Relationship between crime rate and demographic markers in North Carolina

Debalina, Mark, Tina & Vivek

August 1, 2018

Introduction

This report discusses determinants of crime in counties across North Carolina. The data within was obtained through our political campaign partnership and contains geographic, demographic, economic, and crime data. The data was collected by the campaign in 1987 and is reported on a county level basis. The goal is to leverage the data to provide policy suggestions that are applicable to local government to reduce crime based on the available data we have. We will explore the drivers of the crime rate in terms of the number of crimes committed per person.

We believe there are **2** variables of interest to our political partners:

A. Overall crime rate as captured by the **crmrte** variable.
B. Face to face crime rate as captured by a combination of the **crmrte** and **mix** variables. We posit that “face to face”" crime is more concerning to the general public than other varieties. “Mix” is defined as the ratio of face-to-face crime to other crime:

$$mix = \frac{face - to - face}{other}$$

Thus, the face-to-face crime rate can be simplified to:

$$face - to - face\ crime\ rate = crime\ rate * \frac{mix}{mix + 1}$$

We will seek to make recommendations to our political partners regarding reducing crime rates, discuss limitations of this observational study, caution against unintended consequences of societal changes, and suggest future areas of study.

For each variable of interest we will do 3 models, first based on intuition and raw coorelations, second based on rigorous stepwise regression, minimizing Adjusted R-square and BIC, and third by taking all variables.

Exploratory Data Analysis

Data cleaning

We performed the following cleaning operations to the data:

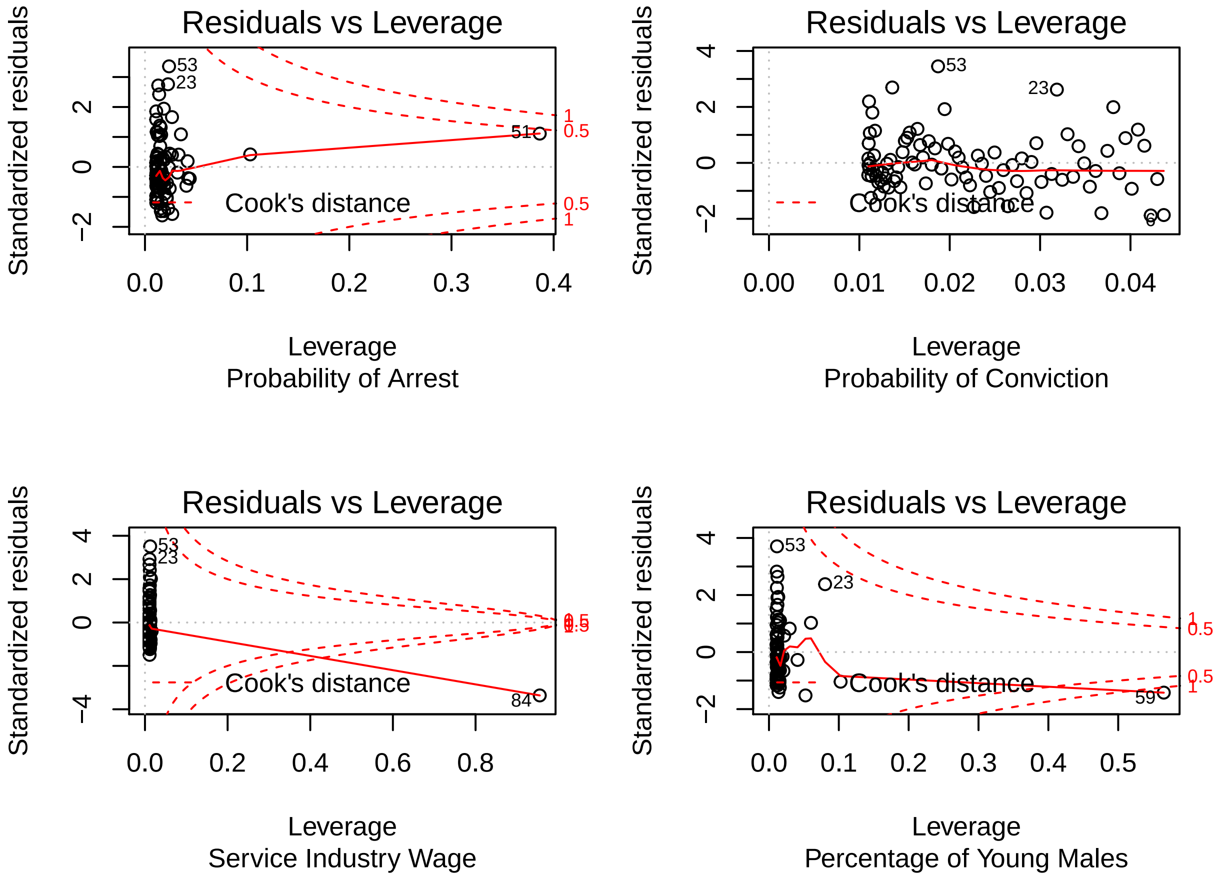
- Removed the ‘year’ attribute as it has no variation
- Removed final 6 rows of NAs as they contain no information
- Certain numeric attributes were imported as strings, so those were converted to numeric data

```r
# Cleaning
df$year <- NULL
df <- df[!is.na(df$county), ]
df <- data.frame(sapply(df, as.numeric))
```

Discussion of Outliers

We noticed the following outliers during EDA and attribute the following explanations and/or rectifications:

```r
modelprbarr <- lm(crmrte ~ prbarr, df)
modelprbconv <- lm(crmrte ~ prbconv, df)
modelwser <- lm(crmrte ~ wser, df)
modelpctymle <- lm(crmrte ~ pctymle, df)

par(mfrow = c(2, 2))
plot(modelprbarr, which = 5)
title(sub = "Probability of Arrest")
plot(modelprbconv, which = 5)
title(sub = "Probability of Conviction")
plot(modelwser, which = 5)
title(sub = "Service Industry Wage")
plot(modelpctymle, which = 5)
title(sub = "Percentage of Young Males")
```



a. Probability > 1 in prbconv and in prbarr. We believe it is plausible that there were more arrests than offenses in a given year, so we chose not to modify these data. Above residual vs leverage plots also confirm that these outliers are not that influencial to be removed from our data set.

b. One of the wser value is approximately 10 times the average in one of the counties. Also from the above plot we see that cook's diatance is way bigger than 1. We believe this to be an error, given that it is not an individual's wage but an average wage for a county. It may have been off by an order of 10 due to a typo. We chose to impute this data point with the mean service wage as follows:

```r
df$wser[which(df$wser == max(df$wser))] <- mean(df$wser)
```

c. pctymle is exceptionally high for one of the counties. We believe this is plausible, if it for example contains a large university, so we chose not to alter the outlier. This variable does show very heavy skew, so we will transform it later which will reduce its influence.

Transformations

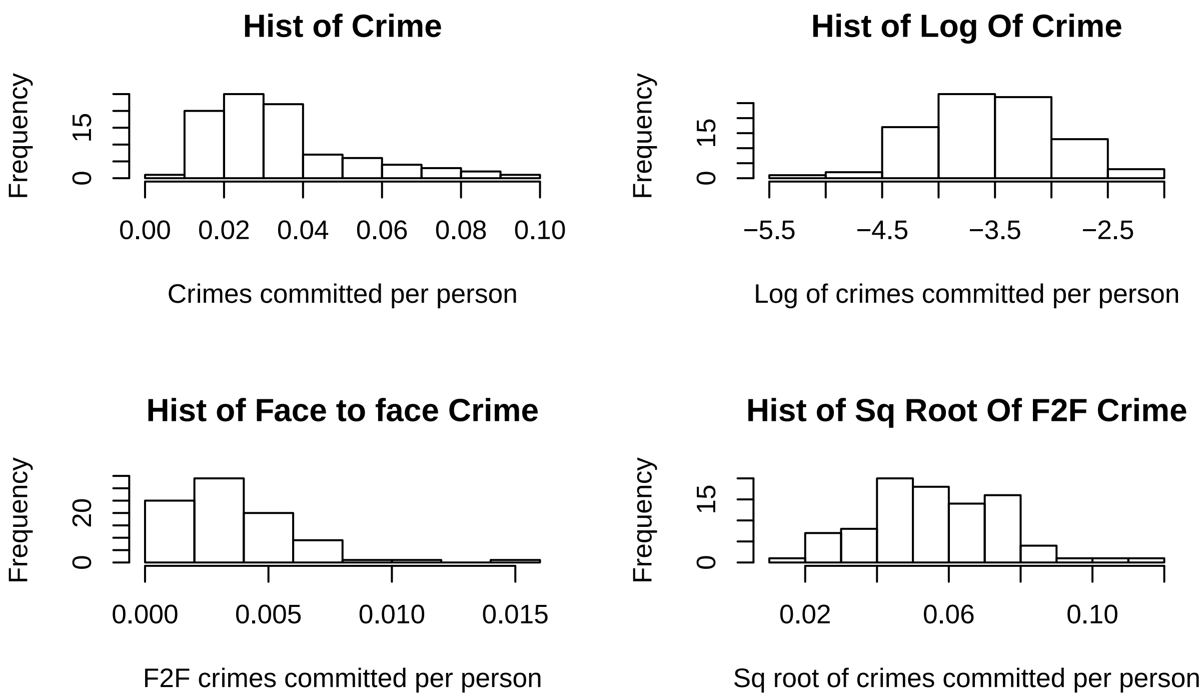
We adhered to the following principles in determining what transformations were appropriate:

1. It is desired that the outcome variable is normally distributed. This increases the chance that we will get normally-distributed errors.
2. It is not necessary that independent variables be normally distributed, but for highly skewed data, it is often beneficial to perform a transformation so as to have a better spread of data across the estimation interval.

Both crime rate and face-to-face crime rates exhibit right skew. We see that log of crime rate variable brings the distribution closer to normal, however log transformation of face to face crime rate makes it skewed in the other direction. Taking square root gives a distribution more appropriate for modelling.

We also noticed that log transformations of police per capita, tax revenue per capita, probability of arrest, and probability of conviction helped to reduce heavy skew.

```r
par(mfrow = c(2, 2))
hist(df$crmrte, main = "Hist of Crime", xlab = "Crimes committed per person")
hist(df$logcrmrte, main = "Hist of Log Of Crime", xlab = "Log of crimes committed per person")
hist(df$fcrmrte, main = "Hist of Face to face Crime", xlab = "F2F crimes committed per person")
hist(df$sqrtfcrmrte, main = "Hist of Sq Root Of F2F Crime", xlab = "Sq root of crimes committed per pers
```

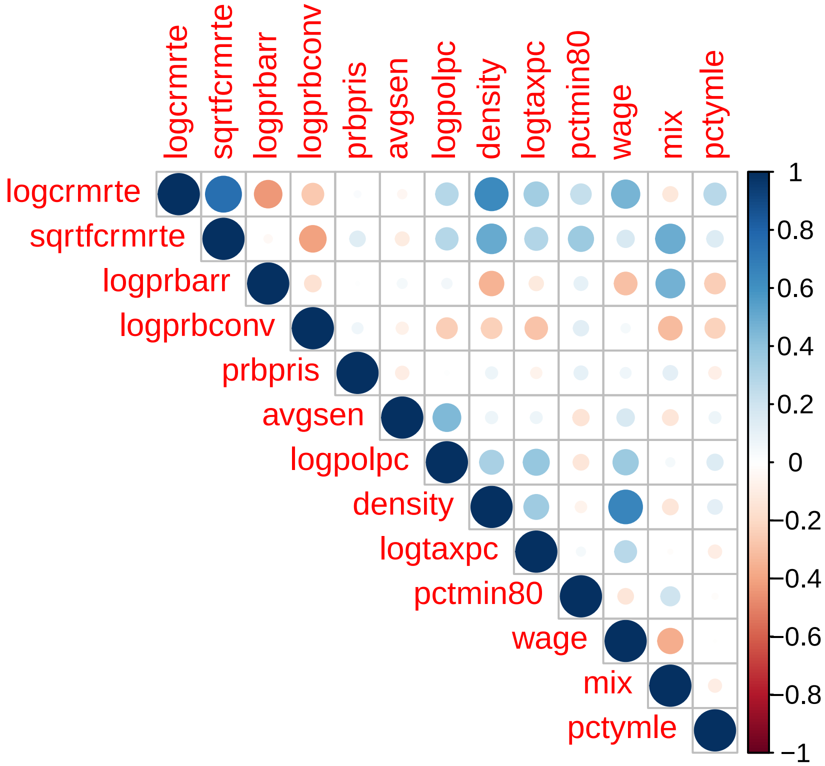


Wage variables show a high degree of multicollinearity. We combined them using a simple mean for now but an additional improvement could be to weight them by sector employment. For clarity, we also coalesced various regions into a **region** factor. Note that no information or degrees of freedom were lost in this process, but it improved clarity during exploration.

```
df$wage <- (df$wcon + df$wtuc + df$wtrd + df$wfir + df$wser + df$wmfg + df$wfed +
    df$wsta + df$wloc)/9
df$region <- factor(ifelse(df$west == 1, "west", ifelse(df$central == 1, "central",
    "other")))

# Misc. transformations
df$urban <- factor(ifelse(df$urban == 1, "urban", "not urban"))
```

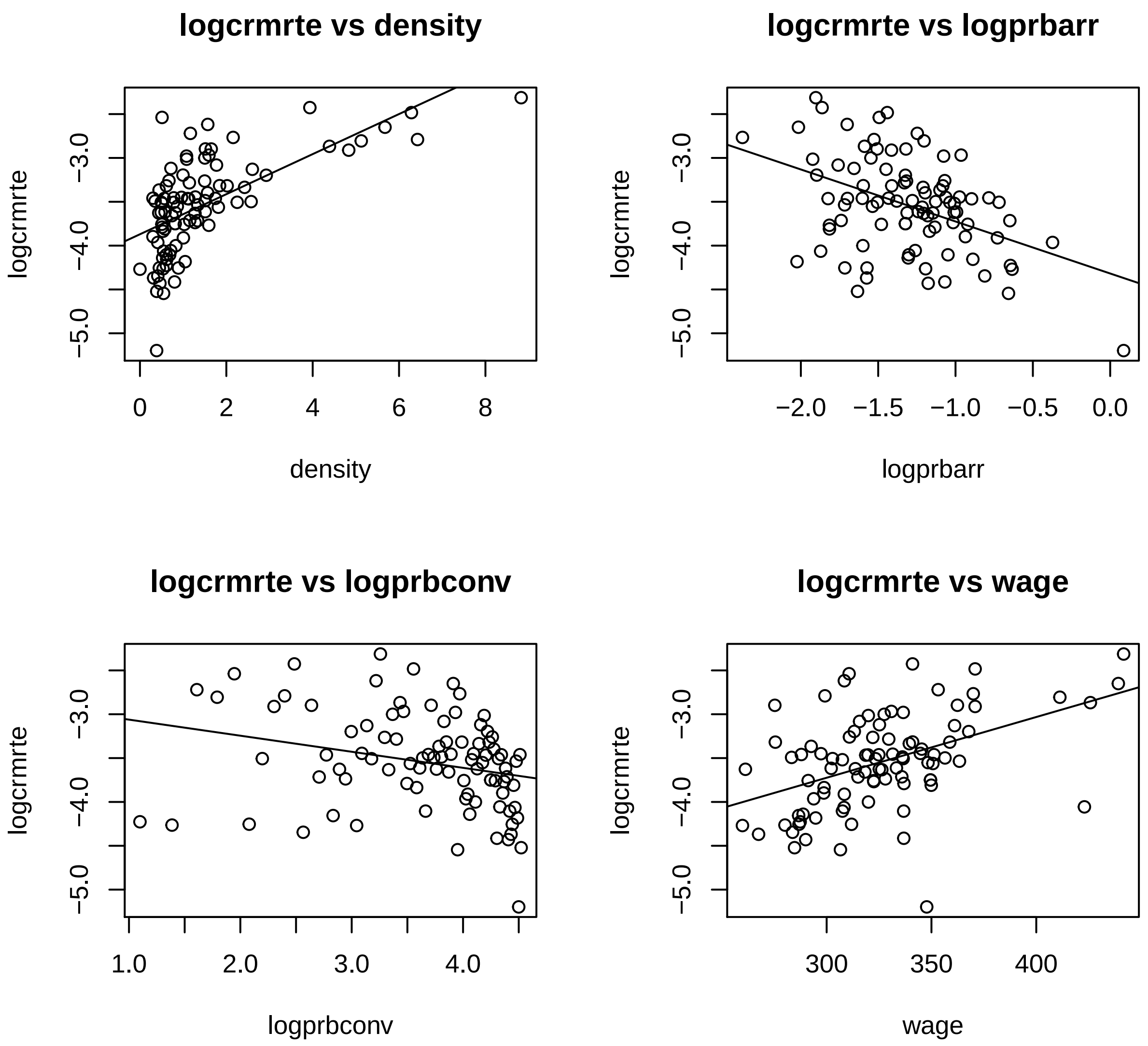
Checking pairwise correlation helps identify some preliminary covariates of interest.

```
corrplot(cor(df[, c("logcrmrte", "sqrtfcrmrte", "logprbarr", "logprbconv", "prbpris",
    "avgsen", "logpolpc", "density", "logtaxpc", "pctmin80", "wage", "mix", "pctymle")]),
    method = "circle", type = "upper")
```



Looking at the correlations we identified the covariates that show high correlation with crime rate, leaving aside those that seem to be collinear with other covariates. For example both density and urban have a high correlation with crime rate but they seem to be also highly correlated with themselves.

```
par(mfrow = c(2, 2))
plot(df$density, df$logcrmrte, main = "logcrmrte vs density", xlab = "density", ylab = "logcrmrte")
abline(lm(df$logcrmrte ~ df$density))
plot(df$logprbarr, df$logcrmrte, main = "logcrmrte vs logprbarr", xlab = "logprbarr",
    ylab = "logcrmrte")
abline(lm(df$logcrmrte ~ df$logprbarr))
plot(df$logprbconv, df$logcrmrte, main = "logcrmrte vs logprbconv", xlab = "logprbconv",
    ylab = "logcrmrte")
abline(lm(df$logcrmrte ~ df$logprbconv))
plot(df$wage, df$logcrmrte, main = "logcrmrte vs wage", xlab = "wage", ylab = "logcrmrte")
abline(lm(df$logcrmrte ~ df$wage))
```



Using this argument, our first basic specification uses the following covariates:

- **density** which is strongly positively correlated with crime rate.
- **prbarr** which is negatively correlated with crime rate.
- **prbconv** which is negatively correlated with crime rate.
- **wages** which is average wage and positively correlated with crime rate.

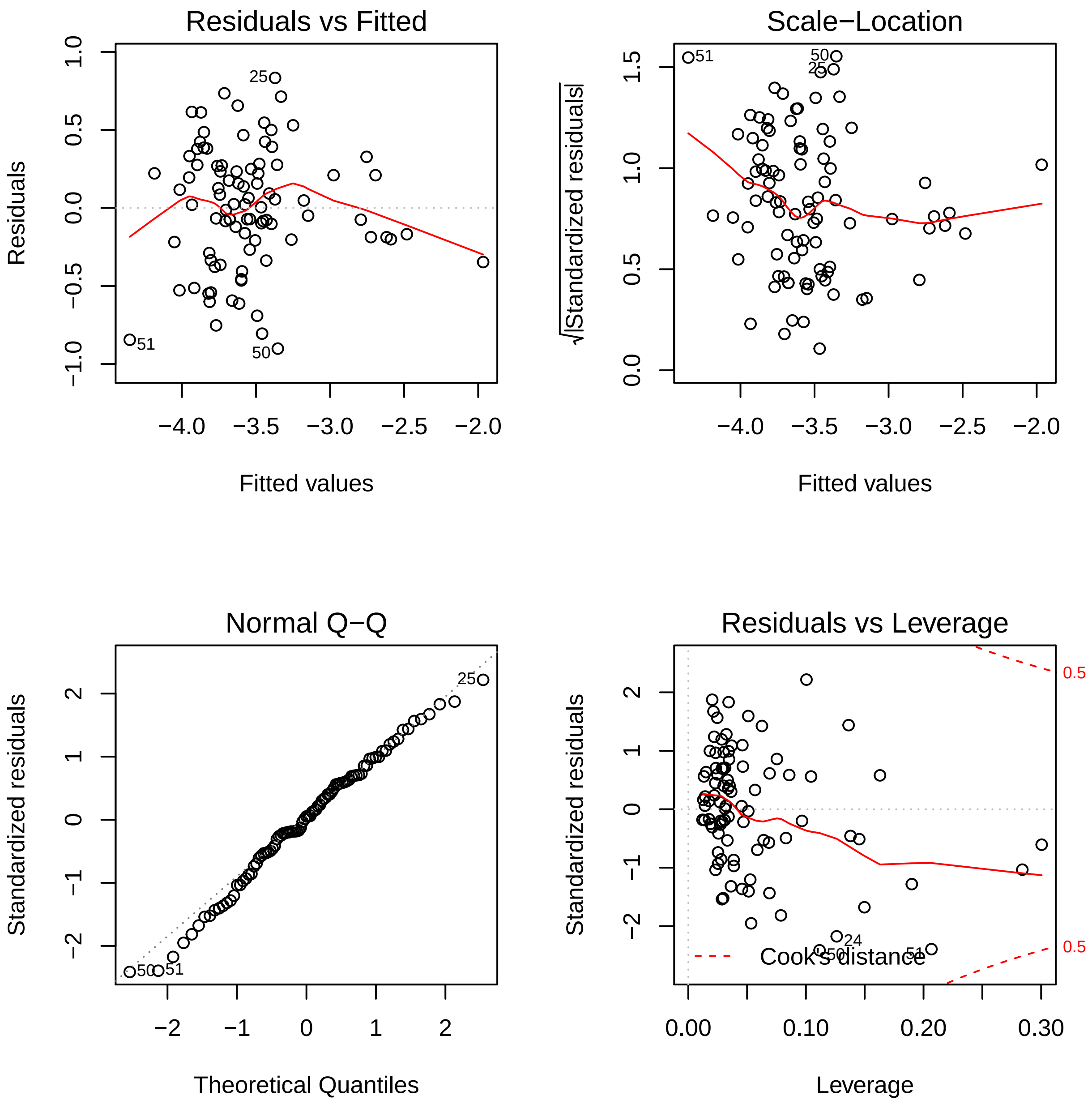
Model Specifications and Assumptions

In our exploratory analysis, we identified key independent variables that were positively and negatively correlated with log of crime rate. To create our first and simplest model that contains variables of key interest that we hypthesized might be the most important determinants of crime. We created three specifications for each of our two response variables of interest (6 total). The specifications iteratively add complexity, representing the tradeoff between accuracy and conciseness.

Crime Rate

Model 1

```
model1df <- df %>% select(logcrmrte, density, logprbarr, logprbconv, wage)

model1 <- lm(formula = logcrmrte ~ density + logprbarr + wage + logprbconv, data = model1df)

layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(model1)
```

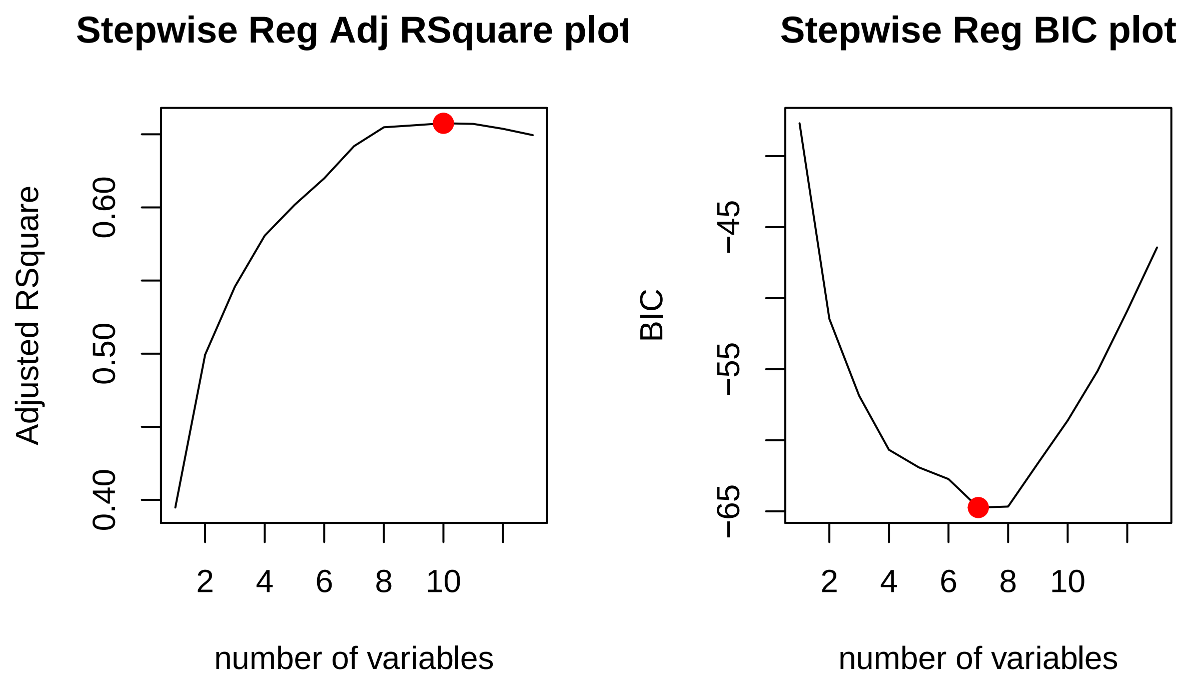


Model 2

For our second specification, we utilize stepwise regression to determine the input variables

```
modeldf <- df %>% select(logcrmrte, logprbarr, logprbconv, prbpris, avgsen, logpolpc,
    density, logtaxpc, urban, pctmin80, wage, region, mix, pctymle)

stepmodel = regsubsets(logcrmrte ~ ., modeldf, nvmax = 13)

stepsummary <- summary(stepmodel)
```

```
# Where is the highest Adjusted RSquare and minimum BIC
par(mfrow = c(1, 2))
plot(stepsummary$adjr2, ylab = "Adjusted RSquare", xlab = "number of variables",
    main = "Stepwise Reg Adj RSquare plot", type = "l")

points(which.max(stepsummary$adjr2), stepsummary$adjr2[which.max(stepsummary$adjr2)],
    col = "red", cex = 2, pch = 20)

plot(stepsummary$bic, ylab = "BIC", xlab = "number of variables", main = "Stepwise Reg BIC plot",
    type = "l")

points(which.min(stepsummary$bic), stepsummary$bic[which.min(stepsummary$bic)], col = "red",
    cex = 2, pch = 20)
```



Both graphs show rapid improvement in explainability for first 7 variables before flattening off. We use the best 7 variables as recommended by stepwise regression.

```
stepsummary$outmat[1:7, ]
```

```
##           logprbarr logprbconv prbpris avgsen logpolpc density logtaxpc
## 1  ( 1 ) " "       " "        " "     " "    " "      "*"     " "
## 2  ( 1 ) " "       " "        " "     " "    " "      "*"     " "
## 3  ( 1 ) "*"       " "        " "     " "    " "      "*"     " "
## 4  ( 1 ) "*"       "*"        " "     " "    " "      "*"     " "
## 5  ( 1 ) "*"       "*"        " "     " "    " "      "*"     " "
## 6  ( 1 ) "*"       " "        " "     "*"    "*"      "*"     " "
## 7  ( 1 ) "*"       "*"        " "     "*"    "*"      "*"     " "
##          urbanurban pctmin80 wage regionother regionwest mix pctymle
## 1  ( 1 ) " "        " "      " "  " "         " "        " " " "
## 2  ( 1 ) " "        " "      " "  "*"         " "        " " " "
## 3  ( 1 ) " "        " "      " "  "*"         " "        " " " "
## 4  ( 1 ) " "        "*"      " "  " "         " "        " " " "
## 5  ( 1 ) " "        " "      " "  "*"         "*"        " " " "
## 6  ( 1 ) " "        " "      " "  "*"         "*"        " " " "
## 7  ( 1 ) " "        " "      " "  "*"         "*"        " " " "
```

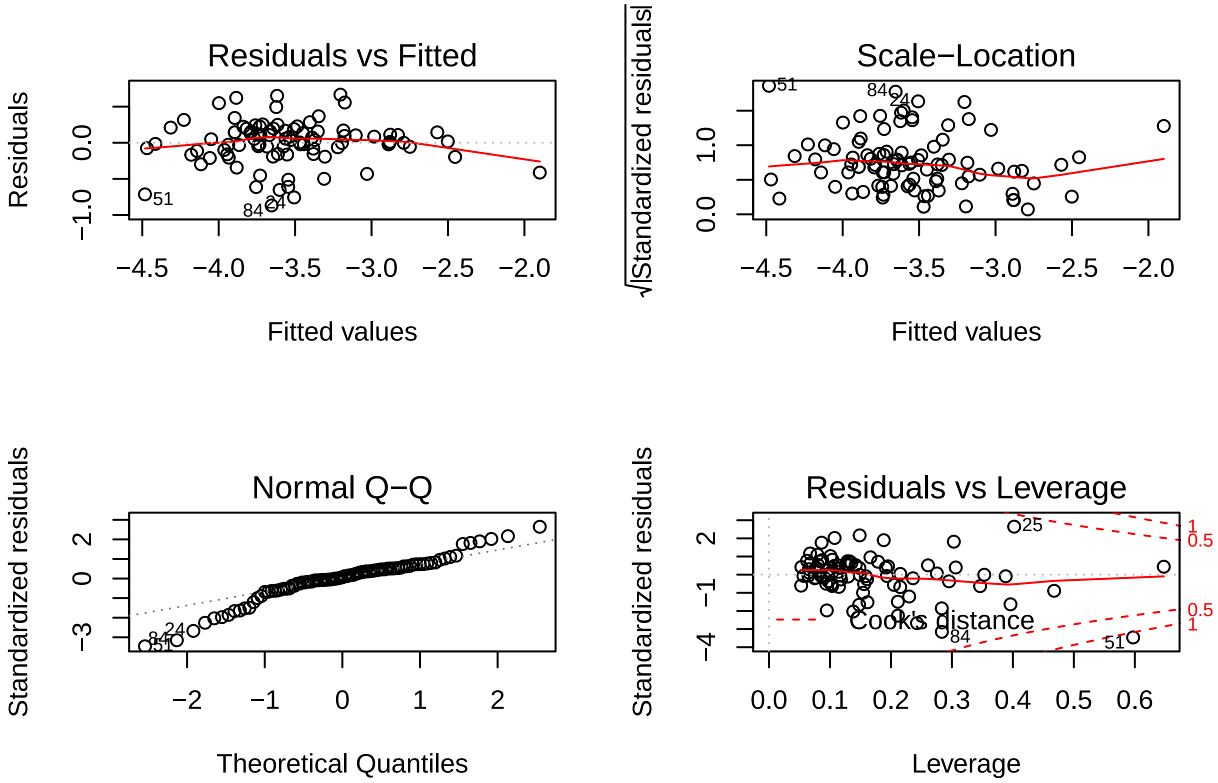
The most effective explanatory variables are: density, region, logprbconv, logprbarr, logpolpc, pctmin80, wage

```
model2df <- df %>% select(logcrmrte, density, logprbarr, logprbconv, wage, region,
    logtaxpc, pctymle)

model2 <- lm(formula = logcrmrte ~ density + logprbarr + wage + logprbconv + region +
    logtaxpc + pctymle, data = model2df)
```

Model 3

This includes all the parameters

```
model3 <- lm(formula = logcrmrte ~ logprbarr + logprbconv + prbpris + avgsen + logpolpc +
    density + logtaxpc + urban + pctmin80 + wage + region + mix + pctymle, data = modeldf)

layout(matrix(c(1, 2, 3, 4), 2, 2))

plot(model3)
```



Aside on Classical Linear Model Assumptions

While model building, we ensured to verify that our regression models satisfied all necessary assumptions. As an aside, we chose our second specification (as it is our best attempt at optimization) for a deep dive into how we tested the assumptions and addressed any concerns.

a. Linear population model

This assumption does not require testing because we have defined our model as linear in beta coefficients and have not constrained the error term.

b. Random Sampling

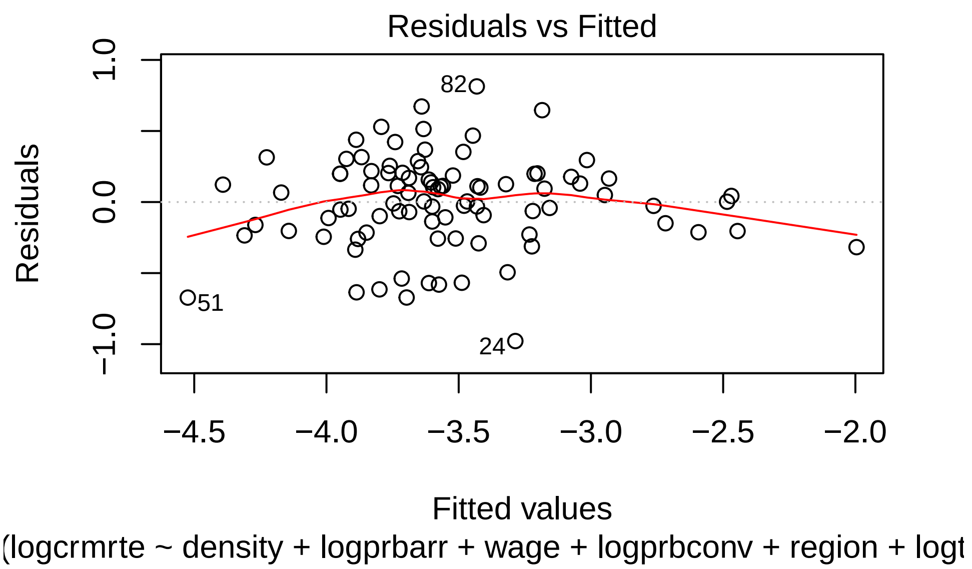
The data set includes 91 counties. A Google search suggests North Carolina has 100 counties, so this sample seems to encompass roughly all counties in the state. We will have to take on faith that the characteristics of each county were developed from a random sampling or otherwise representative data collection process.

c. No perfect multicollinearity

Correlation figure on page 4 show at a glance the collinearity among variables. There are a few noticeable “hot spots” on the matrix. While determining our specifications, we took these collinearities into account, selecting subsets that provided correlation to output variable while not imposing significant collinearity on input variables. We believe the amount of remaining collinearity is consumable for model building.

d. Zero-conditional mean

```r
plot(model2, which = 1)
```

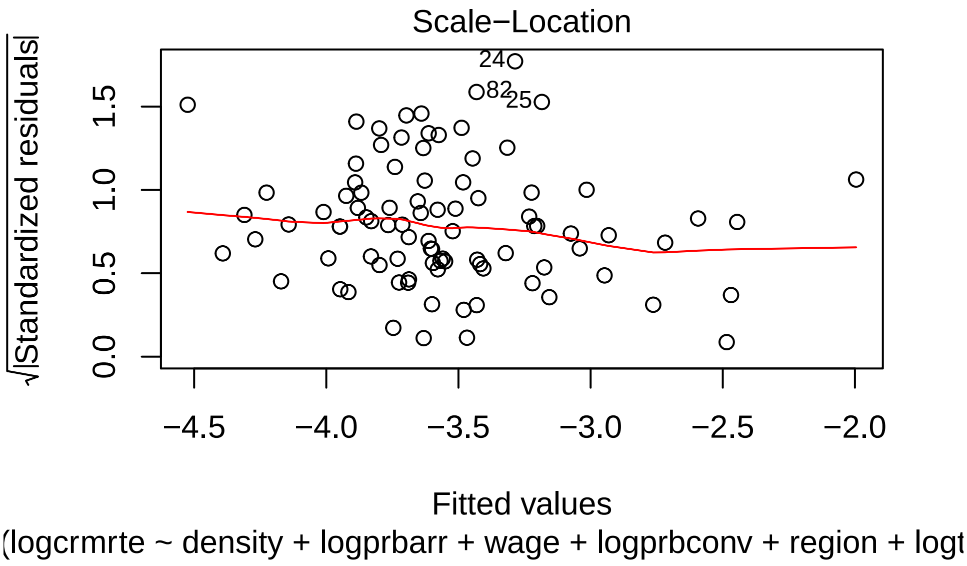


The model seems to satisfy zero conditional mean to a reasonable degree, despite the small perturbations. Additionally, we have a relatively large sample and would not expect coefficients to be biased as a result of these factors.

e. Homoskedasticity

The errors appear to be slightly heteroskedastic. This is confirmed by checking the scale-location plot, which has small deviations from a horizontal trend.

```r
plot(model2, which = 3)
```



We can further test with the studentized Breusch-Pagan test for Heteroskedasticity.

```
bptest(model2)
```

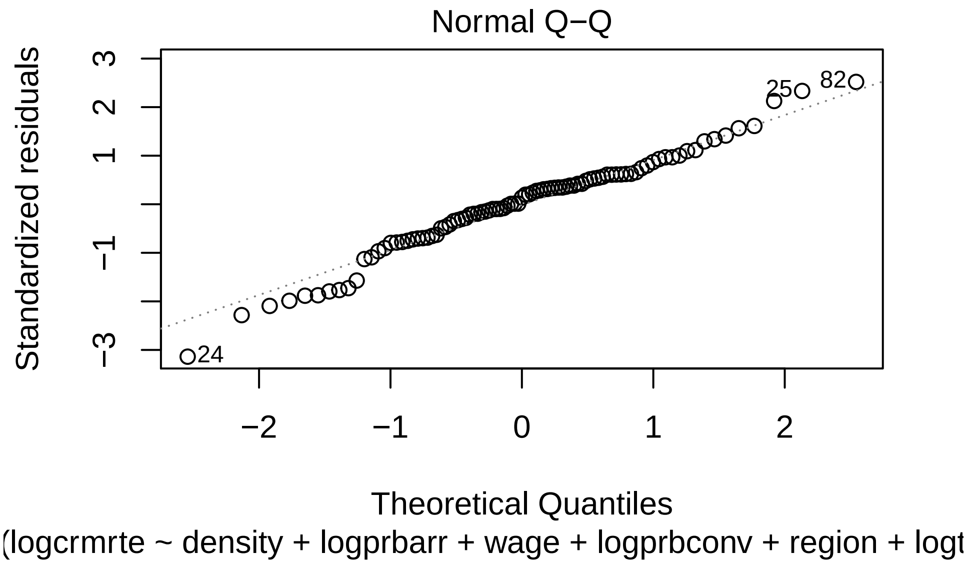
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 16.018, df = 8, p-value = 0.04212
```

Since the p-value is significant at the level of 0.01, we reject the homoskedasticity null hypothesis. Note that this is expected with a relatively large sample size. All things considered, it would be a good idea to use heteroskedasticity-consistent standard errors.

f. Normality of Errors

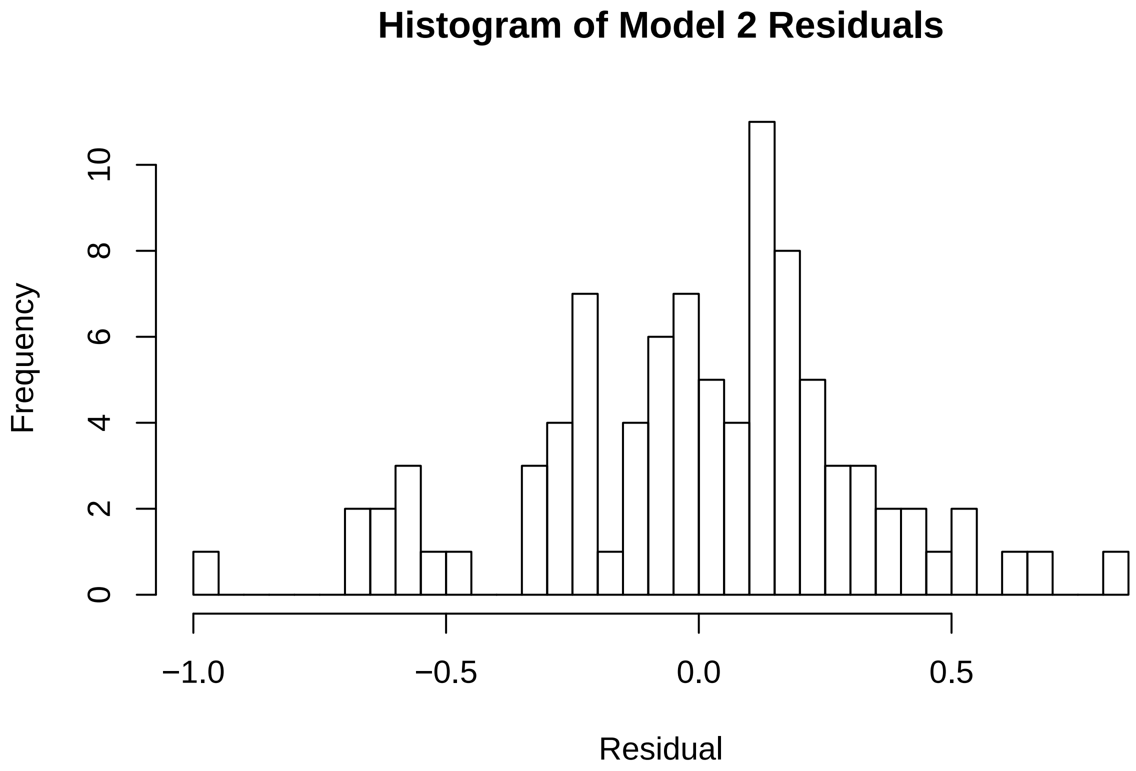
The easiest way to assess normality of errors is through a Normal Q-Q plot:

```
plot(model2, which = 2)
```



The plot looks roughly linear except at the extremes of our prediction inverval. This may indicate that we aren't fitting our data as well at the extreme ends of our model. However, the sample size is large enough to invoke the central limit theorem for slightly skewed residuals, so our estimators will still tend toward a normal distribution.

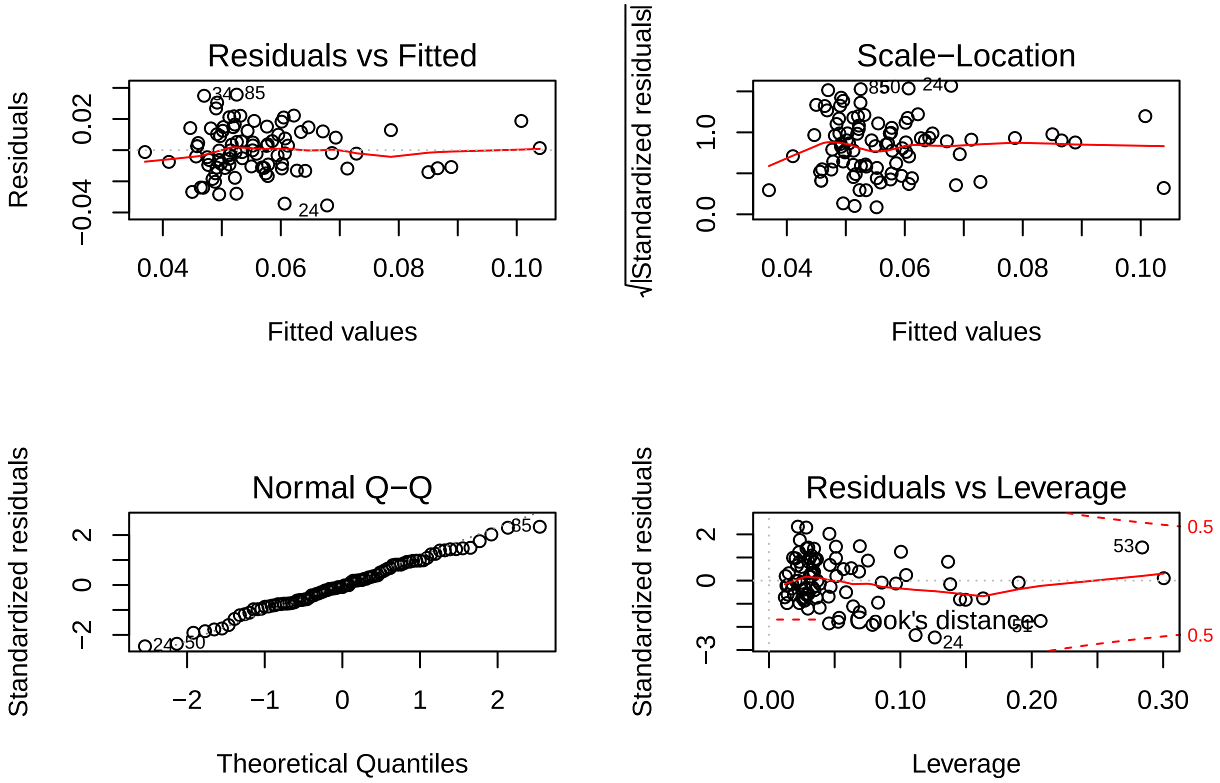
```r
hist(model2$residuals, breaks = 50, main = "Histogram of Model 2 Residuals", xlab = "Residual")
```



Face-to-face Crime Rate

Model 1

```
model1.1df <- df %>% select(sqrtfcrmrte, density, logprbarr, logprbconv, wage)

model1.1 <- lm(formula = sqrtfcrmrte ~ density + logprbarr + wage + logprbconv, data = model1.1df)

layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(model1.1)
```

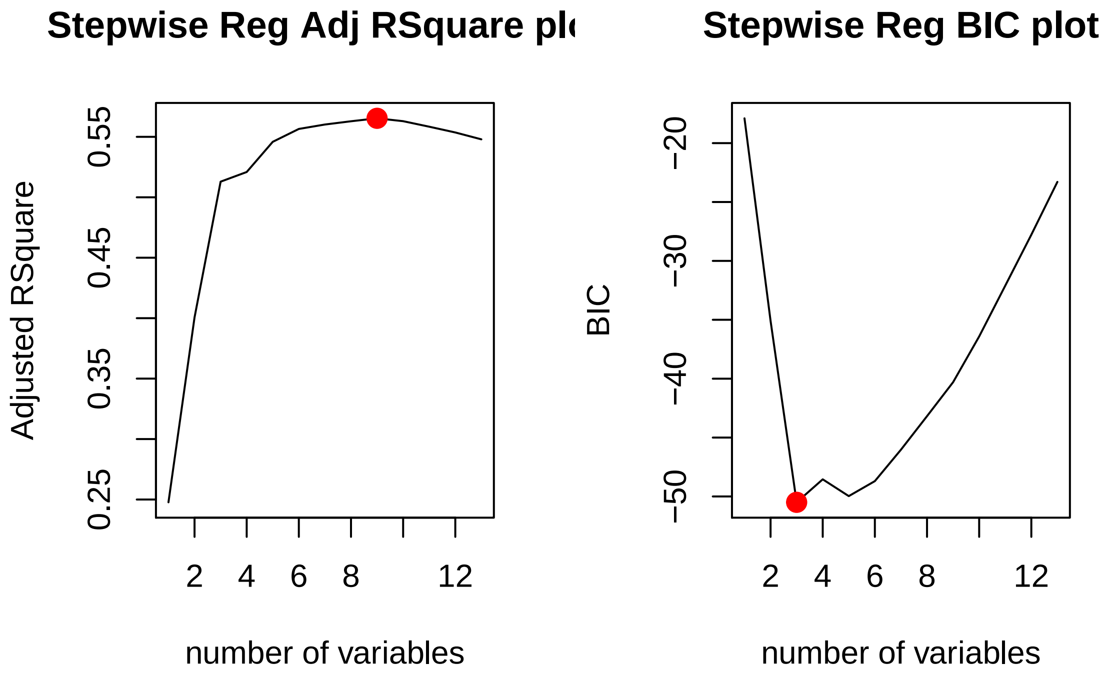


####Model 2

For our second specification, we utilize stepwise regression to determine the input variables

```
fmodeldf <- df %>% select(sqrtfcrmrte, logprbarr, logprbconv, prbpris, avgsen, logpolpc,
    density, logtaxpc, urban, pctmin80, wage, region, pctymle)

fstepmodel = regsubsets(sqrtfcrmrte ~ ., fmodeldf, nvmax = 13)

fstepsummary <- summary(fstepmodel)
```

```
# Where is the highest Adjusted RSquare and minimum BIC
par(mfrow = c(1, 2))
plot(fstepsummary$adjr2, ylab = "Adjusted RSquare", xlab = "number of variables",
    main = "Stepwise Reg Adj RSquare plot", type = "l")
points(which.max(fstepsummary$adjr2), fstepsummary$adjr2[which.max(fstepsummary$adjr2)],
    col = "red", cex = 2, pch = 20)
plot(fstepsummary$bic, ylab = "BIC", xlab = "number of variables", main = "Stepwise Reg BIC plot",
    type = "l")
points(which.min(fstepsummary$bic), fstepsummary$bic[which.min(fstepsummary$bic)],
    col = "red", cex = 2, pch = 20)
```



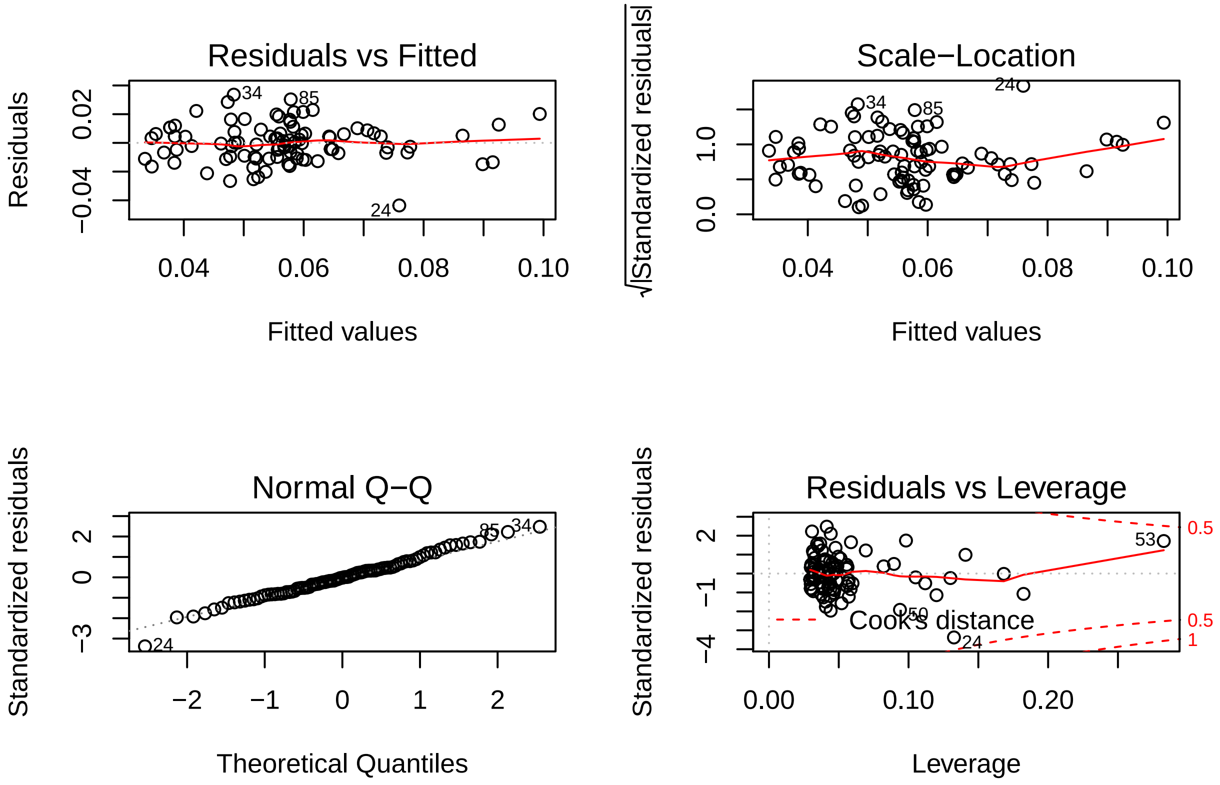
Both graphs show rapid improvement in explainability for first 3 variables before flattening off. We use the best 3 variables as recommended by stepwise regression.

```
stepsummary$outmat[1:3, ]
```

```
##          logprbarr logprbconv prbpris avgsen logpolpc density logtaxpc
## 1  ( 1 ) " "       " "        " "     " "    " "      "*"     " "     
## 2  ( 1 ) " "       " "        " "     " "    " "      "*"     " "     
## 3  ( 1 ) "*"       " "        " "     " "    " "      "*"     " "     
##          urbanurban pctmin80 wage regionother regionwest mix pctymle
## 1  ( 1 ) " "        " "      " "  " "         " "        " " " "    
## 2  ( 1 ) " "        " "      " "  "*"         " "        " " " "    
## 3  ( 1 ) " "        " "      " "  "*"         " "        " " " "    
```

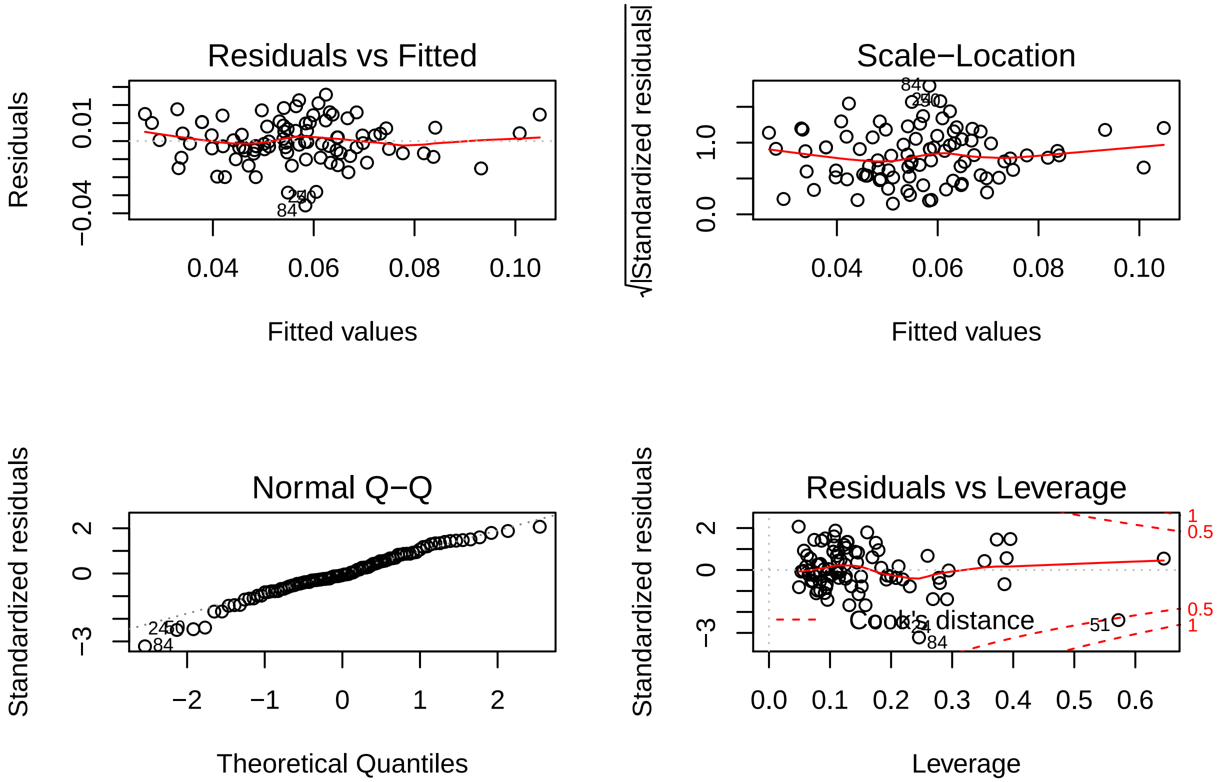
The most effective explanatory variables are: density, region and logprbconv

```
model2.1df <- df %>% select(sqrtfcrmrte, logprbconv, density, region)

model2.1 <- lm(formula = sqrtfcrmrte ~ logprbconv + density + region, data = model2.1df)

layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(model2.1)
```



####Model 3 **This model includes all the parameters**

```
model3.1 <- lm(formula = sqrtfcrmrte ~ logprbarr + logprbconv + prbpris + avgsen +
    logpolpc + density + logtaxpc + urban + pctmin80 + wage + region + pctymle, data = fmodeldf)

layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(model3.1)
```



Regression Tables

Crime Rate

```
cov1 <- vcovHC(model1, type = "HC")
robust.se1 <- sqrt(diag(cov1))
cov2 <- vcovHC(model2, type = "HC")
robust.se2 <- sqrt(diag(cov2))
cov3 <- vcovHC(model3, type = "HC")
robust.se3 <- sqrt(diag(cov3))

stargazer(model1, model2, model3, se = list(NULL, robust.se1, robust.se2, robust.se3),
    type = "latex", report = "vc", font.size = "small", star.cutoffs = c(0.05, 0.01,
        0.001), title = "Table 1:Linear Models to predict log crime rate", add.lines = list(c("AIC",
        round(AIC(model1)), round(AIC(model2)), round(AIC(model3))), c("BIC", round(BIC(model1)),
        round(BIC(model2)), round(BIC(model3)))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 05, 2018 - 8:38:52 PM

Table 1: Table 1:Linear Models to predict log crime rate

	Dependent variable:		
	logcrmrte		
	(1)	(2)	(3)
density	0.145	0.131	0.148
logprbarr	−0.406	−0.330	−0.386
wage	0.002	0.003	0.003
logprbconv	−0.152	−0.144	−0.158
prbpris			−0.040
avgsen			−0.035
logpolpc			0.299
regionother		0.255	0.249
regionwest		−0.269	−0.187
mix			−0.109
logtaxpc		0.011	−0.084
urbanurban			−0.177
pctmin80			0.004
pctymle		2.149	1.501
Constant	−4.307	−4.899	−2.308
AIC	97	71	70
BIC	112	96	110
Observations	91	91	91
R^2	0.497	0.651	0.700
Adjusted R^2	0.474	0.617	0.645
Residual Std. Error	0.396 (df = 86)	0.338 (df = 82)	0.325 (df = 76)
F Statistic	21.269^{***} (df = 4; 86)	19.142^{***} (df = 8; 82)	12.673^{***} (df = 14; 76)
Note:			$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Face-to-Face Crime Rate

```r
cov4 <- vcovHC(model1.1, type = "HC")
robust.se4 <- sqrt(diag(cov4))
cov5 <- vcovHC(model2.1, type = "HC")
robust.se5 <- sqrt(diag(cov5))
cov6 <- vcovHC(model3.1, type = "HC")
robust.se6 <- sqrt(diag(cov6))

stargazer(model1.1, model2.1, model3.1, se = list(NULL, robust.se4, robust.se5, robust.se6),
    type = "latex", report = "vc", font.size = "small", star.cutoffs = c(0.05, 0.01,
        0.001), title = "Table 2:Linear Models to predict log f2f crime rate", add.lines = list(c("AIC"
        round(AIC(model1.1)), round(AIC(model2.1)), round(AIC(model3.1))), c("BIC",
        round(BIC(model1.1)), round(BIC(model2.1)), round(BIC(model3.1)))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 05, 2018 - 8:38:53 PM

Table 2: Table 2:Linear Models to predict log f2f crime rate

	Dependent variable:		
	sqrtfcrmrte		
	(1)	(2)	(3)
density	0.008	0.005	0.007
logtaxpc			−0.004
urbanurban			−0.003
pctmin80			0.0003
logprbarr	0.004		0.001
wage	−0.0001		−0.0001
regionother		0.008	0.006
regionwest		−0.010	−0.006
pctymle			−0.009
logprbconv	−0.006	−0.007	−0.007
prbpris			0.020
avgsen			−0.001
logpolpc			0.014
Constant	0.106	0.075	0.199
AIC	-494	-514	-521
BIC	-479	-499	-483
Observations	91	91	91
R^2	0.368	0.492	0.613
Adjusted R^2	0.339	0.468	0.548
Residual Std. Error	0.015 (df = 86)	0.014 (df = 86)	0.013 (df = 77)
F Statistic	12.529^{***} (df = 4; 86)	20.803^{***} (df = 4; 86)	9.391^{***} (df = 13; 77)
Note:			*p<0.05; **p<0.01; ***p<0.001

Omitted Variable Analysis

Multicollinearity in the data set provided

Plot of pairwise correlations shows varying degrees of correlations between the covariates. To respect parsimony, we have chosen a subset of covariates available to us. This gives us good level of prediction power with least number of variables, however care is needed when making policy decisions. For example our model 2 for face to face crime rate, shows strong effect of probability of conviction. Probability of conviction, however, is correlated with other variables related to enforcement. So this one variable is likely to be working overtime and including effects of related variables as well. When making policy, simply improving this one covariate might not give the effect policy makers might be looking for.

Variables not present in the dataset

There are potentially other factors that impact crime rate which are not included in the data set provided. Some of these factors are:

1. **Unemployment** - The following paper discussed relationship between levels and fluctuations in rate of unemployment to several crime indexes: Cantor, David, and Kenneth C. Land. "Unemployment and Crime Rates in the Post-World War II United States: A Theoretical and Empirical Analysis." American Sociological Review, vol. 50, no. 3, 1985, pp. 317-332. JSTOR, JSTOR, www.jstor.org/stable/2095542.
2. **Weather** - The following paper discusses evidence of positive correlation between temperature and crime rate.

http://drexel.edu/now/archive/2017/September/Violent-Crime-Increases-During-Warmer-Weather-No-Matter-the-Season/

3. **Income Inequality** - The following worldbank report explores the relationship of income inequality on crime rate after controlling for overall economic conditions of a region.

http://documents.worldbank.org/curated/en/236161468299090847/pdf/WPS6935.pdf

Conclusions

As a political organization we want to focus on variables that substantially impact our response variables, but also on which we can affect change. We also want to be mindful of unintended consequences of change. While reducing percent minority, for example, seems to reduce both crime rate and face-to-face crime rate, it will likely have unintended consequences toward diversity and culture. It is also quite hard to quantify what changes in each of these variables would cost, both monetarily and in tradeoffs to society. That being said, our conclusions will remain largely qualitative.

Probability of arrest, probability of conviction, and average sentence are all negatively correlated with our response variables. This would suggest that focusing on enforcement in our judicial system could promise lower crime rates. Even if one or all of these variables are confounded, improvements to the judicial system seem to be the most promising.

Interestingly, police per capita is highly correlated with crime rate and face-to-face crime rate. This is almost certainly not causal in nature, based on intuition. It is much more reasonable that high crime rates *require* more police force and the direction of causality goes in the other direction. It would be very interesting to run an experimental setup with police distribution, perhaps an AB test, to see its causal effect on our response variables. However, we should also keep in mind the ethics behind this type of experiment and not leave a community at risk as we try to determine optimal distribution of law enforcement.