`region` - North East	-0.132	-0.237
`region` - West	0.209	0.0273
`region` - South	0.136	0
`isUrban`	0.157	0.0292
`pctWdiv`	-0.0116	0
`pctKids2Par`	-0.0338	-0.0250
`pctKidsBornNevrMarr`	0.774	0
`pctHousOccup`	-0.0000165	-0.0000114
`pctHousOwnerOccup`	0	-0.00258
`pctVacantBoarded`	0.0279	0.00579
`pctVacant6up`	0	-0.000767
`ownHousQrange`	0.0142	0
`rentQrange`	0	-0.0265
`pctForeignBorn`	0.107	0.0188

A problem with these models is that they both select pairs of variables which have very high correlation. For example the model for violent crime picks both `pctKids2Par` and `pctKidsBornNevrMarr` and the non-violent model picks both `pctVacantBoarded` and `pctVacant6up` and as such these models are likely to have problems regarding multicollinearity. Additionally we see that not all levels of

the categorical variable `region` have been given coefficients in the model for non-violent crime which is problematic since we know there are differences in crime rates across regions.
To address this, we try taking the the variables picked by the LASSO regression and fit a normal least squares regression on those variables (Appendix 10):
The violent crime linear model with variables chosen by LASSO has an AIC of 3520.328 and the non-violent crime linear model with variables chosen by LASSO has an AIC of 1731.793

We find that these models are not an improvement on our previous models as they have a higher AIC and as such we abandon these models.

Comparing the Two Models

After completing the two model creation methods we now have to decide which models are better for each of the outcome variables. To do this we calculated the Mean Square Error (MSE) for each of the models (Appendix 11).
Getting the following results:

- Stepwise Regression Model for Violent Crime - MSE = 0.398177113804347
- Stepwise Regression Model for Non-Violent Crime - MSE = 0.145304970809664
- LASSO Regression Model for Violent Crime - MSE = 0.42072263243735
- LASSO Regression Model for Non-Violent Crime - MSE = 0.157273523662924

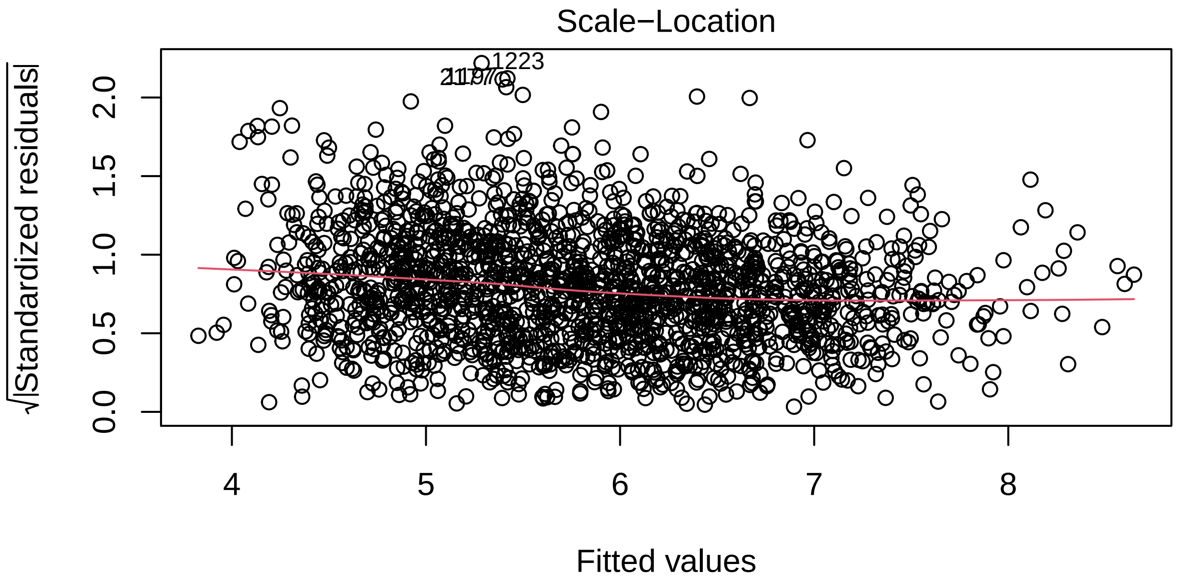
From this we see that the models using step-wise regression (minimising AIC) to select variables give the lowest MSE.

Outliers and Multicollinearity

After finalising the model parameters, there may be outliers or multicollinearity affecting the parameter coefficients of our model, thus we need to investigate further into the variables.

Violent Crime Model

Figure C



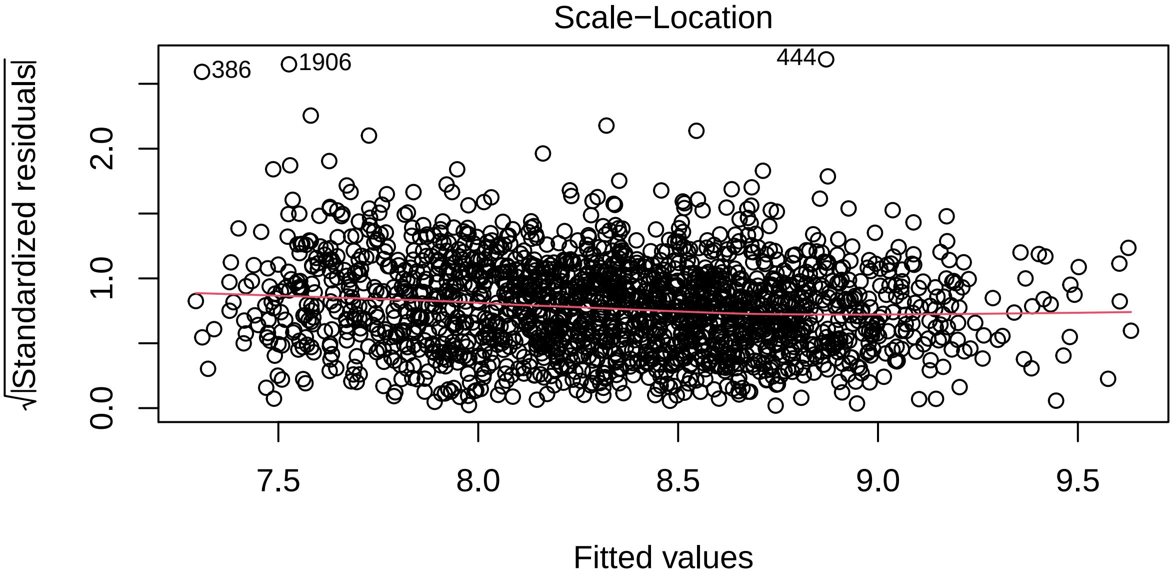
Through the standardised residuals plot, we conclude that there are no outliers. However, before finalising the parameter coefficients, we noticed that the model fitted certain variables which had some correlation with each other, such as `rentMed` with `medIncome`.

Including variables which are highly correlated with each other in the linear model may lead to unreliable parameter coefficients. Therefore, we compute the variance inflation factors to check for multicollinearity in our model.

With `pctKids2Par`, `pctKidsBornNevrMarr`, `medIncome` and `rentMed` having variance inflation factor >5, it is evidence of multicollinearity being present in our model. To mitigate this, we decided to set our parameter coefficients using cross validation ridge regression. (Appendix 12)

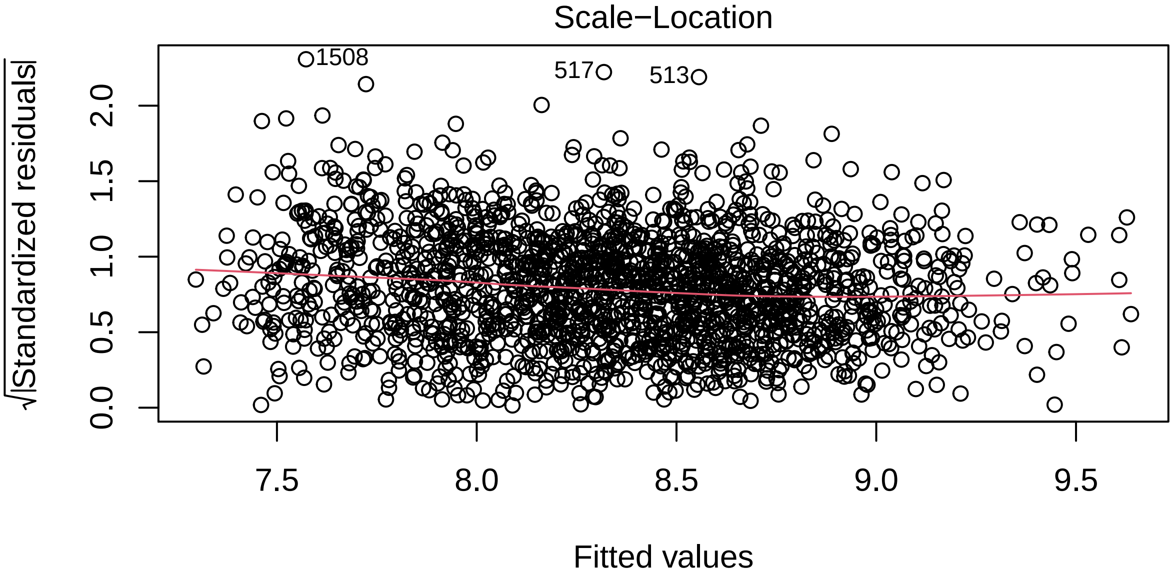
Non Violent Crime Model

Figure D



The standardized residuals plot above shows that there are 3 outliers. Although further investigation suggests that the 3 observations are valid, we would still remove them from the data set before fitting a linear model to prevent the 3 outliers from affecting the model.

Figure E



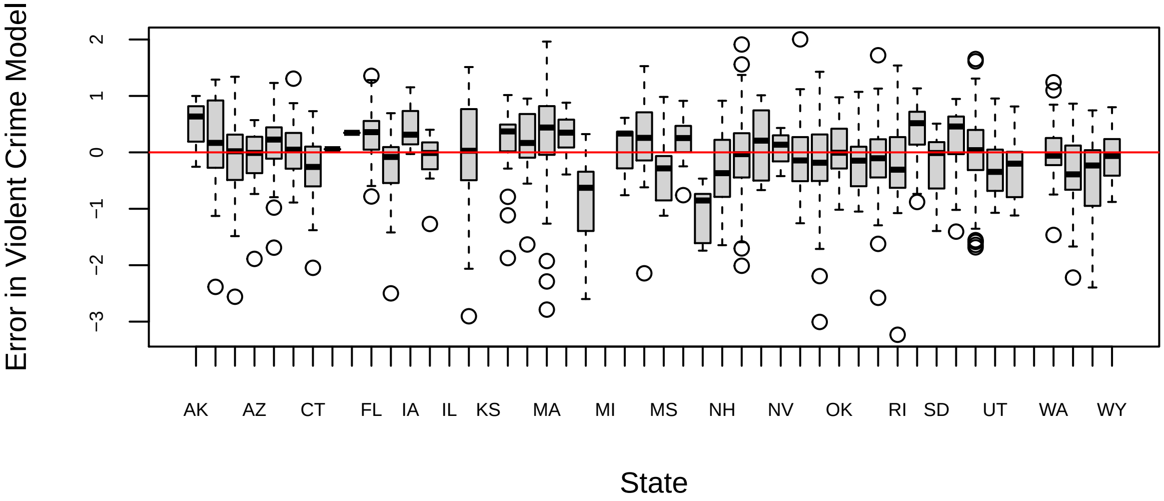
From the new standardised residual plot we conclude there are no more outliers, and proceed to the next stage of our model investigation, multicollinearity. The VIF value provides some evidence of multicollinearity as `ownHousMed` and `pctWdiv` has a variance inflation factor >5. Thus, similar to the violent crime model, we mitigate this by applying cross validation ridge regression to find the appropriate coefficients for our non violent crime model. (Appendix 13)

Final Model

Using 2 step-wise regression to finalise the model parameters, together with cross validation ridge regression to manage multicollinearity in our model parameters and maximise the predictive power of our model concurrently, we have found our final model parameters for both violent and non violent crime.

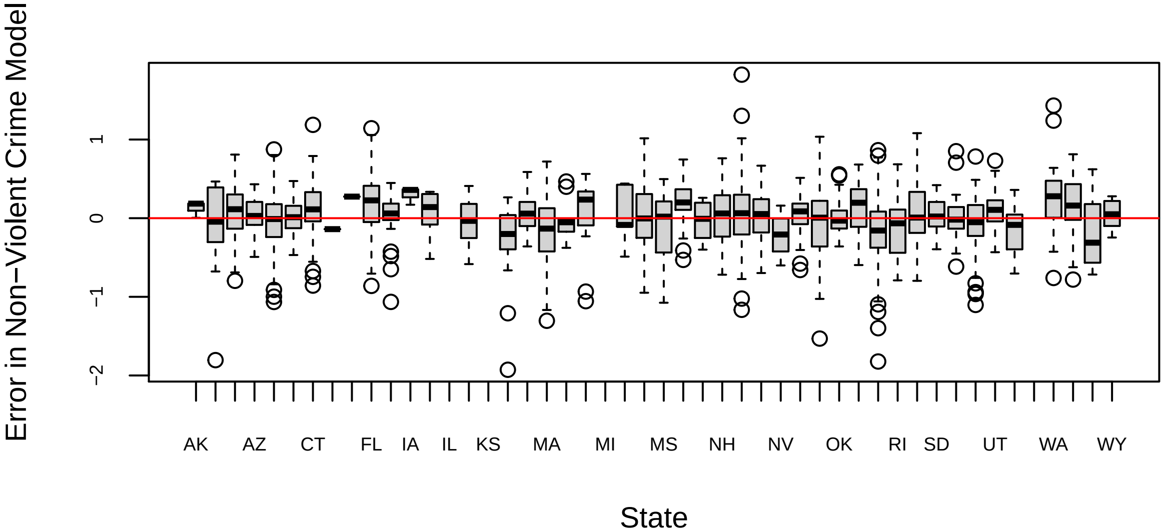
As we discussed in our findings, after computing the difference between actual log(`violentPerPop`) with the predicted values, we create a box plot (Figure F) of the errors for each state which are distributed mostly around zero, indicating the model predicts the violent crime rates for those states well. However, North Dakota does not seem to conform with the general pattern of the model, consistently having actual violent crime rates lower than the predicted violent crime rates.

Figure F



For non-violent crime rates on the other hand, the difference between actual and predicted values for all states does appear to be distributed around zero, so we can conclude that all states conform to the general pattern. (Figure G)

Figure G



Appendix

Packages

```
library(dplyr)
library(e1071)
library(RcmdrMisc)
library(glmnet)
load("~/Downloads/USACrime.Rda")
```

Appendix 1 - Removing unexpected values and NA values

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

Appendix 2 - Changing Pacific region to West

```
levels(USACrime$region) = c("MidWest", "NorthEast", "West", "South", "West")
```

Appendix 3 - Changing the pctUrban variable to factor

```
USACrime = transform(USACrime, pctUrban = +(pctUrban > 85))
names(USACrime)[3] <- "isUrban"
```

Appendix 4 - Finding variables that need to be transformed

```
vars<- c(c(4:22))
posVars <- c()
negVars <- c()
for (i in vars) {
  if(skewness(USACrime[,i]) > 0.7){
    posVars <- c(posVars, i)
  }
  if(skewness(USACrime[,i]) < -0.7){
    negVars <- c(negVars, i)
  }
}
```

Appendix 5 - Finding best transformations for positively skewed variables

```
posVars2 <- posVars[posVars != 12]
posVars2 <- posVars2[posVars2 != 15]
for (i in posVars2){
  a <- abs(skewness(USACrime[,i]))
  b <- abs(skewness(log(USACrime[,i])))
  c <- abs(skewness(sqrt(USACrime[,i])))
  z <- NA
  y <- min(a,b,c)
  if (y == a){
    z <- "normal"
  }
  else if(y == b){
    z <- "log"
  }
  else{
    z <- "sqrt"
  }
  print(paste(colnames(USACrime[i]),z,min(a,b,c)))
}
```

Appendix 6 - Transformations

```
USACrime2 = transform(USACrime,
                      medIncome = log(medIncome),
                      pctLowEdu = log(pctLowEdu),
                      pctNotHSgrad = pctNotHSgrad^0.5,
                      pctCollGrad = log(pctCollGrad),
                      pctUnemploy = log(pctUnemploy),
                      pctKidsBornNevrMarr = pctKidsBornNevrMarr^0.25,
                      pctHousOccup = exp(pctHousOccup/10),
                      pctVacantBoarded = log(pctVacantBoarded+1),
                      ownHousMed = log(ownHousMed),
                      ownHousQrange = log(ownHousQrange),
                      rentMed = log(rentMed),
                      rentQrange = log(rentQrange),
                      popDensity = log(popDensity),
                      pctForeignBorn = log(pctForeignBorn),
                      violentPerPop = log(violentPerPop),
                      nonViolPerPop = log(nonViolPerPop))
```

Appendix 7 - RSS minimization for EDA Variables

```
RSSVMin <- Inf
modelDataVMinRSS <- NULL
for (i in 1:2) {
  for(j in 1:2){
    for (k in 1:2) {
      for (l in 1:2) {
        modelData <- subset(USACrime2, select =
                    c(region,isUrban,pctNotHSgrad,pctWdiv,medIncome,rentQrange,
                    c(pctUnemploy, pctEmploy)[i],
                    c(pctKids2Par,pctKidsBornNevrMarr)[j],
                    c(pctHousOccup, pctHousOwnerOccup)[k],
                    c(ownHousMed, rentMed)[l]))
        lmA <- lm(USACrime2$violentPerPop ~ ., modelData)
        if (sum(residuals(lmA)^2) < RSSVMin) {
          RSSVMin <- sum(residuals(lmA)^2)
          modelDataVMinRSS <- modelData
        }
      }
    }
  }
}
lmVMinRSS <- lm(USACrime2$violentPerPop ~ ., modelDataVMinRSS)
RSSNVMin <- Inf
modelDataNVMinRSS <- NULL
for (i in 1:2) {
  for(j in 1:2){
    for (k in 1:2) {
      for (l in 1:2) {
        modelData <- subset(USACrime2, select =
                    c(region,isUrban,pctNotHSgrad,pctWdiv,medIncome,rentQrange,
                    c(pctUnemploy, pctEmploy)[i],
                    c(pctKids2Par,pctKidsBornNevrMarr)[j],
                    c(pctHousOccup, pctHousOwnerOccup)[k],
                    c(ownHousMed, rentMed)[l]))
        lmA <- lm(USACrime2$nonViolPerPop ~ ., modelData)
        if (sum(residuals(lmA)^2) < RSSNVMin) {
          RSSNVMin <- sum(residuals(lmA)^2)
          modelDataNVMinRSS <- modelData
        }
      }
    }
  }
}
```

```
lmNVMinRSS <- lm(USACrime2$nonViolPerPop ~ ., modelDataVMinRSS)
```

Appendix 8 - Stepwise regression with AIC

```
#REMOVING STATES AND VIOL/NONVIOL
USACrime2V = USACrime2[,c(-1,-24)]
USACrime2NV = USACrime2[,c(-1,-23)]
#VIOL AIC
violentModelAICMax = stepwise(lm(violentPerPop ~ ., data = USACrime2V),
        direction = "backward/forward", criterion = "AIC", trace = FALSE)
violentModelAICMin = stepwise(lm(violentPerPop ~ ., data = USACrime2V),
        direction = "forward/backward", criterion = "AIC", trace = FALSE)
#NONVIOL AIC
nonViolModelAICMax = stepwise(lm(nonViolPerPop ~ ., data = USACrime2NV),
        direction = "backward/forward", criterion = "AIC", trace = FALSE)
nonViolModelAICMin = stepwise(lm(nonViolPerPop ~ ., data = USACrime2NV),
        direction = "forward/backward", criterion = "AIC", trace = FALSE)
```

Appendix 9 - LASSO Regression (Coefficient Shrinkage)

```
set.seed(1234)
X1 <- model.matrix(lm(violentPerPop ~ 1 + ., data = USACrime2[-c(1,24)]))
X2 <- model.matrix(lm(nonViolPerPop ~ 1 + ., data = USACrime2[-c(1,23)]))
y1 <- log(USACrime$violentPerPop)
y2 <- log(USACrime$nonViolPerPop)
fitA <- glmnet(X1, y1, alpha = 1)
fitB <- glmnet(X2, y2, alpha = 1)
{plot(fitA, xvar = "lambda", label = TRUE,
      sub = "LASSO Regression on Violent Crime", cex.main = .7)
plot(fitB, xvar = "lambda", label = TRUE,
     sub = "LASSO Regression on Non-Violent Crime", cex.main = .7)}
```

Appendix 10 - LASSO Regression (Find optimal lambda and get coefficients)

```
set.seed(1234)
model5A <- cv.glmnet(X1, y1, alpha = 1)
model5B <- cv.glmnet(X2, y2, alpha = 1)
{par(mfrow = c(1, 2))
plot(model5A, sub = "Violent Crime")
plot(model5B, sub = "Non-Violent Crime")}
coef(model5A)
coef(model5B)
```

Appendix 11 - Linear models AIC using LASSO regression variables

```
model6A <- lm(violentPerPop ~ region + isUrban
              + pctWdiv + pctKids2Par + pctKidsBornNevrMarr
              + pctHousOccup + pctVacantBoarded + pctForeignBorn
              + ownHousQrange, data = USACrime2)
model6B <- lm(nonViolPerPop ~ region + isUrban + pctKids2Par
              + pctKidsBornNevrMarr + pctHousOccup + pctHousOwnerOccup
              + pctVacantBoarded + pctVacant6up
              + pctForeignBorn + rentQrange, data = USACrime2)
AIC(model6A)
AIC(model6B)
```

Appendix 12 - MSE of the models from step-wise regression with AIC and LASSO

```
paste("violentModelAIC MSE =",mean(violentModelAICMin$residuals^2))
paste("nonViolModelAIC MSE =",mean(nonViolModelAICMin$residuals^2))
paste("violentModelLASSO MSE =",model5A$cvm[model5A$lambda == model5A$lambda.1se])
paste("nonViolModelLASSO MSE =",model5B$cvm[model5B$lambda == model5B$lambda.1se])
```

Appendix 13 - Outliers, Multicollinearity and Ridge Regression for violent crime model

```
plot(violentModelAICMin, which = 3, sub.caption = "")
ols_vif_tol(violentModelAICMin)
set.seed(3562)
violentModelRidge = cv.glmnet(model.matrix(violentModelAICMin),
                          USACrime2V$violentPerPop, alpha = 0)
coef(violentModelRidge)
```

Appendix 14 - Outliers, Multicollinearity and Ridge Regression for non violent crime model

```
plot(nonViolModelAICMin, which = 3), sub.caption = ""
USACrime2NV = USACrime2NV[c(-320,-363,-1548),]
nonViolModelAICMin = lm(nonViolPerPop ~ region + pctKids2Par
 + pctHousOccup + pctForeignBorn + rentQrange + pctEmploy
 + pctCollGrad + ownHousMed + pctVacant6up + pctLowEdu
 + isUrban + pctWdiv + pctVacantBoarded, data = USACrime2NV)
plot(nonViolModelAICMin, which = 3, sub.caption = "")
ols_vif_tol(nonViolModelAICMin)
set.seed(6375)
nonViolModelRidge = cv.glmnet(model.matrix(nonViolModelAICMin),
                         USACrime2NV$nonViolPerPop, alpha = 0)
coef(violentModelRidge)
```

Appendix 15 - Box-plot of errors across states (Violent Crime) (Figure F)

```
USACrime2$violentError = USACrime2$violentPerPop -
  predict(violentModelRidge, newx = model.matrix(violentModelAICMin))
boxplot(USACrime2$violentError ~ USACrime2$State, cex.axis = 0.6, cex.lab = 0.9)
abline(h=0, col = "red")
```

Appendix 16 - Box-plot of errors across states (Non-Violent Crime) (Figure G)

```
USACrime3 = USACrime2[c(-320,-363,-1548),]
USACrime3$nonViolError = USACrime3$nonViolPerPop -
  predict(nonViolModelRidge, newx = model.matrix(nonViolModelAICMin))
boxplot(USACrime3$nonViolError ~ USACrime3$State, cex.axis = 0.6, cex.lab = 0.9)
abline(h=0, col = "red")
```

Mark Allocations

- 1806987 - 103%
- 1800673 - 104%
- 1802212 - 103%
- 1824442 - 90%