W203 Lab 3

Armand Kok, Adam Yang, James De La Torre

Introduction

Our team has been hired by a local political campaign to provide research on North Carolina crime statistics and to generate policy suggestions for reducing crime. Our candidate seeks to portray herself as being "pro-cop" and "tough on crime", and she espouses strong policing and enforcement. She also has a strong desire to understand the situations faced by the minority population within the state, and has expressed a keen interest in understanding how minority communities are impacted by crime.

The crime statistics dataset provided for analysis is a subset of the data used by Cornwell and W. Trumball in their 1994 study. The dependent variable of our study is the crimes commited per capita, given as `crmrate`. There are 24 other variables in the dataset, each of which can be potential modulators of the crime rate. We aim to build a linear model that regresses `crmrate` on the key variables in the dataset. In particular, we are interested in examining the potential of the following policies in reducing crime rate:

- Policy to increase the police per capita of a county
- Policy to implement a more stringent arrest protocol
- Policy to enhance community outreach in high density and minority communities

In addition, we aim to identify other factors that may influence crime and attempt to fully explore other possible political strategies. Not all correlating variables will have an actionable solution, though their inclusion in the regression model will contribute to its accuracy.

2.0 Data Loading and Cleaning

TO DO: Top Coding?

The data provided is a sample from 91 counties in North Carolina, containing information from 1987. The variables in the dataset and their meanings are shown below:

Variable	Label	Variable	Label
county	county identifier	**urban**	=1 if in SMSA
year	1987	**pctmin80**	perc. minority, 1980
crmrte	crimes committed per person	**wcon**	weekly wage, construction
prbarr	'probability' of arrest *	**wtuc**	wkly wge, trns, util, commun
prbconv	'probability' of conviction *	**wtrd**	wkly wge, whlesle, retail trade
prbpris	'probability' of prison sentence *	**wfir**	wkly wge, fin, ins, real est
avgsen	avg. sentence, days	**wser**	wkly wge, service industry
polpc	police per capita	**wmfg**	wkly wge, manufacturing
density	people per sq. mile	**wfed**	wkly wge, fed employees
taxpc	tax revenue per capita	**wsta**	wkly wge, state employees
west	=1 if in western N.C.	**wloc**	wkly wge, local gov emps
central	=1 if in central N.C.	**mix**	offense mix: face-to-face/other
pctymle	percent young male		

***** *These are not true probabilities that are limited between 0 and 1, but are ratios instead.* `probarr` *is the ratio of arrests to offenses,* `probconv` *is the ratio of convictions to arrests, and* `prbpris` *is the ratio of convictions resulting in an prison sentence to total convictions. Therefore, some of these values can be greater than 1.*

For example, the offender can be convicted of multiple crimes after his arrest.

2.1 Loading the Data

The data file, `crime_v2.csv` was opened and found to contain 97 rows.

```
# Import all libraries that will be used in the lab
library(car)
library(reshape2)
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Set directory based on who is running code
if (file.exists("/Users/adamyang/")) {
    mydir <- "/Users/adamyang/Desktop/w203/Lab3/w203-Lab3/"
} else if (file.exists("C:/Users/ak021523/")) {
    mydir <- "C:/Users/ak021523/Documents/GitHub/mids-repos/W203/Homework/w203-Lab3/"
} else {
    mydir <- "F:/users/jddel/Documents/DATA_SCIENCE_DEGREE_LAPTOP/W203_Stats/Lab_03/"
}

# read df
crime_df = read.csv(paste0(mydir, "crime_v2.csv"))
```

2.2 Data Cleanup

Immediate inspection of the data revealed a few requirements for data cleanup.

- The last 6 rows of the data set were blanks. These empty records were deleted.
- One row had values of 1 for both `west` and `central`, placing that county in two regions simultaneously. It is unknown whether this is possible, but currently there has been no reason to delete this particular row so the data will be kept for now.
- The `prbconv` variable, representing the "probability of conviction" was read in as a factor (a cateogorical variable) instead of a numeric variable. This variable was converted to numeric.

```
# get rid of rows with missing values (this only kills the 6
# blank rows)
crime_df <- crime_df[complete.cases(crime_df), ]
```

```
# convert prob of conviction to numeric
crime_df$prbconv <- as.numeric(as.character(crime_df$prbconv))
```

A summary was created for each of the variables and `probarr` and `probconv` stood out as having maximum values over 1. We believe this is valid because these variables are not true probabilities. Instead, they are proxied by the ratios of arrest:offenses and convictions:arrests, respectively. Therefore we decided not to omit these variables.

The `density` variable also stood out as having an extremely low minimum value. Upon inspection, the county in row 79 is revealed to have a density of 0.0000203422 person per sq. mile. This translates to 1 person in 49,159 square miles. Given that North Carolina is only 53,819 square miles, we believe this is an invalid value. The rest of the variables for this county seemed to have valid values so only the `density` observation was replaced with an `NA`.

```
summary(crime_df[, c("prbarr", "prbconv", "density")])
```

```
##      prbarr           prbconv           density
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.00002
##  1st Qu.:0.20568   1st Qu.:0.34541   1st Qu.:0.54741
##  Median :0.27095   Median :0.45283   Median :0.96226
##  Mean   :0.29492   Mean   :0.55128   Mean   :1.42884
##  3rd Qu.:0.34438   3rd Qu.:0.58886   3rd Qu.:1.56824
##  Max.   :1.09091   Max.   :2.12121   Max.   :8.82765
```

```
crime_df$density[79] = NA
```

A historgram was plotted for each of the variables and no evidence of top or bottom coding was found. However, we did notice a few other strong outliers that will be discussed in the next section.

2.3 Extreme Value Identification

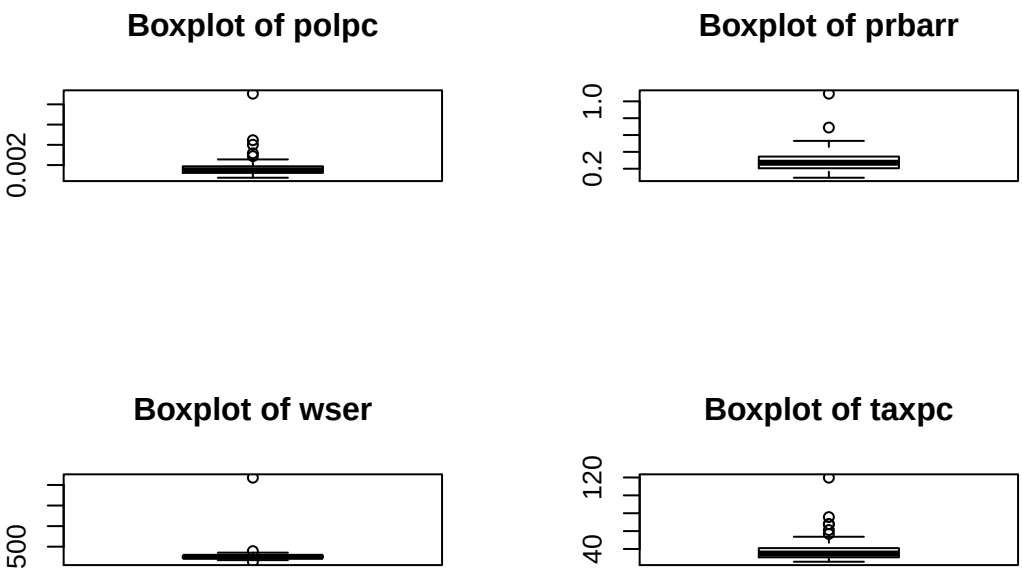
After generating histograms to review the distributions of the different variables, four were found to have extremely skewed data points:

- polpc - row 51
- prbarr - row 51
- wser - row 84
- taxpc row 25

The boxplots shown below

After reviewing further, there was no reason for the extremely skewed data points to be removed from the data set. boxplots of the variables above are shown below.

```
m <- rbind(c(1, 2), c(3, 4))
layout(m)
boxplot(crime_df$polpc, main = "Boxplot of polpc")
boxplot(crime_df$prbarr, main = "Boxplot of prbarr")
boxplot(crime_df$wser, main = "Boxplot of wser")
boxplot(crime_df$taxpc, main = "Boxplot of taxpc")
```



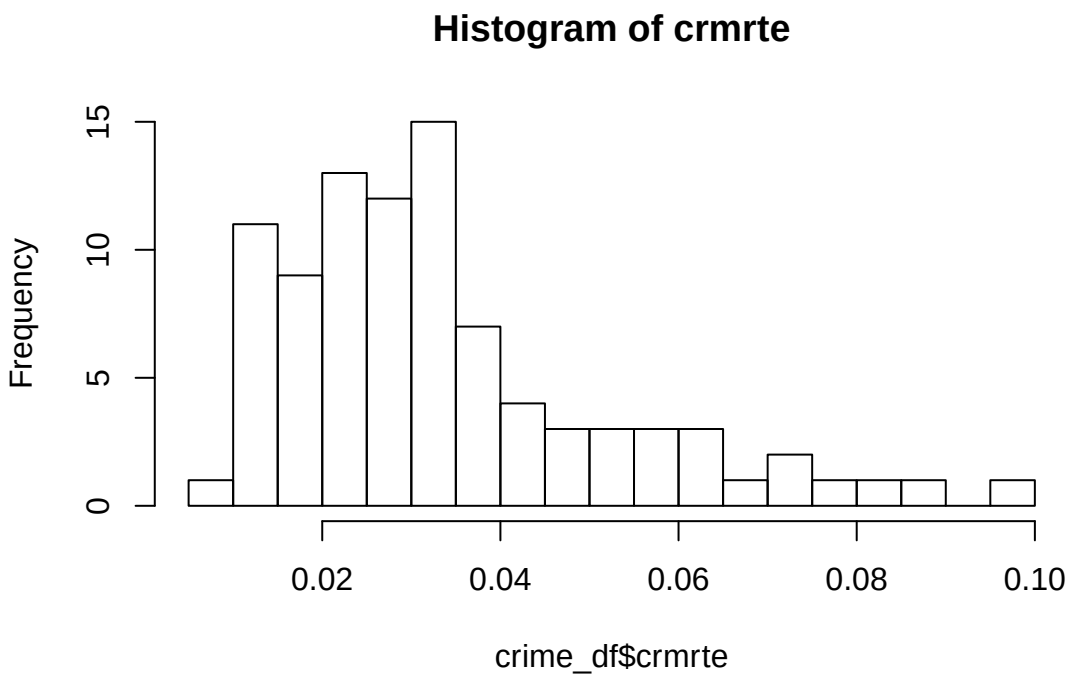
3.0 Model Building Process

TO DO: remove this instruction text

Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Is there a thorough univariate analysis of the outcome variable. Did the team identify at least two key explanatory variables and perform a thorough univariate analysis of each? Did the team clearly state why they chose these explanatory variables, does this explanation make sense in term of their research question? Did the team consider available variable transformations and select them with an eye towards model plausibility and interperability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

It is important to examine our outcome variable, `crmrte` before building any models.

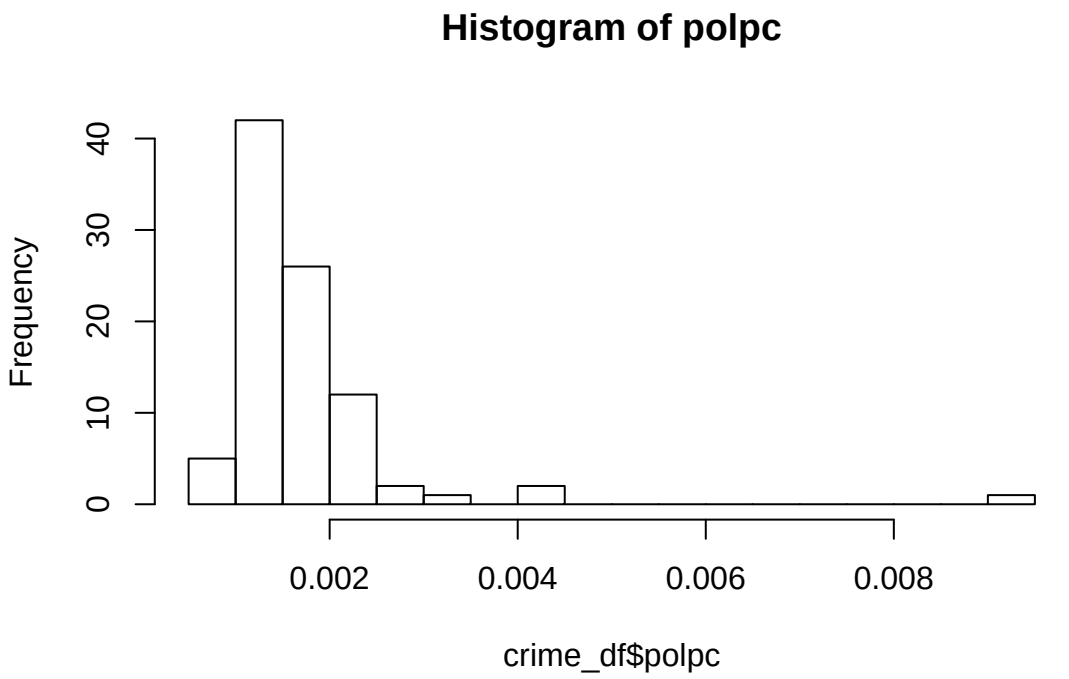
```
hist(crime_df$crmrte, main = "Histogram of crmrte", breaks = 30)
```



The histogram of `crmrte` shows some positive skew, but there are no extreme outliers.

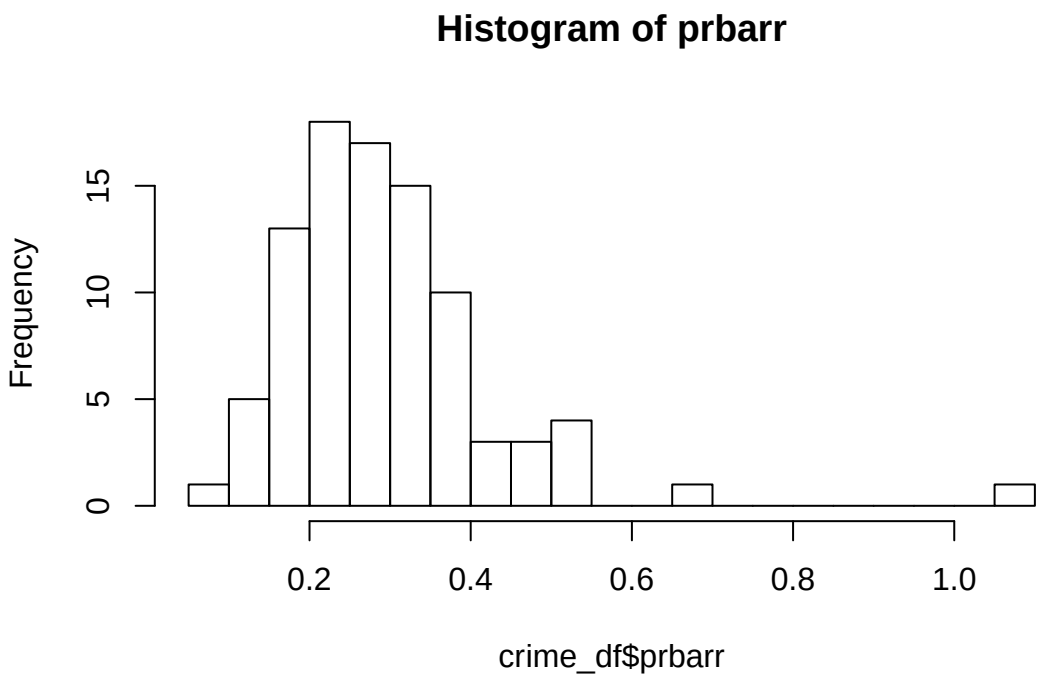
Given that our candidate is considering policies involving increasing the size of the police force, instituting stricter arrest protocols, and addressing issues of minorities in densely populated areas, the police per capita (`polpc`), probability of arrest, (`prbarr`), population per square mile (`density`), and percent minority (`pctmin80`) variables will be examined more closely. Histograms of these variables are shown below.

```
hist(crime_df$polpc, main = "Histogram of polpc", breaks = 20)
```



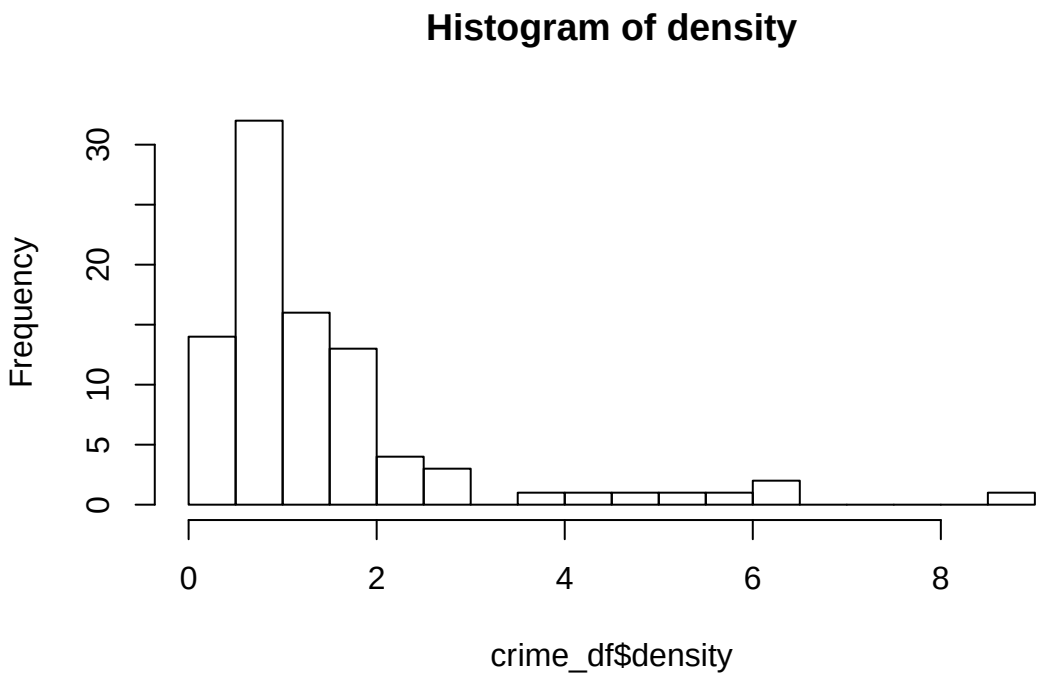
The histogram of `polpc` shows the point with an extreme value mentioned in section 2.3.

```
hist(crime_df$prbarr, main = "Histogram of prbarr", breaks = 20)
```



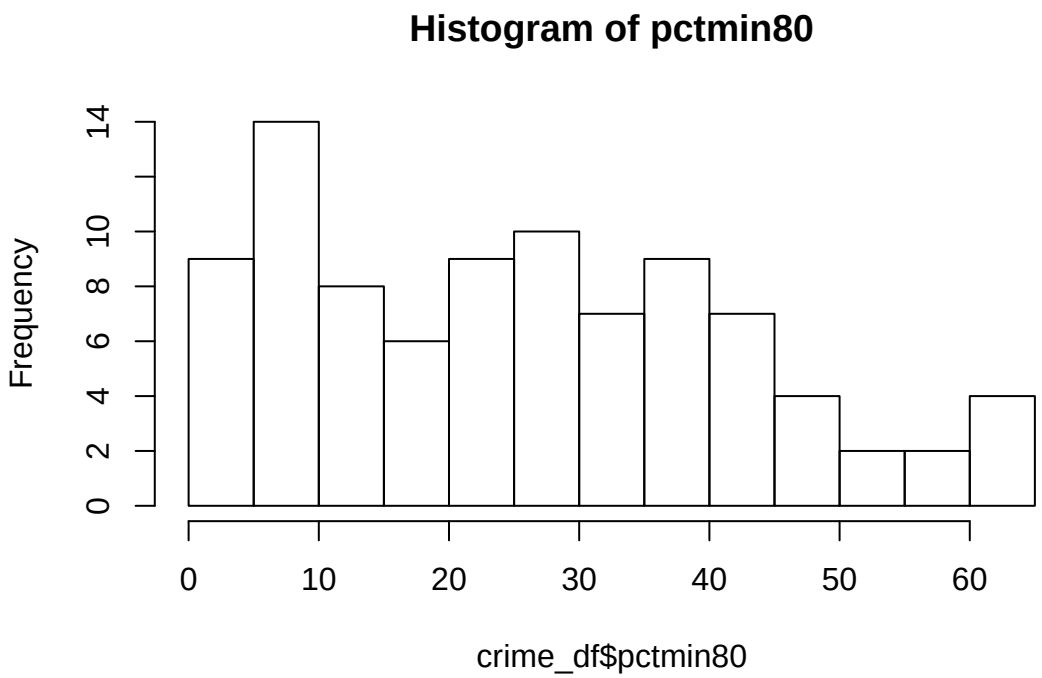
The histogram of `prbarr` also shows a point with an extreme value, which is the same record that has an extreme value for `polpc`.

```
hist(crime_df$density, main = "Histogram of density", breaks = 20)
```



Population density has a positive skew, which is likely due to the few counties with large cities.

```
hist(crime_df$pctmin80, main = "Histogram of pctmin80", breaks = 20)
```



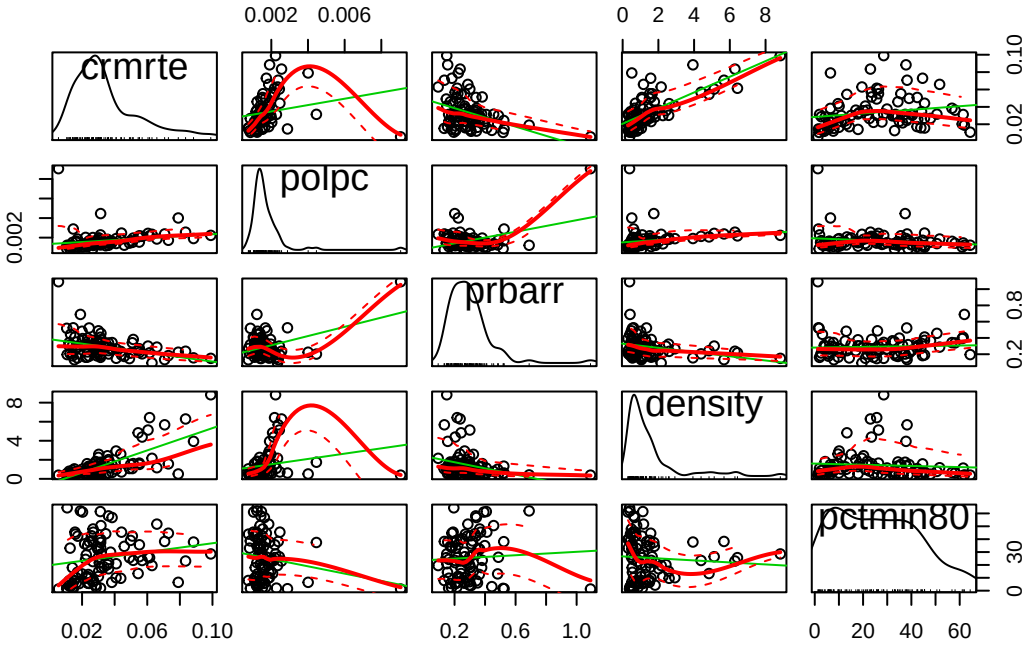
There are no extreme values in the histogram of `pctmin80`.

With the exception of the one record that has extreme values for both `polpc` and `prbarr`, the key variables in our dataset that most closely relate to our candidate's policy interests appear to have distributions that can be used for modeling without the need for any transformations. As we build models, we will watch for high influence from the record with extreme values (#51).

3.1 Scatterplot Matrix

To visualize the relationship between crime rate and our explanatory variables of interest, a scatterplot matrix was generated.

```
spm(~crmrte + polpc + prbarr + density + pctmin80, data = crime_df)
```



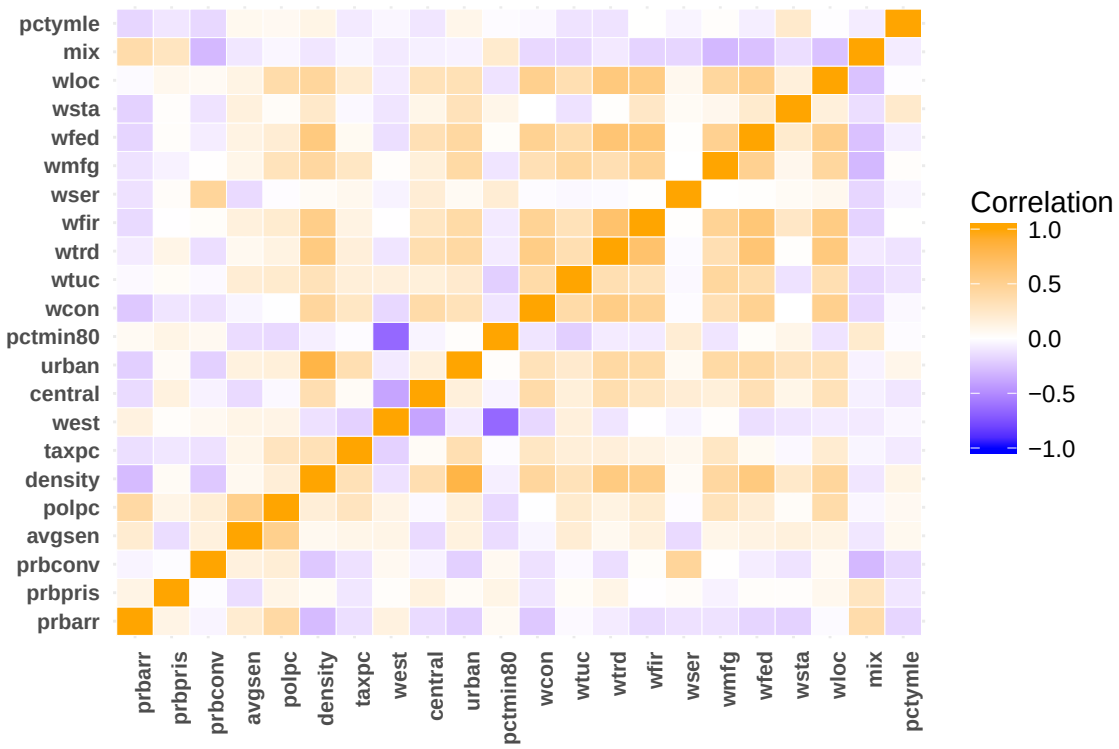
The plots reveal that each of the selected explanatory variables shows a relationship with crime rate. There is some degree of nonlinear relationship between `polpc` and `crmrte` and between `pctmin80` and `crmrte`. However, transforming these variables would distort the practical interpretability of any model slope coefficients. Therefore, the variables will not be transformed.

3.2 Check for multicolinearity

Since additional variables may be included in other models for crime rate, it is important to identify those explanatory variables with a high degree of colinearity. Perfectly colinear variables are prohibited in OLS regression, so in the unlikely event that any such variables are found, only 1 from each perfectly colinear group will be kept. Less-than-perfect colinearity can still be problematic, adding variance to a model, so if any highly colinear groups of variables are identified, only one variable from each group will be kept. This analysis will narrow down the set of candidate variables for inclusion in any models we may choose to build.

To identify colinear variables, a correlation matrix was generated as shown below.

```
# TO DO - fix matrix sizing

# correlation matrix for top 4 correlation and bottom 4
# correlation
cor_dr = cor(crime_df[c("prbarr", "prbpris", "prbconv", "avgsen",
    "polpc", "density", "taxpc", "west", "central", "urban",
    "pctmin80", "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg",
    "wfed", "wsta", "wloc", "mix", "pctymle")], use = "complete.obs")

# Heatmap
ggplot(data = melt(cor_dr, na.rm = TRUE), aes(Var2, Var1, fill = value)) +
    theme_minimal() + geom_tile(color = "white") + scale_fill_gradient2(low = "blue",
    high = "orange", mid = "white", midpoint = 0, limit = c(-1,
        1), name = "Correlation") + theme(axis.text.x = element_text(face = "bold",
    angle = 90, vjust = 1, size = 8, hjust = 1), axis.text.y = element_text(face = "bold",
    size = 8), axis.title.x = element_blank(), axis.title.y = element_blank())
```

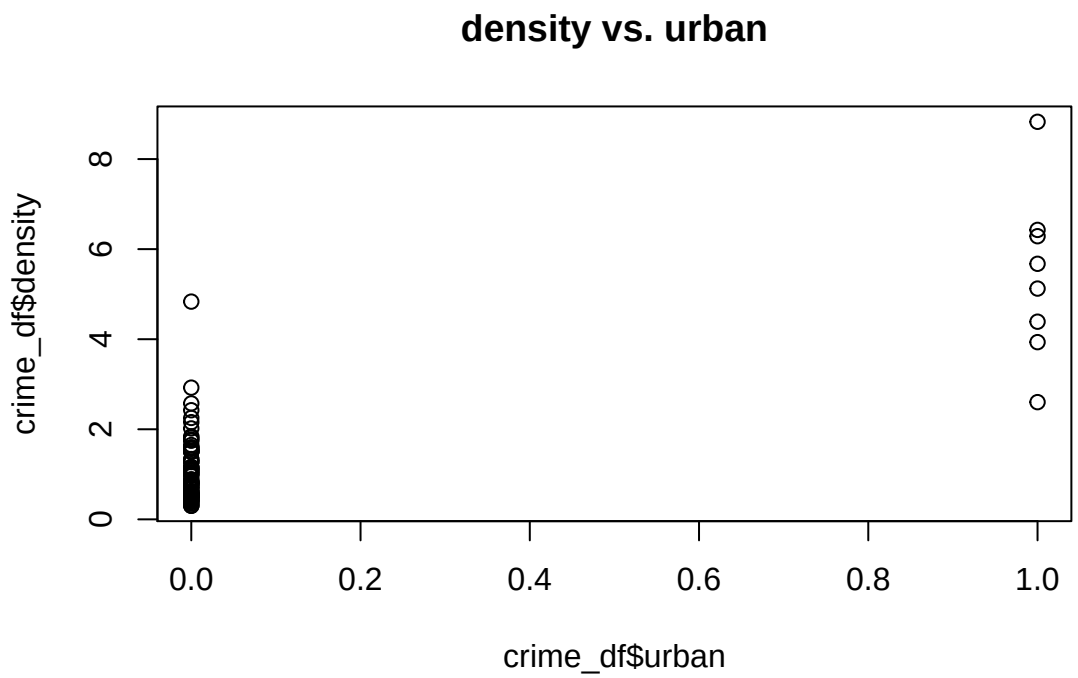


After reviewing the correlation matrix in detail, there were 5 pairs of variables that have a somewhat strong correlation to each other (i.e. has correlation > 0.6), which are listed below:

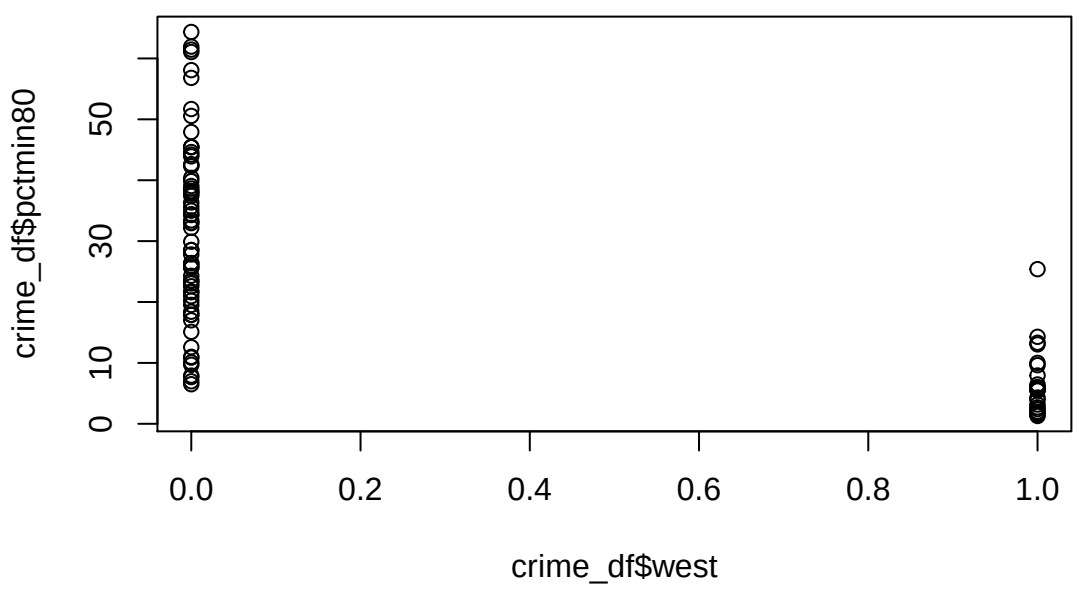
- `urban` (82% correlation with `density`. Kept `density` because it is a continuous variable providing more information than the categorical `urban` variable)
- `west` (-64% correlation with `pctmin80`. Kept `pctmin80` because it is one of the main drivers of our policies.)
- `wtrd`, `wfed`, `wfir` (each of these had correlations >60% with eachother and/or with `density` or other wage columns. Kept `density` as it can act as a proxy for the greatest number of other variables.)

Below are the scatterplots of the different correlated variables.

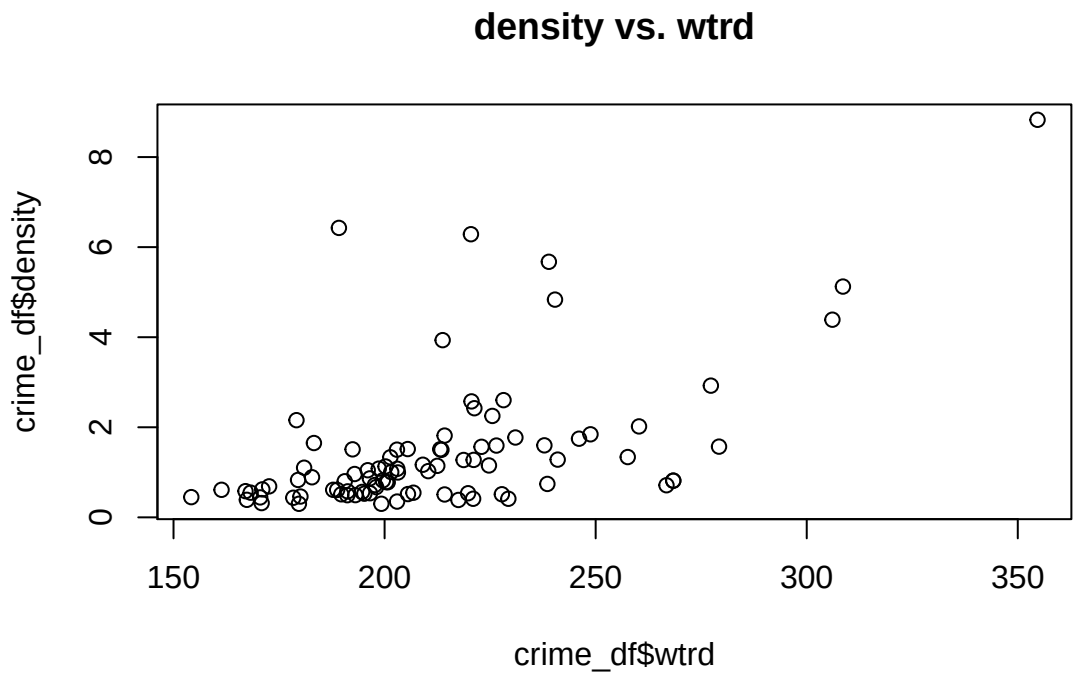
```
plot(crime_df$urban, crime_df$density, main = "density vs. urban")
```



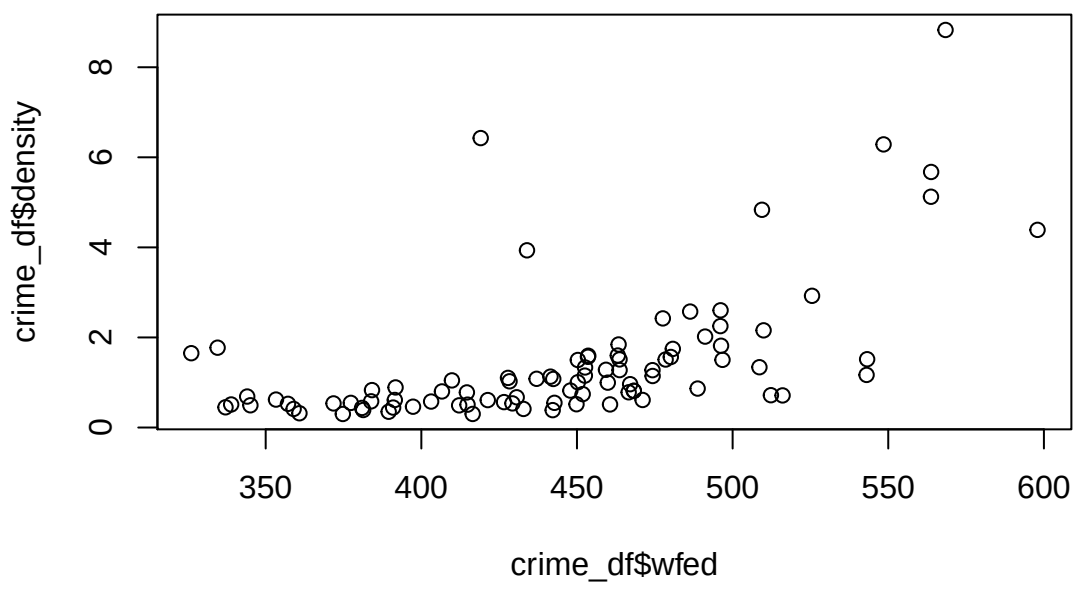

```
plot(crime_df$west, crime_df$pctmin80, main = "pctmin80 vs. west")
```



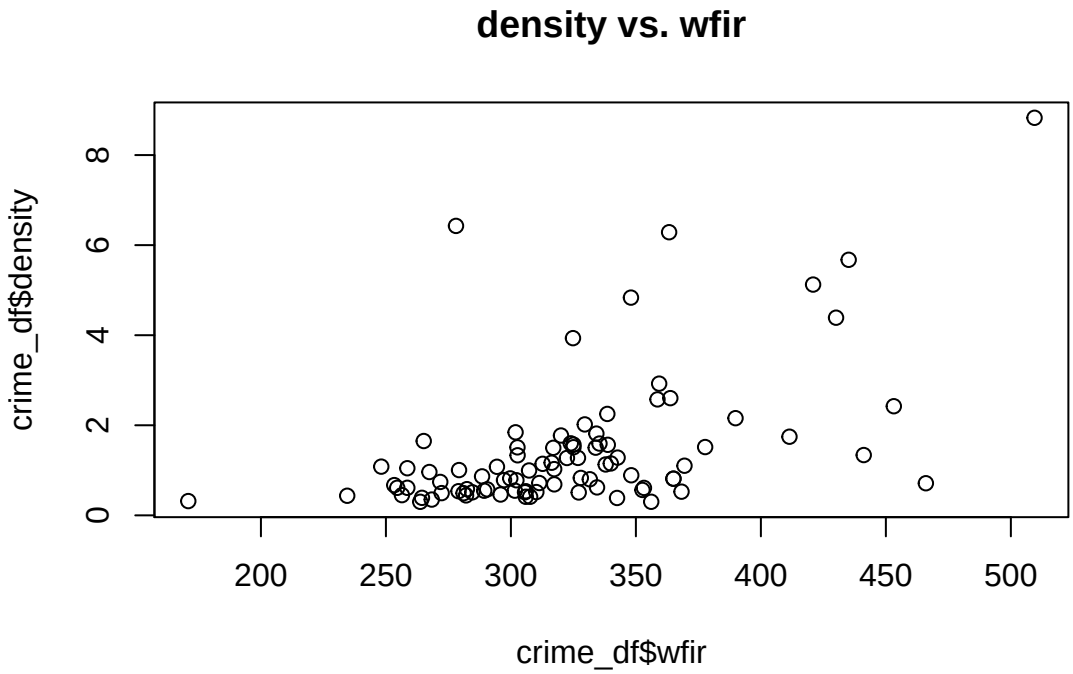

```
plot(crime_df$wtrd, crime_df$density, main = "density vs. wtrd")
```




```
plot(crime_df$wfed, crime_df$density, main = "density vs. wfed")
```




```
plot(crime_df$wfir, crime_df$density, main = "density vs. wfir")
```



4.0 Regression Models: Base Model

TO DO: create a new section with 6 CLM assumptions for the best model, and trim down this section.

The initial model created contains only those variables directly related to the candidate's positions on being pro-police, for strict enforcement, and concern with inner city and minority communities. Therefore, the variables we have chosen to represent these positions are: probability of arrest (prbarr), density, police per capita (polpc), and the percentage of minorities (pctmin80).

```
# Creating initial model
model1 <- lm(crmrte ~ prbarr + density + polpc + pctmin80, data = crime_df)
```

After creating the model, we will start by evaluating it against the six Classical Linear Model assumptions.

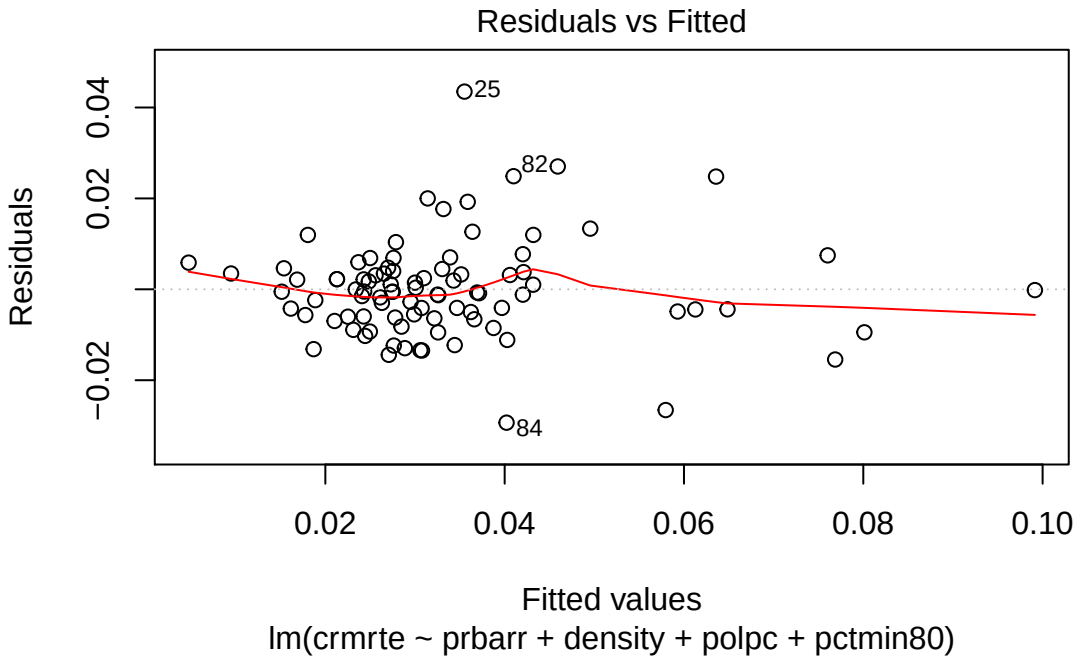
CLM 1. Linear population model: We do not have to worry about this assumption at the moment because we haven't constrained the error term.

CLM 2. Random Sampling: To check random sampling, we need domain knowledge and an understanding of how the data were collected. There are 100 counties in North Carolina, and there are data for 91 of them. Without knowledge of the 9 excluded counties, no statement regarding the validity of random sampling can be made.

CLM 3. No perfect multicollinearity: There is no need to explicitly check for perfect collinearity, because R would've reported a warning if this occurred. Furthermore, the correlation matrix shown in section "**TO DO**___" also shows that there is no perfect collinearity.

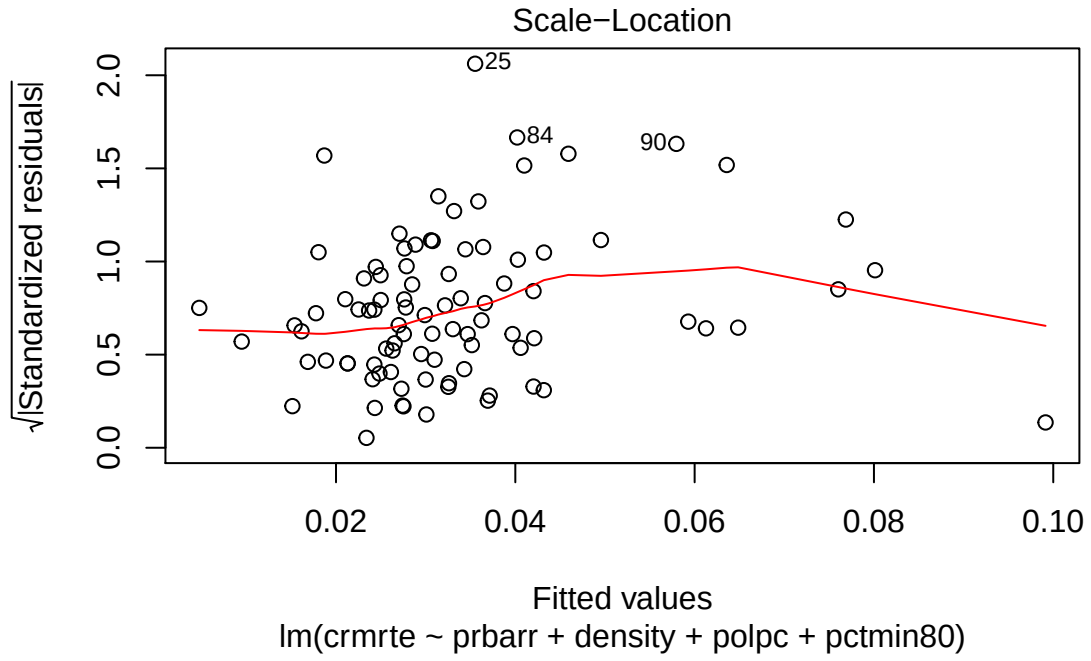
CLM 4. Zero Conditional Mean: $E(u|x) = 0$. For this model, the residuals vs. fitted values plot shown below reveals a relatively flat spline centered around zero. Therefore, there does not seem to be a clear deviation from the zero conditional mean and the assumption holds.

```
# Residuals vs. Fitted Plot
plot(model1, which = 1)
```



CLM 5. Homoscedasticity: In the residuals vs. fitted values plot shown in **CLM 4**, the data points seem to form a cone shape which suggests some heteroscedasticity. In the scale-location plot below, there seems to be a slight positive slope across the range of fitted values between 0.02 and 0.04. Furthermore, the Breusch-Pagan test shown below has a p-value of 6.278e-05 which indicates that the null hypothesis of homoscedasticity can be rejected. When evaluating the statistical significance of calculated model coefficients, heteroscedastic-robust standard errors will be used.

```
# Scale-Location Plot
plot(model1, which = 3)
```

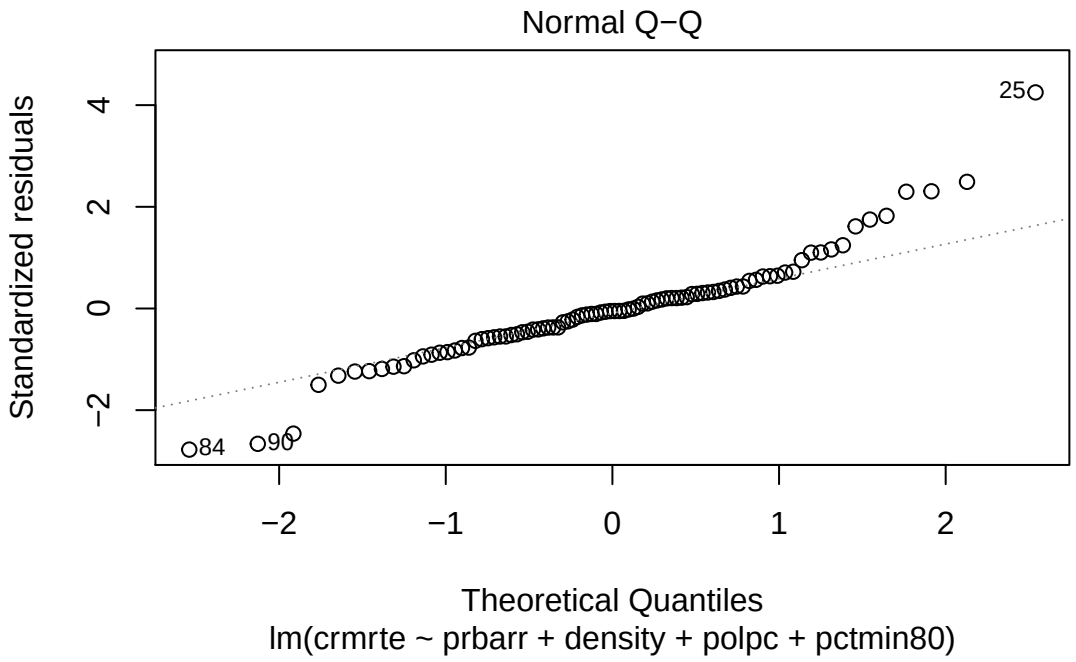



```
# Breusch-Pagan
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 25.072, df = 4, p-value = 4.866e-05
```

CLM 6. Normality of errors: In the Q-Q plot shown below, the bulk of the error terms seem to follow the straight line which suggests a fairly normal distribution. However, the standardized residuals show some deviation from the straight line at the extreme ends of the distribution. This suggests some skew at the extreme ends of our residuals. Furthermore, the Shapiro test shown below has a p value of 0.0002 which means we can reject the null hypothesis of the residuals having a normal distribution.

```
# Q-Q plot of Standardized Residuals
plot(model1, which = 2)
```

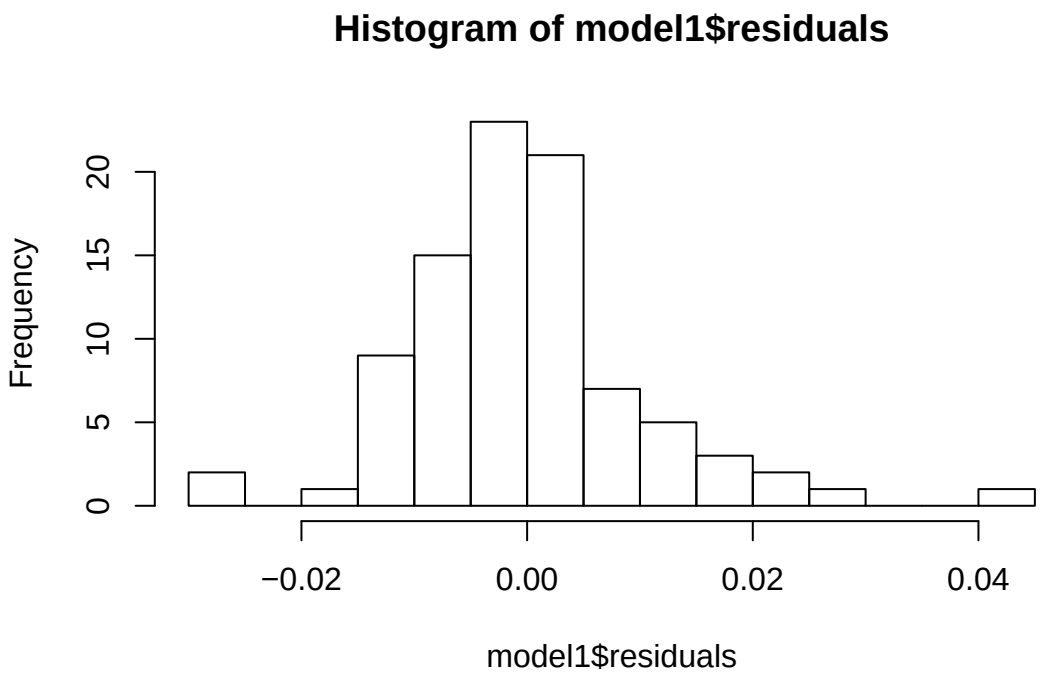



```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.93859, p-value = 0.0003529
```

To further verify this observation, a histogram of this model's residuals is shown below. The histogram shows approximate normality near the center of the distribution, but also some evidence of skewness; especially on the positive end. However, the Central Limit Theorem (CLT) claims that if the sample size is large enough we can assume that the residuals have a normal sampling distribution. For distributions with a very strong skew, a much larger sample size may be required, but for minor skews as in this case, the rule of thumb is that the CLT can be applied when the sample size is greater than 30. The sample size used for this model is 91 which should be enough for the CLT to hold.

```
hist(model1$residuals, breaks = 20)
```



Based on our review of the six CLM assumptions, this is a valid linear model. We replaced the regular standard errors with the heteroskedasticity-robust standard errors. The resulting coefficients and parameters of the model are shown below:

```
linearHypothesis(model1, c("prbarr = 0", " density = 0", "polpc = 0",
    "pctmin80 = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## density = 0
## polpc = 0
## pctmin80 = 0
##
## Model 1: restricted model
## Model 2: crmrte ~ prbarr + density + polpc + pctmin80
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     89
## 2     85  4 43.797 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
paste("adj.r.square:", summary(model1)$adj.r.squared)
```

```
## [1] "adj.r.square: 0.653401844637933"
```

The adjusted r-squared of the model is relatively high at 0.66. This means that 66% of the variation in crime rate is explained by our input variables. In addition, the omnibus test reveals that the model is statistically

significant, indicating that the model has some predictive power.

```
coeftest(model1, vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.9575e-02  9.7639e-03  2.0049  0.048160 *
## prbarr      -4.6038e-02  2.1484e-02 -2.1428  0.034985 *
## density      7.4746e-03  1.2128e-03  6.1631 2.303e-08 ***
## polpc        5.0860e+00  5.0122e+00  1.0147  0.313120
## pctmin80     3.1864e-04  8.6015e-05  3.7045  0.000376 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

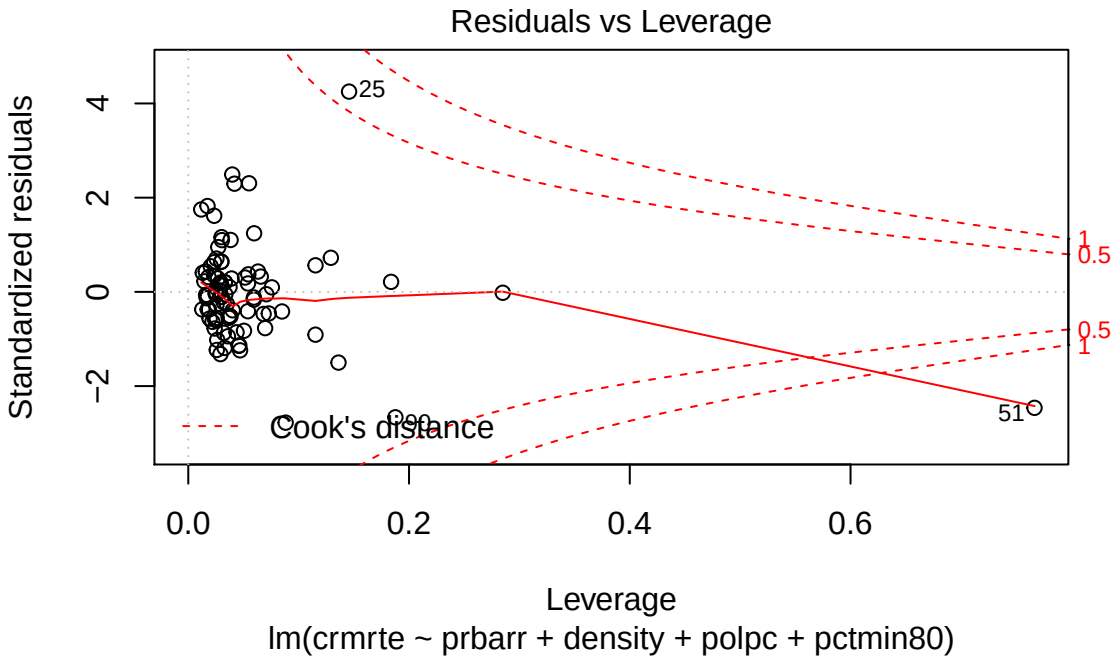
Furthermore, the results of our initial model shows that the probability of arrest is statistically significant as a modulator of crime, while the density and minority percentage of each county are strongly statistically significant. The police per capita, on the other hand, is not. The slope coefficients tell us that for every 1 unit increase in `prbarr`, there is a corresponding 0.046 decrease in the crime rate. The model also suggests that by increasing the density of a county by 1 person per square mile, crime commited per person may rise by 0.008. Finally, for every percentage point increase of minorities in a county, crime commited per person may rise by 0.0003. The model also suggests that by increasing the police per capita by 1 will result in 5 additional crimes commited per person. However, this slope coefficeint is shown to be statistically insignificant.

To further assess the strength of our model, we can take a look at the residuals vs. leverage plot shown below. Here we can see that data point 51, has a Cook's distance greater than 1, meaning it has high influence over the model. As shown in section 2.3 this data point has `polpc` and `prbarr` values multiple times higher than the next highest values for these variables. If this data point is not representative of the general population in North Carolina, then it may hurt the accuracy of our model. However, we investigated the other values of this county and could not justify removing this data point without further information.

Furthermore, a general rule is that if 1 % (or more) data points have standardized residuals > 2.5, the model contains too much error. If 5% (or more) of data points have residuals > 2, the model has too much error and represents our data poorly. In the residual vs. leverage plot below, we see that 7.7% of our data points have standardized residuals over 2. Therefore, our model has too much error and may represent our data poorly.

Because of this, we will now incorporate additional covariates that might increase the accuracy of our results.

```
plot(model1, which = 5)
```



4.1 Regression Model: Second Model

Does this model include covariates meant to increase the accuracy of the regression? Has the team justified inclusion of each of these additional variables? Does the team identify what they want to measure with each coefficient? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

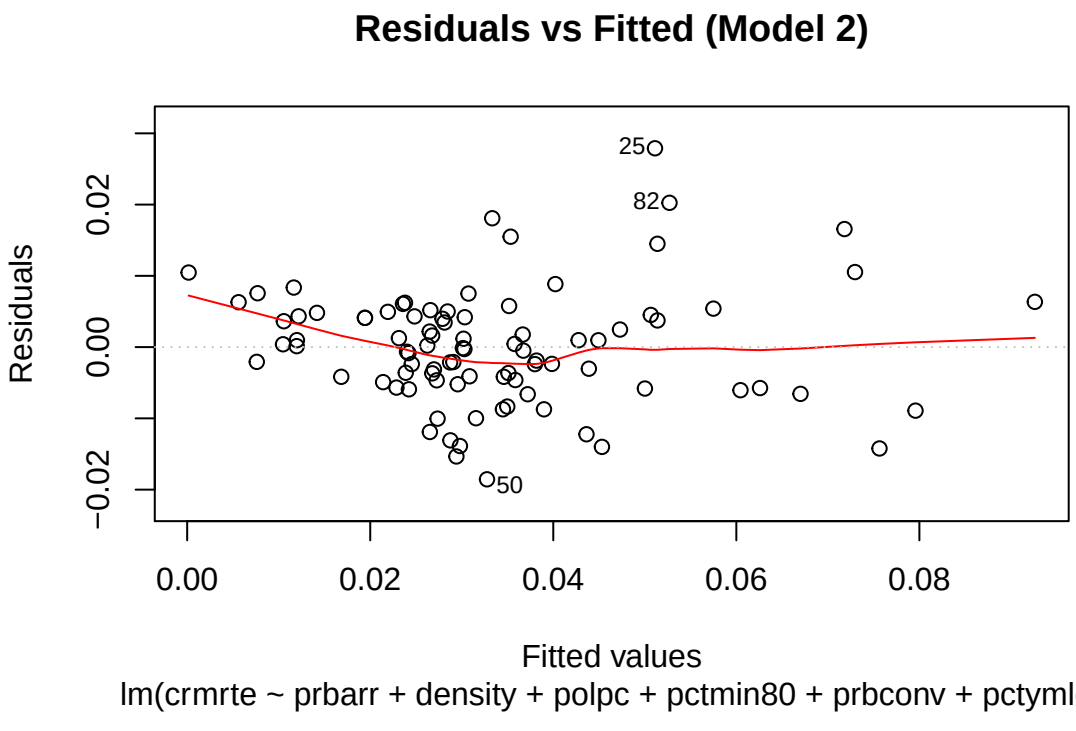
One model that includes key explanatory variables and only covariates that you believe increase the accuracy of your results without introducing substantial bias (for example, you should not include outcome variables that will absorb some of the causal effect you are interested in). This model should strike a balance between accuracy and parsimony and reflect your best understanding of the determinants of crime.

A second model was created which included the three original explanatory variables (probability of arrest, `prbarr`; population per square mile, `density`; and police per capita, `polpc`) plus two additional variables–the "probability" of conviction, `probconv` and percentage of a county's population comprised of young males `pctymle`. The probability of conviction was selected based on the thought that if someone believes he is more likely to be convicted if he commits a crime, he may be less inclined to take the risk of committing a crime. The percent young males variable was included because we believe that young males are responsible for a disproportionally large share of total crimes committed. Including these variables should improve the accuracy of our inferences for crime rate.

```
# new: prbconv pctymle
model2 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + prbconv +
    pctymle, data = crime_df)
```

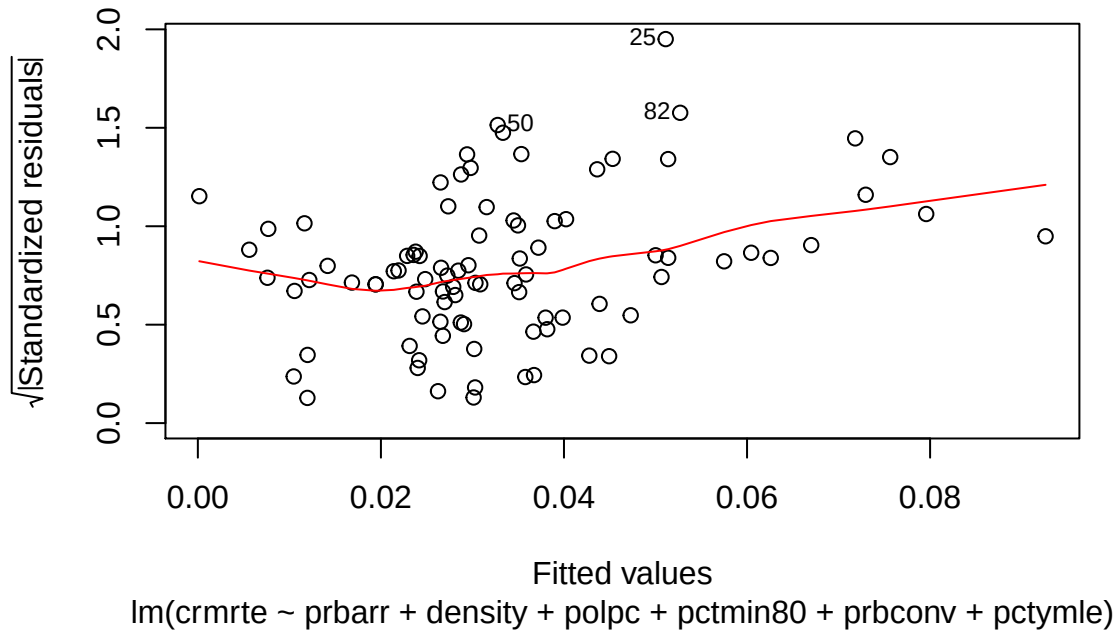
This second model produced some unexpected results with respect to the linear model assumptions. First, the residuals show more deviation from the zero conditional mean assumption than our previous model exhibited. The residuals vs. fitted values plot below shows positive residuals for fitted values less than 0.015. Perhaps there is an omitted variable responsible for this, or perhaps there is a nonlinear relationship between some of the variables in the model and crime rate.

```
plot(model2, which = 1, caption = "", main = "Residuals vs Fitted (Model 2)")
```



Judging from the Scale-Location plot below, model 2 still exhibits heteroscedasticity, albeit slighthly less heteroscedastic than model 1. Furthermore, the Breusch-Pagan test has a p-value of 0.0005, which reaffirms the findings from the Scale-Location plot. Using heteroscedastic-robust standard errors when evaluating the model coefficients should prevent this from being a problem.

```
plot(model2, which = 3, caption = "", main = "Scale-Location (Model 2)")
```



```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 23.399, df = 6, p-value = 0.0006734
```

The summary of the second model's output is shown below.

```
paste("Model 2 adj.r.square:", summary(model2)$adj.r.squared)
```

```
## [1] "Model 2 adj.r.square: 0.800320933946948"
```

```
coeftest(model2, vcovHC)
```

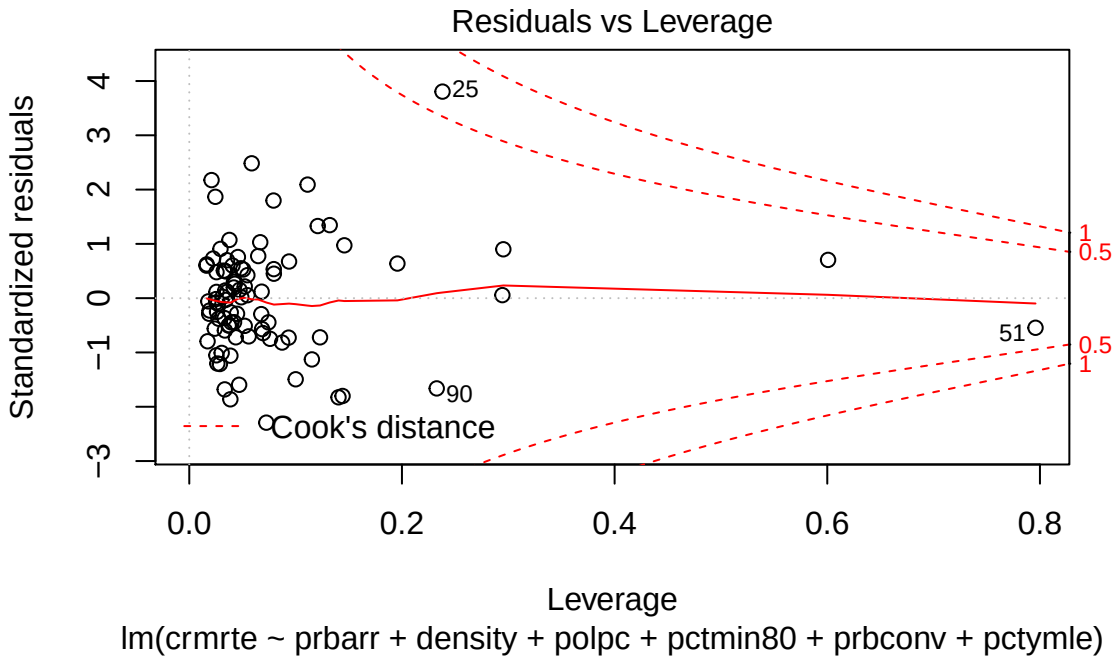
```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.02756913  0.00810917  3.3997 0.0010389 **
## prbarr      -0.06205084  0.01534721 -4.0431 0.0001174 ***
## density      0.00549532  0.00124583  4.4110 3.066e-05 ***
## polpc        8.00114865  2.57465856  3.1077 0.0025822 **
## pctmin80     0.00036814  0.00005417  6.7960 1.520e-09 ***
## prbconv     -0.02112277  0.00464877 -4.5437 1.859e-05 ***
## pctymle      0.06026608  0.05486783  1.0984 0.2752118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second model has an adjusted r-squared value of 0.80, meaning 80% of the variation in crime rate is explained by the explanatory variables in the model. The police per capita variable, `polpc`, which was not statistically significance in the first model is now significant in this second model. As for the two variables added in the second model, the probability of conviction, `prbconv`, is highly statistically significant, while the percent young male, `pctymle`, variable, suprisingly, is not.

In the second model, the slope coefficients can be interpreted as follows:

- For every 1 percentage point increase in the probability of arrest, crime decreases by 0.0006 crimes per person.
- For every 1 additional person per square mile, crime increases by 0.0056 crimes per person.
- For every 1 additional police officer per person, crime increases by 7.7 crimes per person.
- For every 1 percentage point increase in minority population, crime increases by 0.00037 crimes per person
- For every 1 percentage point increase in probability of conviction, crime decreases by 0.00021 crimes per person.

```
plot(model2, which = 5)
```



Judging from the Residuals vs. Leverage plot above, point 51 no longer has Cook's distance greater than 1 like it did in model 1. This indicates that point 51 is not influential on the new regression model. This is due to the additional information that the new variables bring to the model.

In addition, the Residual vs. Leverage plot above, we see that there are fewer data points that have standardized residuals over 2. Therefore, model 2 represents our data better than model 1 does.

4.2 Regression Third model

The following is the model that contains almost all available variables as explanatory variables with the exception of variables we excluded due to high level of multi-collinearity.

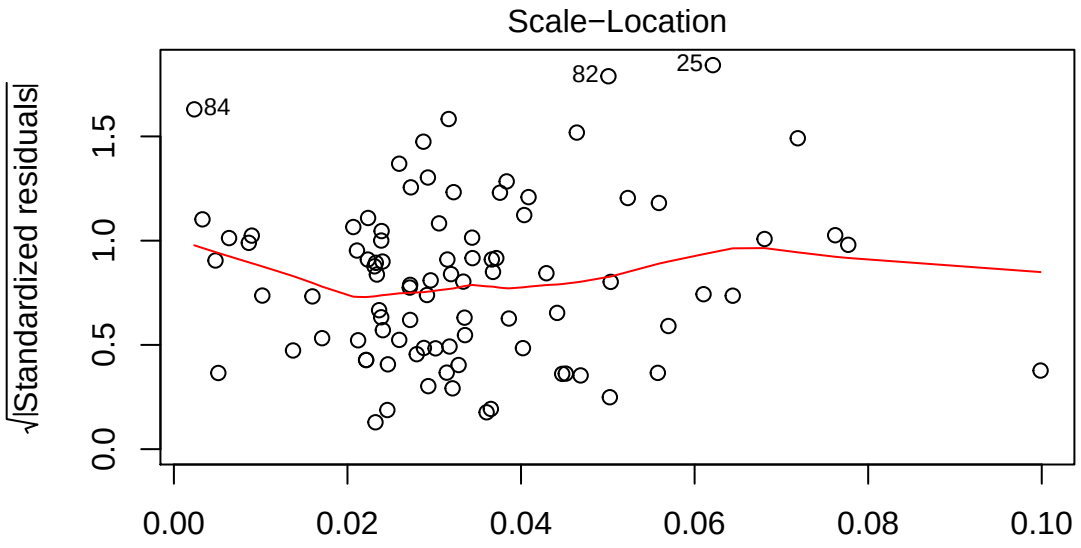
```
model3 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + prbconv +
    pctymle + log(wcon) + log(wtuc) + log(wser) + log(wmfg) +
    log(wsta) + log(wloc) + log(wser) + taxpc + central + mix +
    prbpris + avgsen, data = crime_df)
```

Model 3 for most part follows meets most of the CLM assumptions, however there are exceptions, and some other interesting points discussed below.

CLM 5. Homoscedasticity:

The Scale-Location plot below reveals that there seems to be heteroscedacity given that the spline is not horizontal. Howver, the Breusch-Pagan test has a p-value greater than 0.05, meaning we cannot reject the null hypothesis of homoscedasticity. Based on the conflicting findings, to be conservative, heteroscedastic-robust standard errors will be used to perform any sort of statistical test for the model.

```
# scale location plot
plot(model3, which = 3)
```




```
# breusch-pagan test
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 25.98, df = 17, p-value = 0.07482
```

```
# print adj r squared
paste("adj.r.square:", summary(model3)$adj.r.squared)
```

```
## [1] "adj.r.square: 0.814106405769601"
```

```
# test coefficient significance
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.9148e-03  1.4455e-01  0.0132 0.9894674
## prbarr      -5.3074e-02  1.2944e-02 -4.1002 0.0001071 ***
## density      5.5383e-03  1.4800e-03  3.7421 0.0003640 ***
## polpc        7.2473e+00  2.4971e+00  2.9023 0.0049121 **
## pctmin80     3.9441e-04  8.0479e-05  4.9007 5.702e-06 ***
## prbconv     -1.8942e-02  5.9676e-03 -3.1741 0.0022118 **
## pctymle      9.2327e-02  3.7729e-02  2.4471 0.0168418 *
## log(wcon)    8.3404e-03  9.7747e-03  0.8533 0.3963415
## log(wtuc)    3.3996e-03  8.1728e-03  0.4160 0.6786736
## log(wser)   -5.2561e-03  1.7096e-02 -0.3074 0.7593917
## log(wmfg)   -1.9394e-03  7.8116e-03 -0.2483 0.8046308
## log(wsta)   -9.5841e-03  1.2549e-02 -0.7637 0.4475364
## log(wloc)    8.8453e-03  2.2973e-02  0.3850 0.7013468
## taxpc        1.5628e-04  2.7362e-04  0.5712 0.5696663
```

```
## central     -2.7426e-03  2.6009e-03 -1.0545 0.2951966
## mix         -2.0703e-02  2.1302e-02 -0.9719 0.3343548
## prbpris      6.5066e-04  1.5253e-02  0.0427 0.9660914
## avgsen      -4.6691e-04  4.4409e-04 -1.0514 0.2965946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared to model 2, the adjusted R-squared is only marginally higher, this suggest that we will need to further evaluate the joint significance of the additional variables that were included as part of model 3. The following are the interpretation of the signifcant coefficients:

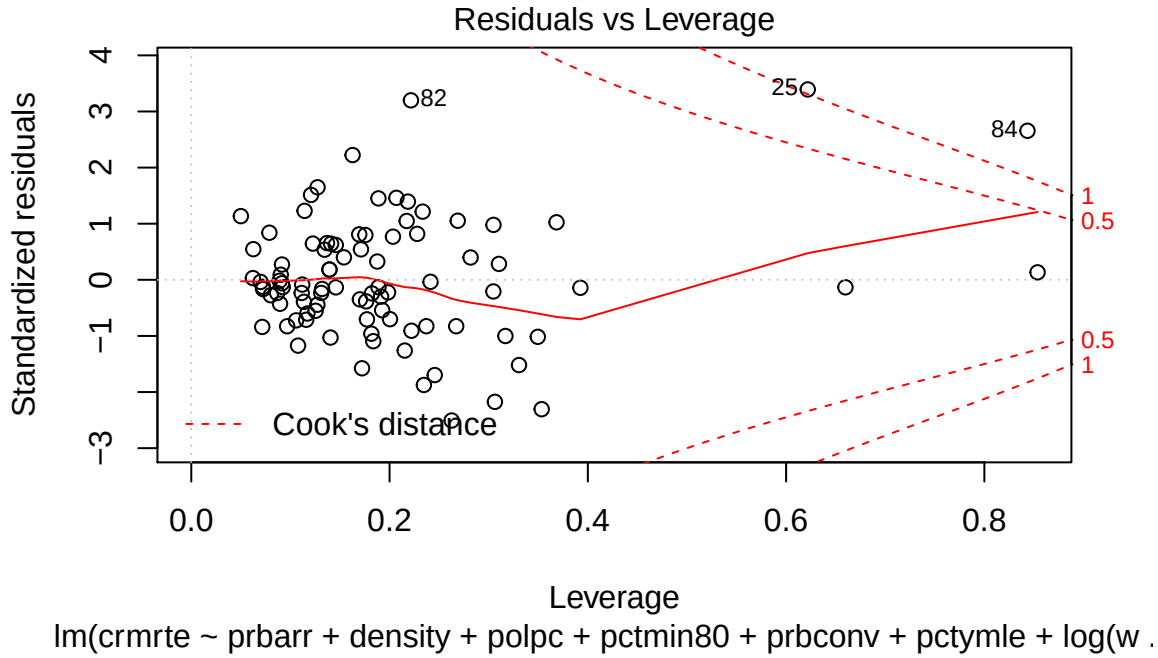
- For every percentage point increase in increase in probability arrest, crime rate decreases by 0.0005.
- For every 1 additional person per square mile, crime increases by 0.0056 crimes per person.
- For every 1 addtional police officer per person, crime increases by 7.7 crimes per person.
- For every percentage point increase in minority populatoin, crime increases by .0004 crimes per person.
- For every percentage point increase in probability of conviction, crime decreases by 0.00019 crimes per person
- For every percentage point increase in young male population, crime increases by 0.00095 crimes per person.

What is interesting here is that `pctymle` variable is now considered significant, unlike what is in model 2. This may be due to the nuances that the additional variables bring for the `pctymle` variable.

Other Analysis The residuals vs. leverage plot shows that there are two data points (25 and 84), that have Cook's distance greater than 1, indicating that they highly influence the regression model. In looking at data point 84 further, there are several things that stand out: it has the highest `wser`, `prbconv`, and `pctmin80`. On the other hand point 25, highest `taxpc`. However, we investigated the other values of this county and could not justify removing this data point without further information.

Furthermore, in the residual vs. leverage plot below, we see that compared to model 2, we now have more data points have standardized residuals over 2. Therefore, compared to model 2, model 3 does not provide a better representation of the data.

```
plot(model3, which = 5)
```



4.3 Regression Table

The regression table below shows the summary of the three models that were created.

```
# Replace regular Standard Errors with the
# heteroskedasticity-robust Standard Errors
se.model1 <- sqrt(diag(vcovHC(model1)))
se.model2 <- sqrt(diag(vcovHC(model2)))
se.model3 <- sqrt(diag(vcovHC(model3)))

# stargazer with all 3 models
stargazer(model1, model2, model3, title = "Regression Models",
    type = "text", report = "vcsp", omit.stat = "f", se = list(se.model1,
        se.model2, se.model3), star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Regression Models
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                        crmrte
##                           (1)             (2)             (3)
## ---------------------------------------------------------------
## prbarr                   -0.046          -0.062          -0.053
##                          (0.021)         (0.015)         (0.013)
##                         p = 0.033      p = 0.0001      p = 0.00005
##
## density                   0.007           0.005           0.006
##                          (0.001)         (0.001)         (0.001)
##                         p = 0.000      p = 0.00002     p = 0.0002
##
## polpc                     5.086           8.001           7.247
##                          (5.012)         (2.575)         (2.497)
##                         p = 0.311       p = 0.002       p = 0.004
##
## pctmin80                 0.0003          0.0004          0.0004
##                         (0.0001)        (0.0001)        (0.0001)
##                        p = 0.0003       p = 0.000      p = 0.00000
##
## prbconv                                  -0.021          -0.019
##                                          (0.005)         (0.006)
##                                        p = 0.00001      p = 0.002
##
## pctymle                                   0.060           0.092
##                                          (0.055)         (0.038)
##                                         p = 0.273       p = 0.015
##
## log(wcon)                                                 0.008
##                                                          (0.010)
##                                                         p = 0.394
##
## log(wtuc)                                                 0.003
##                                                          (0.008)
##                                                         p = 0.678
```

```
##
## log(wser)                                                  -0.005
##                                                            (0.017)
##                                                           p = 0.759
##
## log(wmfg)                                                  -0.002
##                                                            (0.008)
##                                                           p = 0.804
##
## log(wsta)                                                  -0.010
##                                                            (0.013)
##                                                           p = 0.446
##
## log(wloc)                                                   0.009
##                                                            (0.023)
##                                                           p = 0.701
##
## taxpc                                                      0.0002
##                                                           (0.0003)
##                                                           p = 0.568
##
## central                                                    -0.003
##                                                            (0.003)
##                                                           p = 0.292
##
## mix                                                        -0.021
##                                                            (0.021)
##                                                           p = 0.332
##
## prbpris                                                     0.001
##                                                            (0.015)
##                                                           p = 0.966
##
## avgsen                                                     -0.0005
##                                                           (0.0004)
##                                                           p = 0.294
##
## Constant                  0.020           0.028           0.002
##                          (0.010)         (0.008)         (0.145)
##                         p = 0.045       p = 0.001       p = 0.990
##
## ---------------------------------------------------------------------
## Observations               90              90              90
## R2                        0.669           0.814           0.850
## Adjusted R2               0.653           0.800           0.814
## Residual Std. Error 0.011 (df = 85) 0.008 (df = 83) 0.008 (df = 72)
## =====================================================================
## Note:                                   *p<0.05; **p<0.01; ***p<0.001
```

polpc as it can be seen in the table, this variable only became significant after more variables were introduced in model 2 and 3. This may be due to the omitted variable effect that prbconv had on polpc. We believe that higher prbconv would result in lower crmrte while higher polpc would result in higher prbconv. Therefore, the omitted variable bias would be negative and scale the OLS coefficient on polpc towards 0. This would cause the marginal effect of polpc to be underestimated. By introducing prbconv in model 2, we are negating

this omitted variable bias which is why `polpc` became statistically significant. The slope coefficient of `polpc` has a positive slope which suggests that counties with higher police per capita would have a higher crime rate, which seems to be counter intuitive. It is possible that counties with higher crime rates are hiring more policemen in hopes of decreasing crime, but the presence of these extra policemen does not have the impact on crime rate that they hoped. Therefore increasing the police force may not be the most effective policy to decrease the crime rate

`prbarr` became more significant, and the slope coefficient became more negative in model 2 and model 3. This may be due to the omitted variable effect that `prbconv` had on `prbarr`. We believe that higher `prbconv` would lead to lower `crmrte`, while higher `prbarr` would result in lower `prbconv`. This can be interpreted that being overzealous in arrest may lead to a lower conviction rate. Therefore, the omitted variable bias would be positive and scale the OLS coefficient on `prbarr` towards 0. This would cause the marginal effect of `prbarr` to be underestimated. By introducing `prbconv` in model 2, we are negating this omitted variable bias which is why `prbarr` became more statistically significant. So practically, the model suggests that we need to reach a level of balance between arrests and conviction.

`prbconv` when introduced to the model, this variable was significant, and it has a negative slope. This indicates that a higher probability of conviction results in a lower crime rate. Practically, this means that a higher likelihood of conviction may deter people from comitting crime.

`pctymle` this variable did not become significant until model 3, when additional variables were introduced. This may be due to the nuance that the wage variables have, and how they might impact the young men population. Practically, this can be interpreted as: when wages are low and there is a high percentage of young men in a county, then the young men may have more incentive to commit crimes. There may be an omitted variable effect on the `pctymle` by the wages variables. Upon closer inspection of the various wage variables, `wsta`, `wser` and `wmfg` while not significant, do have negative slopes. From a practical stand point,perhaps young men tend to have state, service and/or manufacturing jobs, and the higher the wages for those three industries, the less incentive there is for the young men to commit crime.

4.4 Model Selection and Assessment

To DO: We need justification of why model 2 was chosen TO DO: detailed assessment of the 6 CLM TO DO: other diagnostic tools to assess whether the assumptions appear to be violated. TO DO: create model 4, and discuss why it is not worth it to add additional variables to the model (i.e. because the model do not improve in a statistically significant way), this may be due to omitted variable bias from a variable that is not included in the data set, and then segway from here to the omitted variable discussion.

Justification To help select the best model, we ran a joint significance test to to see if the variables that were added for model 2 and model 3 improved the regression in a statistically significant way.

```
# joint significance between model1 and model2
waldtest(model1, model2, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: crmrte ~ prbarr + density + polpc + pctmin80
## Model 2: crmrte ~ prbarr + density + polpc + pctmin80 + prbconv + pctymle
##   Res.Df Df      F    Pr(>F)
## 1     85
## 2     83  2 16.467 9.483e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# joint significance between model2 and model3
waldtest(model2, model3, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: crmrte ~ prbarr + density + polpc + pctmin80 + prbconv + pctymle
## Model 2: crmrte ~ prbarr + density + polpc + pctmin80 + prbconv + pctymle +
##     log(wcon) + log(wtuc) + log(wser) + log(wmfg) + log(wsta) +
##     log(wloc) + log(wser) + taxpc + central + mix + prbpris +
##     avgsen
##   Res.Df Df      F Pr(>F)
## 1     83
## 2     72 11 0.8457 0.5958
```

The addition of the `pctymle` and `prbconv` in model 2 improved the model in a statistically significant way. However the addition of variables in model 3 are not jointly significant when compared against model 2.

the AIC test, reveals that model 2 has the lowest AIC values, indicating that it is the best fitting model for the data set among the three models that were created.

Based on the Shapiro-Wilkes test, we can reject the null hypothesis of normal distribution for model 1, but not for model 2. While model 1 fails the normality assumption, model 2 does not.

```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.93859, p-value = 0.0003529
```

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97587, p-value = 0.09232
```

Model 2 also has the best standardized residual vs. leverage plot, where it has the least amount of data points that have standardized residuals greater than 2. In addition, unlike model 1 and 3, model 2 do not have any data points that has Cook's distance greater than 1 as it can be seen in the Residuals vs. Leverage plots above. This indicates that model 2 is not strongly influenced by any particular data point.

CLM Assumption Diagnostics

Here we are doing a detailed evaluation of our chosen model (model 2) against the 6 CLM Assumptions.

CLM 1. Linear population model: We do not have to worry about this assumption at the moment because we haven't constrained the error term.

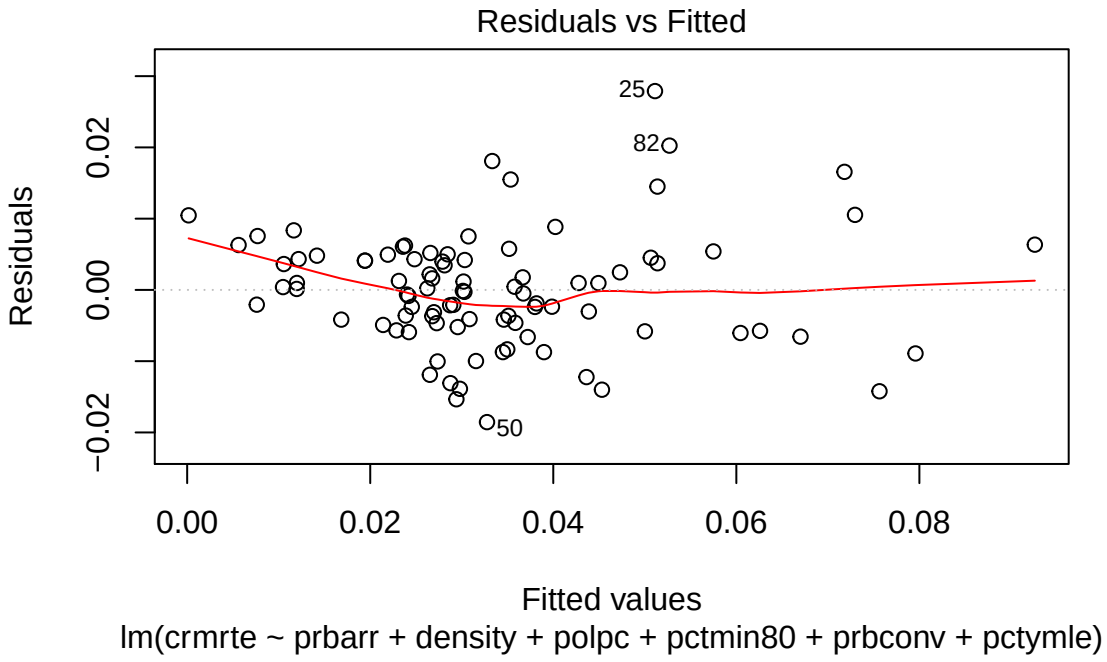
CLM 2. Random Sampling: To check random sampling, we need domain knowledge and an understanding of how the data were collected. There are 100 counties in North Carolina, and there are data for 91 of them. Without knowledge of the 9 excluded counties, no statement regarding the validity of random sampling can be made.

CLM 3. No perfect multicollinearity: There is no need to explicitly check for perfect collinearity, because R would've reported a warning if this occurred. Furthermore, the correlation matrix shown in section 3.2 also shows that there is no perfect collinearity.

CLM 4. Zero Conditional Mean: $E(u|x) = 0$. For this model, the residuals vs. fitted values plot shown below reveals a relatively flat spline centered around zero for most of the fitted values. However, looking at the plot closely the zero conditional mean does not hold for the majority of the fitted values. This may be

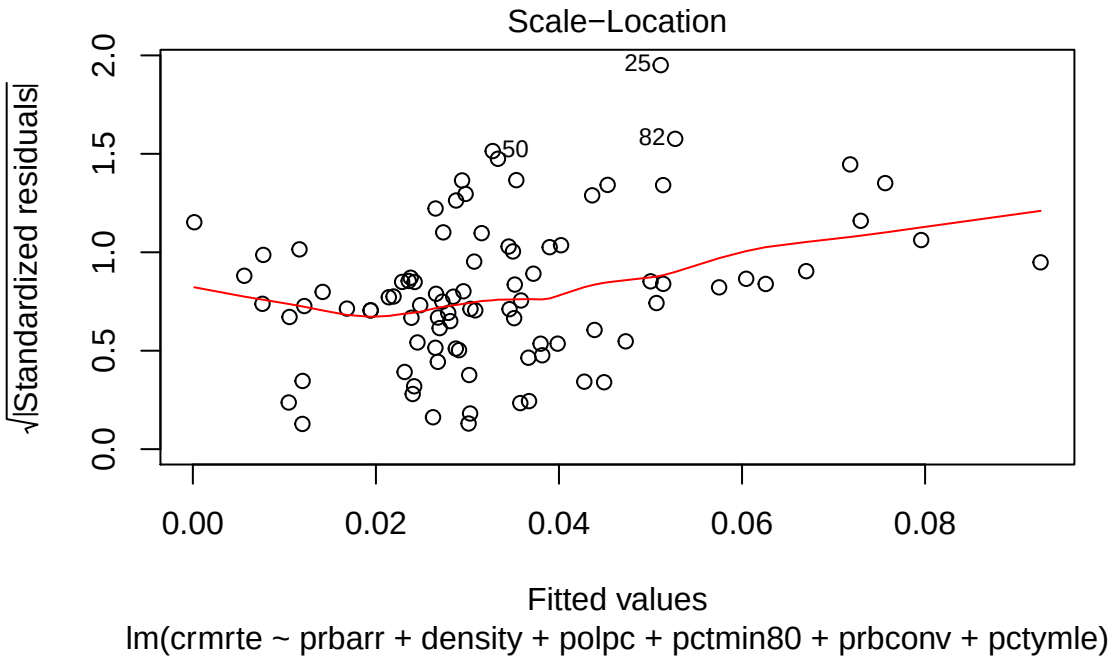
caused by biases introduced by the additional variables or non-linear relationship between the new variables and the output variable. We may need to consider investigating and removing the source of the bias.

```
# Residuals vs. Fitted Plot
plot(model2, which = 1)
```



CLM 5. Homoscedasticity: In the residuals vs. fitted values plot shown in **CLM 4**, the data points seem to form a cone shape which suggests some heteroscedasticity. In the scale-location plot below, there seems to be a slight positive slope across the range of fitted values. Furthermore, the Breusch-Pagan test shown below has a p-value of 0.0004972 which indicates that the null hypothesis of homoscedasticity can be rejected. However, heteroscedastic-robust standard errors were used When evaluating the statistical significance with this model, so the heteroscedacity may not be an issue.

```
# Scale-Location Plot
plot(model2, which = 3)
```

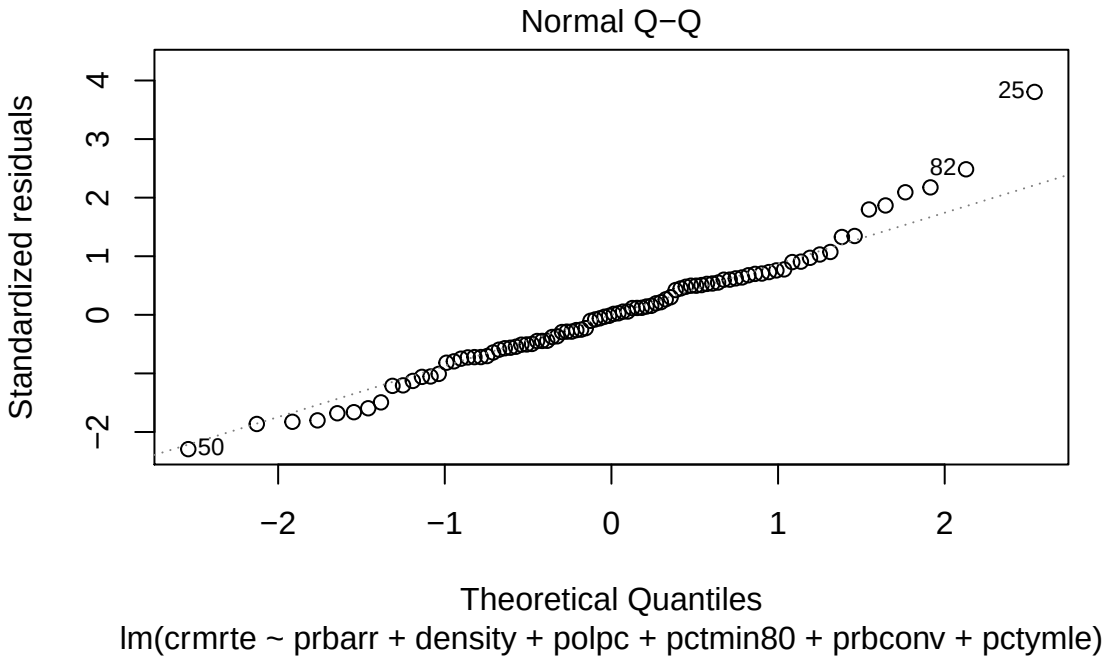



```
# Breusch-Pagan
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 23.399, df = 6, p-value = 0.0006734
```

CLM 6. Normality of errors: In the Q-Q plot shown below, the bulk of the error terms seem to follow the straight line which suggests a fairly normal distribution. The standardized residuals show some deviation from the straight line at the high end of the distribution. This suggests some skew at the high end of our residuals. However, the Shapiro-Wilk test shown below has a p value of 0.05755 which means we cannot reject the null hypothesis of the residuals having a normal distribution.

```
# Q-Q plot of Standardized Residuals
plot(model2, which = 2)
```

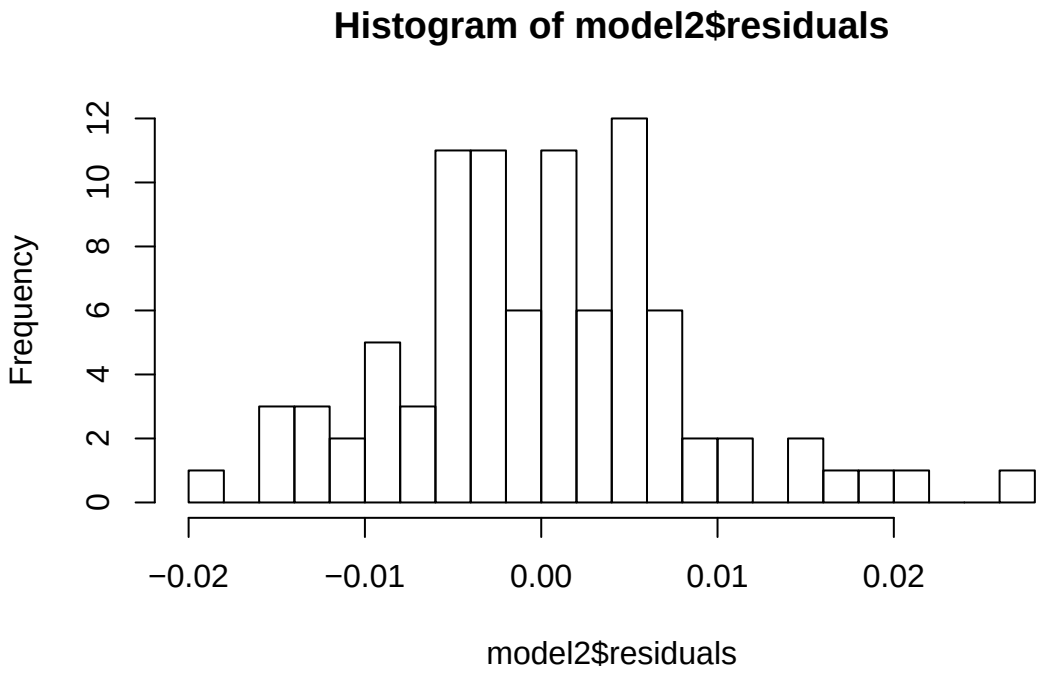



```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97587, p-value = 0.09232
```

To further verify this observation, a histogram of this model's residuals is shown below. The histogram shows approximate normality near the center of the distribution, but also some evidence of skewness on the positive end. However, the Central Limit Theorem (CLT) claims that if the sample size is large enough we can assume that the residuals have a normal sampling distribution. For distributions with a very strong skew, a much larger sample size may be required, but for minor skews as in this case, the rule of thumb is that the CLT can be applied when the sample size is greater than 30. The sample size used for this model is 91 which should be enough for the CLT to hold.

```
hist(model2$residuals, breaks = 20)
```



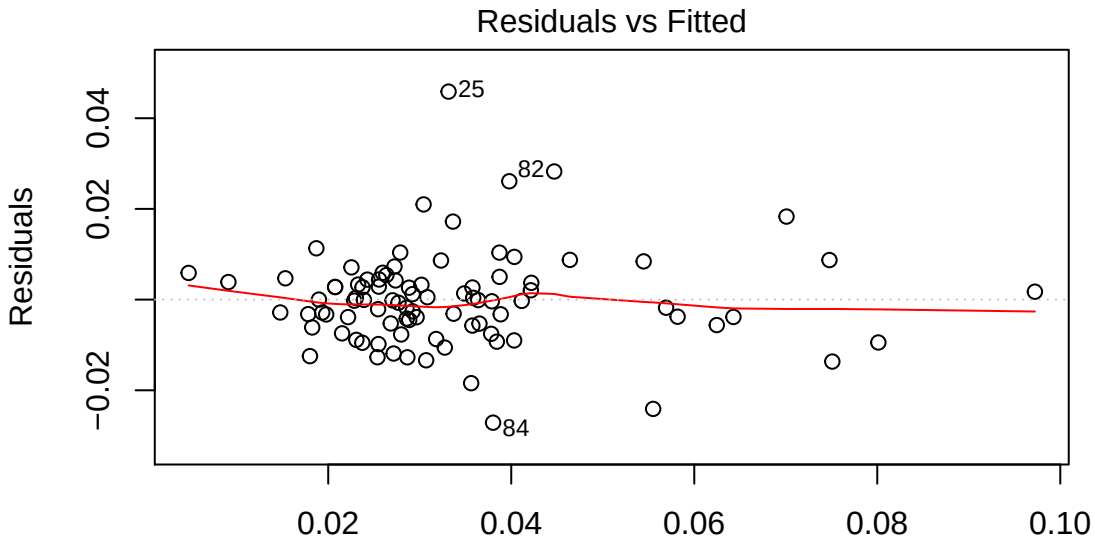
Based on our review of the six CLM assumptions, model 2 may not be considered an unbiased estimator, especially since it violates the zero conditional mean. To draw inferences on the coefficients, we replaced the regular standard errors with the heteroskedasticity-robust standard errors.

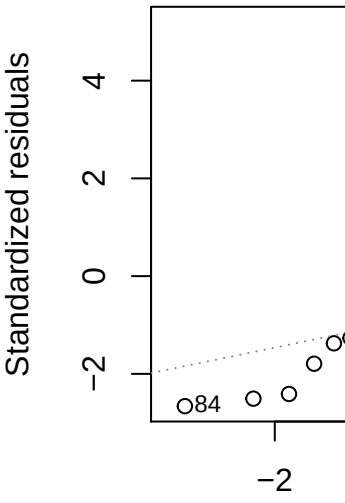
Adjusting Model Specifications After reviewing the 6 CLM assumptions, the zero-conditional mean assumption was violated, upon investigation, we found that the addition of `prbconv` introduced some bias when `prbarr` is also included in the regression model. So to improve the model and meet the zero-conditional mean assumption, we decided to remove `prbconv` from the regression model, and created model 4 below

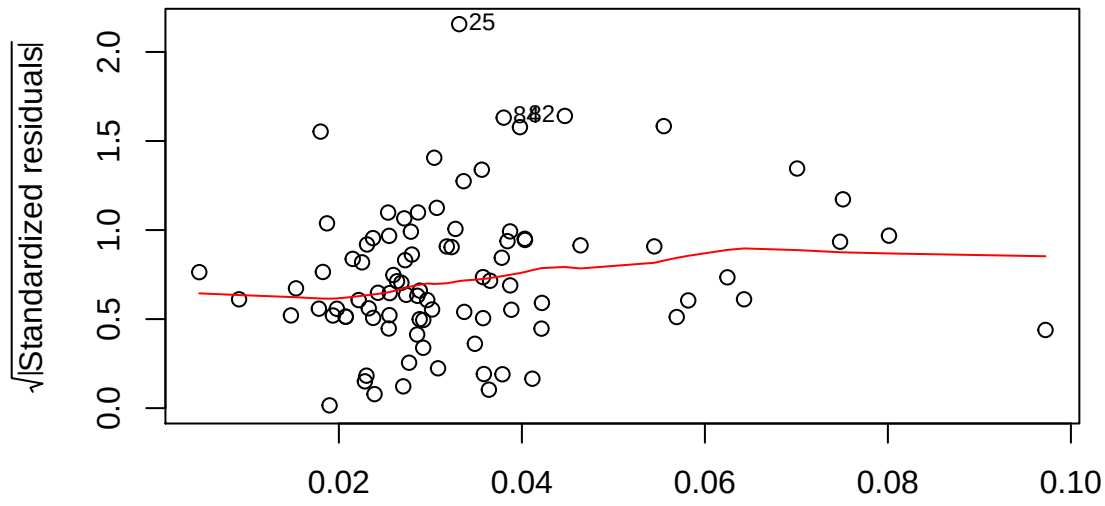
```
model4 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + pctymle,
    data = crime_df)

waldtest(model1, model4, vcov = vcovHC)
```

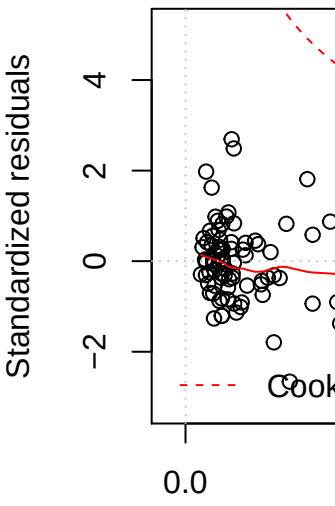
```
## Wald test
##
## Model 1: crmrte ~ prbarr + density + polpc + pctmin80
## Model 2: crmrte ~ prbarr + density + polpc + pctmin80 + pctymle
##   Res.Df Df      F   Pr(>F)
## 1     85
## 2     84  1 7.1316 0.009092 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model4)
```









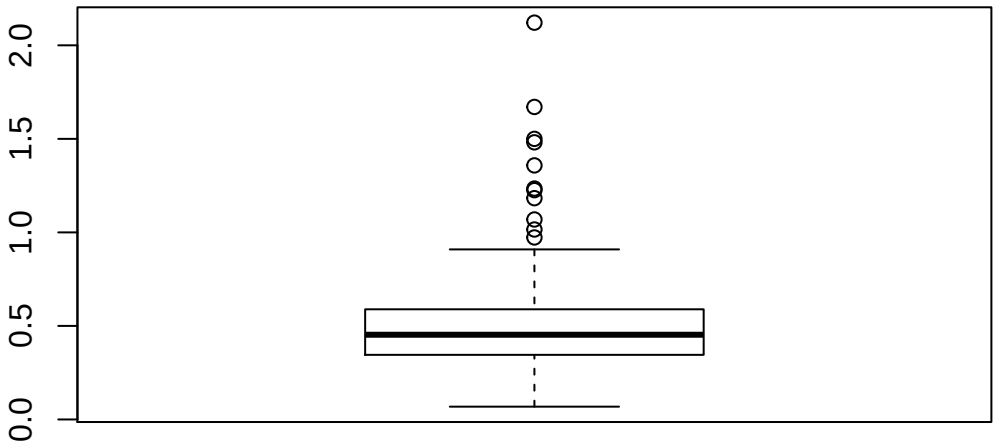

```
crime_df[crime_df$prb_conv > 0.65, ]
```

```
##  [1] county   year     crmrte   prbarr   prbconv  prbpris  avgsen
##  [8] polpc    density  taxpc    west     central  urban    pctmin80
## [15] wcon     wtuc     wtrd     wfir     wser     wmfg     wfed
## [22] wsta     wloc     mix      pctymle
## <0 rows> (or 0-length row.names)
```

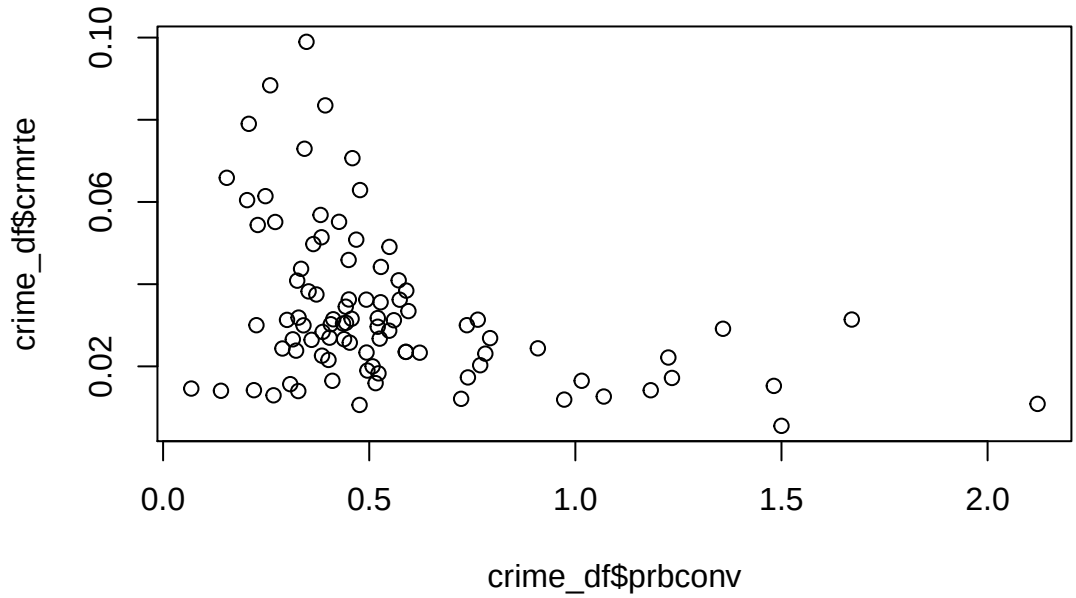
```
crime_df$prb_conv > 0.65
```

```
## logical(0)
```

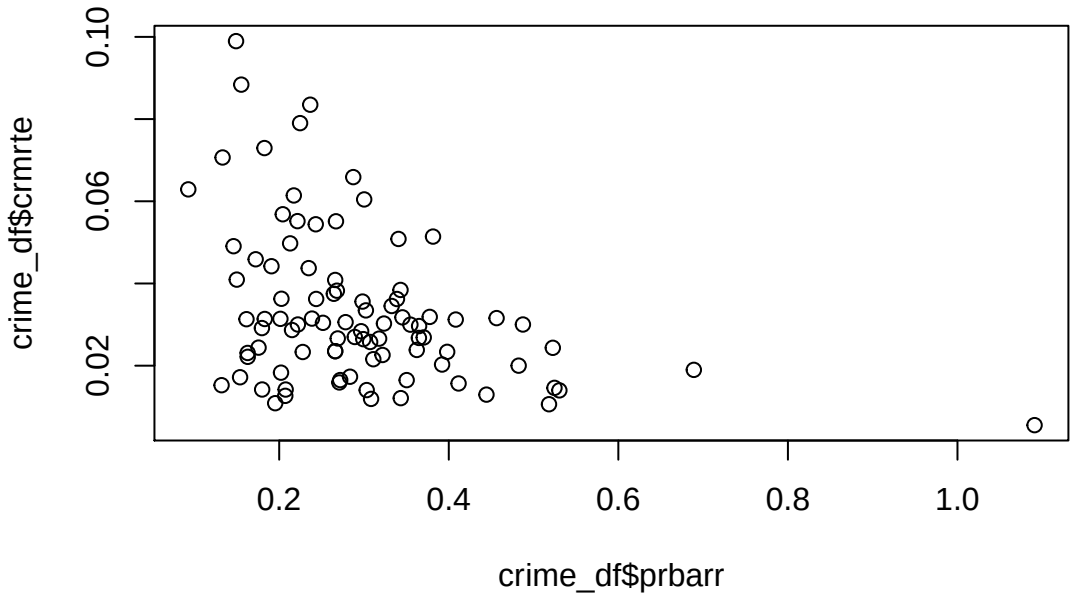
```
boxplot(crime_df$prbconv)
```




```
plot(crime_df$prbconv, crime_df$crmrte)
```




```
plot(crime_df$prbarr, crime_df$crmrte)
```




```
crime_df[51, ]
```

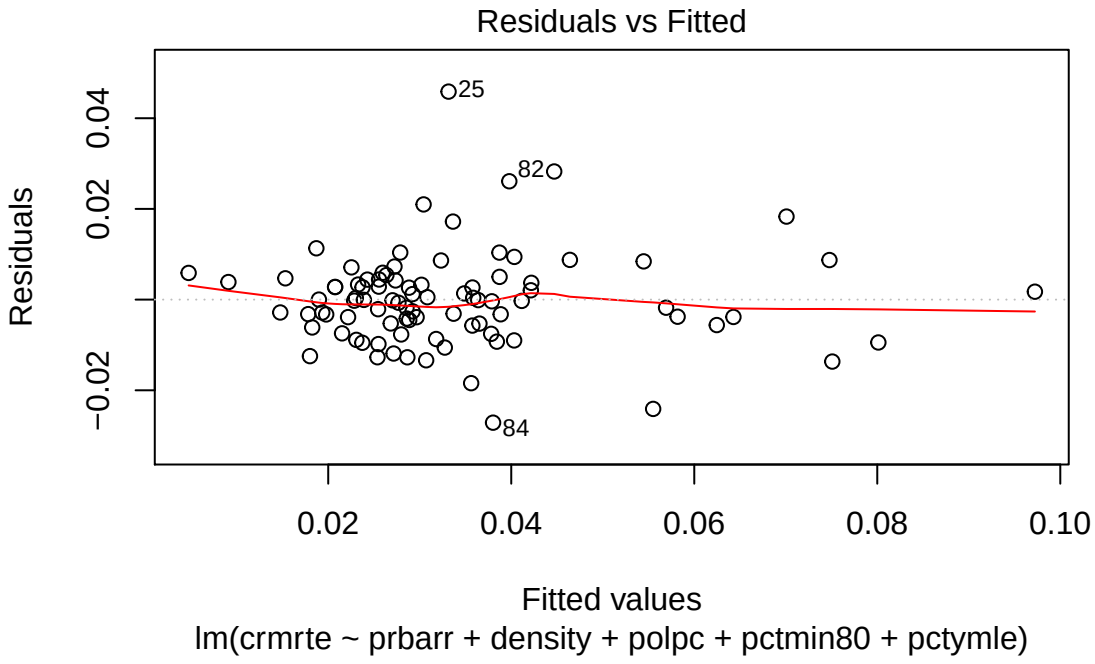
```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

```
summary(crime_df$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

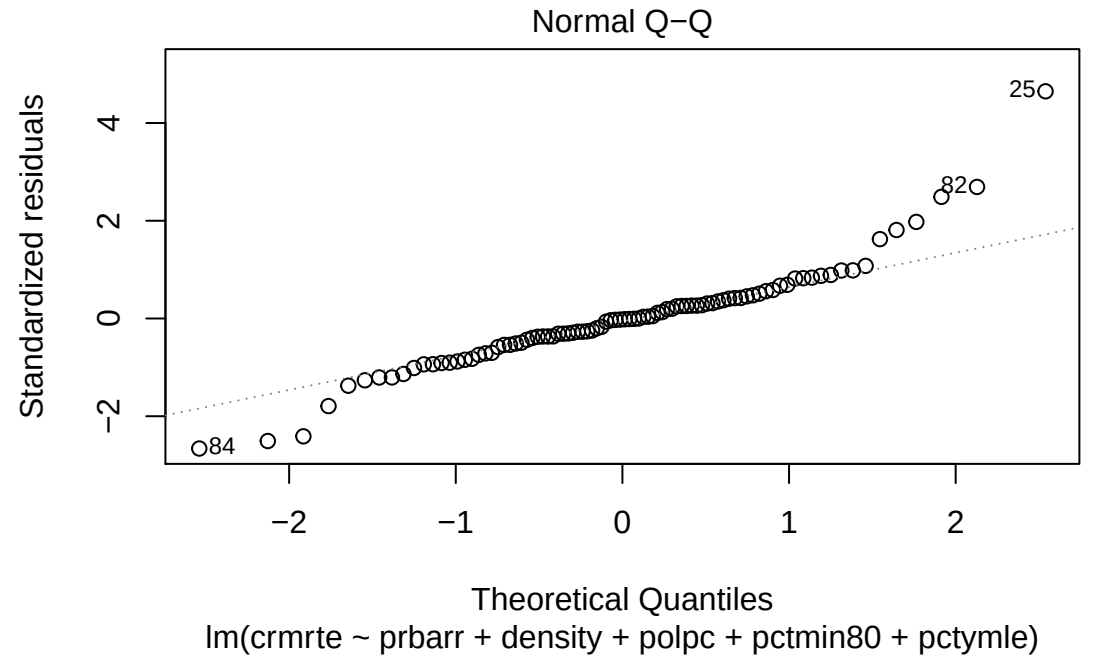
Upon inspection of the residual vs fitted plot below , it appears that model 4 does not violate the zero conditional mean assumption.

```
plot(model4, which = 1)
```



However, as a tradeoff, model 4's Normal Q-Q plot shows that our residuals have lost normality. We confirmed this by comparing the shapiro test of our two models. As shown below, the p-value of model 4 is 4.755e-05, which means we can reject the null hypothesis of normal error distribution, while the p-value for model 2 is 0.05755, which means we cannot reject the null hypothesis. However, since we have a sample size of 91 we can invoke the CLT to rely on the asymptotic properties of OLS.

```
plot(model4, which = 2)
```




```
shapiro.test(model4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
```

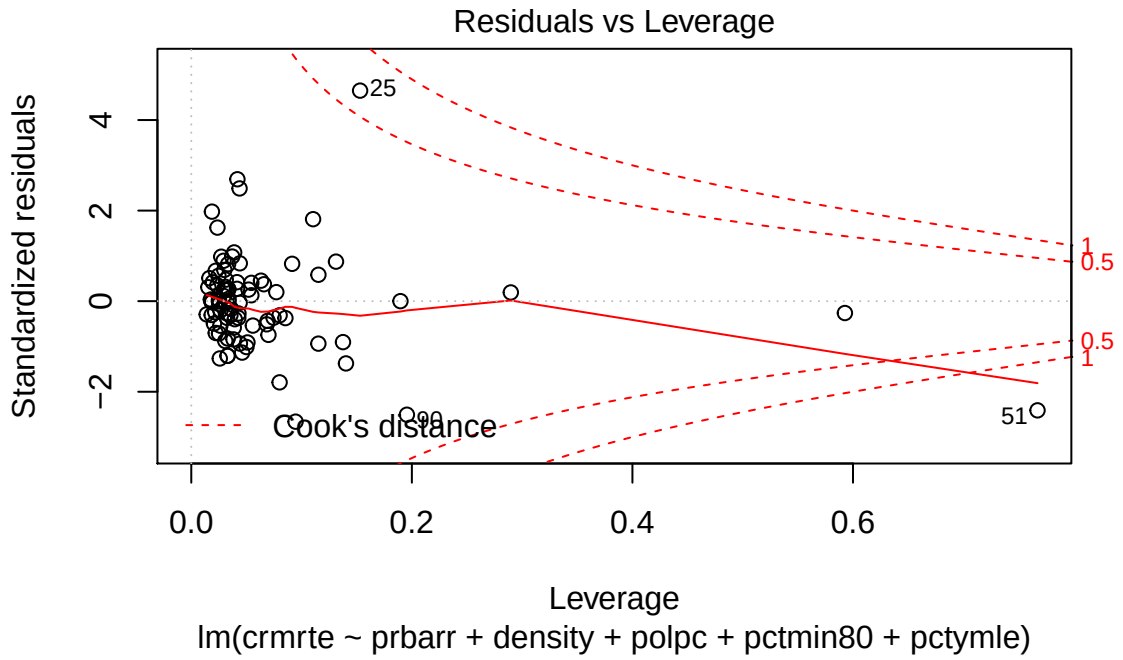
```
## data:  model4$residuals
## W = 0.92493, p-value = 6.399e-05
```

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97587, p-value = 0.09232
```

Also, another trade off is that we now have a highly influential data point (point #51), as we can see below in the Residuals vs. Leverage plot. Lastly, we also increased the number of standardized residuals that are greater than 2, indicating that model 4 is not as good of a represantation of the data set compared to model 2.

```
plot(model4, which = 5)
```



Despite the tradeoffs mentioned above, we think it is more important to meet the zero-conditional mean assumption, since that is the strictest assumption among the 6 in CLM that we need to meet in order to get a unbiased estimator.

Below is the final heteroscedastic robust coefficient test, and the significant coefficients can be interpreted as follows: - For every percentage point increase in increase in probability arrest, crime rate decreases by 0.0004. - For every 1 additional person per square mile, crime increases by .007 crimes per person. - For every percentage point increase in minority population, crime increases by .0003 crimes per person. - For every percentage point increase in young male population, crime increases by 0.00129 crimes per person.

```
# Replace regular Standard Errors with the
# heteroskedasticity-robust Standard Errors
se.model2 <- sqrt(diag(vcovHC(model2)))
se.model4 <- sqrt(diag(vcovHC(model4)))

# stargazer with all 3 models
stargazer(model2, model4, title = "Regression Models", type = "text",
    report = "vcsp", omit.stat = "f", se = list(se.model2, se.model4),
```

```
    star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Regression Models
## ================================================
##                          Dependent variable:
##                     ----------------------------
##                                 crmrte
##                           (1)             (2)
## ------------------------------------------------
## prbarr                   -0.062          -0.041
##                          (0.015)         (0.022)
##                         p = 0.0001       p = 0.059
##
## density                   0.005           0.007
##                          (0.001)         (0.001)
##                         p = 0.00002      p = 0.000
##
## polpc                     8.001           4.607
##                          (2.575)         (4.930)
##                          p = 0.002       p = 0.350
##
## pctmin80                 0.0004          0.0003
##                         (0.0001)        (0.0001)
##                          p = 0.000      p = 0.0002
##
## prbconv                  -0.021
##                          (0.005)
##                         p = 0.00001
##
## pctymle                   0.060           0.128
##                          (0.055)         (0.048)
##                          p = 0.273       p = 0.008
##
## Constant                  0.028           0.008
##                          (0.008)         (0.010)
##                          p = 0.001       p = 0.405
##
## ------------------------------------------------
## Observations               90              90
## R2                        0.814           0.693
## Adjusted R2               0.800           0.675
## Residual Std. Error 0.008 (df = 83) 0.011 (df = 84)
## ================================================
## Note:                 *p<0.05; **p<0.01; ***p<0.001
```

5.0 Omitted Variables Discussion

Even by including all of the relevant variables provided in the data set to the linear regression, the resulting model may still be biased. This is because of potentially influential omitted variables that is either not provided, or is difficult to obtain. Below are some of the omitted variables that might be important along with how their absense may affect our results.

Potential Omitted Variable #1: Financial Wellfare (Poverty Rate and Unemployment)

We believe that an important driver of crime rate is the financial wellfare of the people. The following equations can help us determine how the omitted variable bias would impact our density coefficient:

$$crmrte = \beta_0 + \beta_1 * density + \beta_2 * poverty\_rate + u$$

$$poverty\_rate = \alpha_0 + \alpha_1 * density + u$$

We believe that higher poverty and unemployment rates will result in higher crime rate ($\beta_2 > 0$) as people are more desperate and will resort to crime in order to survive. Furthermore, in areas of high population density, there may be fewer jobs available as well as a higher poverty rate ($\alpha_1 > 0$). Therefore, the omitted variable bias ($\beta_2\alpha_1$) for both poverty rate and unemployment rate is positive, scaling the OLS coefficient on `density` away from zero (more positive). In other words, the marginal effect of `density` on crime rate may be overestimated, resulting in a magnified statistical significance.

Using the same analysis method on `pctmin80`, we theorize that in 1987, minorities tend to be more impoverished than their counterparts. Therefore, a larger percentage of minorities in a county, will likely result in higher poverty and unemployment rates ($\alpha_1 > 0$). In fact, we believe that there is a strong marginal effect of `pctmin80` on poverty rate, which means the omitted variable bias of poverty rate and unemployment rate would scale the OLS coefficient on `pctmin80` by a relatively large amount. This means that the marginal effect and statistical significance of `pctmin80` on crime rate may be highly overestimated.

The tax revenue and various wage variables may help proxy these two omitted variables. However, we believe they are not very accurate proxies, especially for unemployment rate, because the unemployed are not paying income tax and do not have any wages at all.

Potential Omitted Variable #2: Percent of Arrests Driven by Discrimination

In 1987, and arguably even today, discrimination has unfortunately played a big role in the incarceration of certain minority groups. This can come in the form of false arrests, or disproportionate arrests for petty crimes and misdemeanors in minority communities. The higher the number of arrests driven by discrimination, the higher the reported crime rate would be ($\beta_2 > 0$). Furthermore, the higher the percentage of minorities in a county, the higher the number of arrests driven by discrimination would be ($\alpha_1 > 0$). Therefore, the omitted variable bias ($\beta_2\alpha_1$) of discrimination is positive, which would scale the OLS coefficient on `pctmin80` away from zero (more positive). This means that the marginal effect and statistical significance of `pctmin80` on crime rate may be overestimated.

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * discrimination$$

$$discrimination = \alpha_0 + \alpha_1 * pctmin80$$

In addition, we believe that counties with a higher "probability" of arrest would also have a higher number of arrests driven by discrimination ($\alpha_1 > 0$). Therefore, the omitted variable bias is also positive in this case, which would scale the slope coefficient on `prbarr` away from zero (more positive). Therefore, the marginal effect and statistical significance of `prbarr` on crime rate may also be overestimated by omitting the effect of discrimination in the model.

The number of arrests driven by discrimination is very difficult to measure because very few policemen would admit to doing such a thing. Therefore, we unfortunately do not have any proxy variables to represent this omitted variable, except maybe `pctmin80`. However, using `pctmin80` as a representation of discrimination would be imperfect and making a lot of broad assumptions.

Potential Omitted Variable #3: Family Heath (Number of Parents, Amount of Abuse/Neglect, Availability of Positive Role Models)

Another potentially strong influence on crime is family health. This can be possibly represented by the number of parents an individual while growing up in a household, the level of abuse and neglect that the

individual suffers, and the availability of positive role models in the individual's life, among other things. There are so many complicated aspects to family health that it would be hard to accurately predict the effects of this omitted variable on the OLS coefficients. For the sake of simplicity, we will only explore the effects of having a two parent household as our omitted variable.

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * pct\_of\_2parents\_hh$$
$$pct\_of\_2parents\_hh = \alpha_0 + \alpha_1 * pctmin80$$

We do not have a concrete understanding of whether children from two parent households are less likely to commit crime than children in single-parent households and orphans. Our subjective assumption is that it might be easier for two parents to provide good care for a child. For example, with two providers, the child would be less likely to live in poverty as well as possibly have more quality time with at least one of the parents. Therefore, the larger the percentage of two-parent households in a county, the lower the crime rate may be ($\beta_2 < 0$). According to kidscount.org, the percentage of African American children in single-parent households is 3 times larger than the percentage of Caucasian children in single-parent households in the State of North Carolina in 2005. Extrapolating from this, we will assume that counties with higher `pctmin80` would have lower percentage of two-parent households ($\alpha 1 < 0$). Therefore, the omitted variable bias is positive, which would scale the slope coefficient on `pctmin80` away from zero (more positive). Thus, the marginal effect and statistical significance of `pctmin80` on crime rate may also be overestimated by omitting the effect of two-parent households in the model.

Potential Omitted Variable #4: Percentage of Highschool Graduates

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * pct\_hs\_graduates$$
$$pct\_hs\_graduates = \alpha_0 + \alpha_1 * pctmin80$$

The average years of education in a county may also be an important factor that influences crime rate. We assume that more graduation from highschool would result in a higher chance of employment at a higher paying job. Furthermore, time spent in school at a young age is believed to keep children out of trouble and away from bad influences. Therefore, a county with a higher percentage of highschool graduates may possibly have a lower crime rate ($\beta_2 < 0$). According to governing.com, the North Carolina highschool graduation rate of African Americans is 10% lower than the highschool graduation rate of Caucasians in 2011. By extrapolating this information, we assume that counties with higher percentage of minorities will have a lower percentage of highschool graduates ($\alpha_1 < 0$). This means that the omitted variable bias is positive, which would scale the slope coefficient on `pctmin80` away from zero (more positive). Thus, the marginal effect and statistical significance of `pctmin80` on crime rate may also be overestimated by omitting the effect of education in the model.

Potential Omitted Variable #5: Rate of Drug Abuse

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * drug\_abuse$$
$$drug\_abuse = \alpha_0 + \alpha_1 * pctmin80$$

The last omitted variable we considered is the Rate of drug abuse as an indicator of crime. We assume that counties with a higher rate of drug abuse would also have a higher crime rate because the usage of illegal drugs is a crime ($\beta_2 > 0$). The war on drugs lead to a vastly disproportionate rate of arrest in minority groups, which means that in 1987, it may be that a county with a larger percentage of minorities would have a higher rate of substance abuse related arrests (α_1). This means that the omitted variable bias is positive, which would scale the slope coefficient on `pctmin80` away from zero (more positive). Thus, the marginal effect and statistical significance of `pctmin80` on crime rate may also be overestimated by omitting the effect of substance abuse in the model.

Something that strongly stands out in all of these omitted variables is that the OLS coefficient on `pctmin80` is strongly impacted by omitted variable bias. It is possible that the marginal effect of the percentage of minorities on crime rate is entirely an artifact of omitted variable bias.

6.0 Conclusion

The best model for crime rate, have the following coefficients: * prbarr * density * polpc * pctmin80 * pctymle Out of these coefficients, `polpc` was not significant, while the others were. As discussed in section **TO DO** police per capita has a positive slope coefficient with crmrte which is counterintuitive. We believe this is an artifact of counties of higher crime rate are hiring more police officers to combat crime. Therefore a policy of "more cops on the street" may not be very effective. However, higher "probability" of arrest may lead to lower crime rate. Therefore the campaign may want to pursue a policy of strict enforcement of the law by police officers to increase "probability" of arrest to help drive down crime rate.

The other variables (density, pctymle, pctmin) however cannot be directly affected by a policy, but these variables are affected by the ommited variables that were discussed above, which gives us an idea where to focus our efforts. For example, we may want to create a job placement program in dense areas with high percentage of minority population, so that we can decrease unemployment rate and poverty rate and therefore reduce crime rate. Additionally, we may want to do mentorship programs in areas with high percentage of minority that have young males who lack positive role models and have a low high school graduation rate in hopes of reducing crime rate. Lastly, we may want to pursue an anti drug campaign to reduce crime.