W203 Lab 3

Armand Kok, Adam Yang, James De La Torre

Introduction

Is the introduction clear? Is the research question specific and well defined? Could the research question lead to an actionable policy reccomendation? Does it motivate the analysis? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.

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 she 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`, while 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: Look for any top-coding or bottom coding. TO DO: remove this instructiuon line **Did the team notice any anomalous values? Is there a sufficient justification for any data points that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Overall, does the report demonstrate a thorough understanding of the data?**

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

Variable	Label	Variable	Label
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. For example,* `probconv` *is the ratio of the number of convictions to the number of arrests, which can be larger than 1.*

2.1 Loading the Data

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

```r
# 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
```

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

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

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

```r
# TO DO: Eliminate these commented lines Adam's dir mydir <-
# '/Users/adamyang/Desktop/w203/Lab3/w203-Lab3/' Armand's dir
# mydir<-'C:/Users/ak021523/Documents/GitHub/mids-repos/W203/Homework/w203-Lab3/'
# jim's directory mydir<-
# 'F:/users/jddel/Documents/DATA_SCIENCE_DEGREE_LAPTOP/W203_Stats/Lab_03/'

# 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 data cleanup steps were required.

- 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.

```
# Summary was performed to understand data.  Will not include
# in report text, however. summary(crime_df) str(crime_df)


# 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))
```

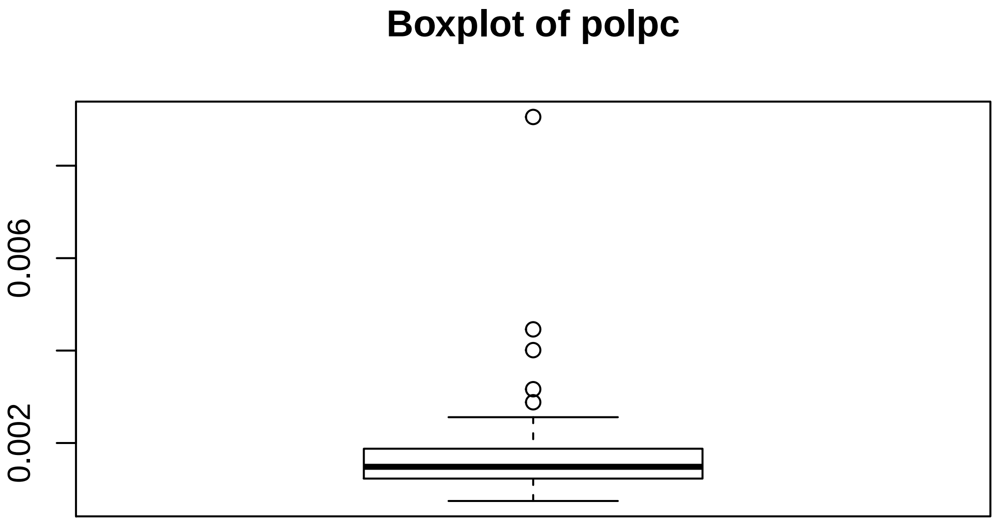
2.3 Outlier Identification

TO DO: Write function that computes outliers by column TO DO: Do we really want to mention this loose definition of outliers? We likely have dozens of them.

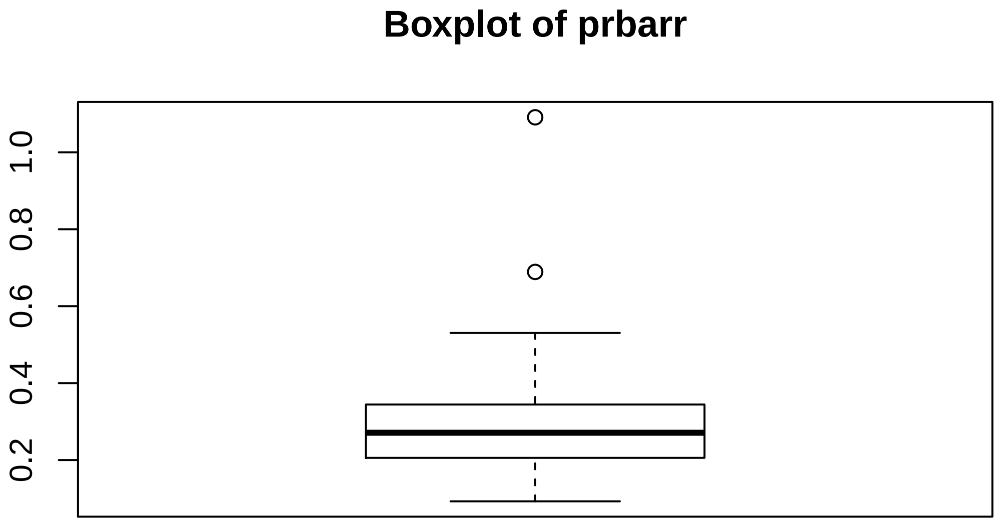
After generating histograms to review the distributions of the different variables, four were found to have outliers (an outlier is defined as a data value that is more than Q3 + 1.5 IQR or Q1 - 1.5 IQR): - polpc - row 51 - prbarr - row 51 - wser - row 84 - taxpc row 25

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

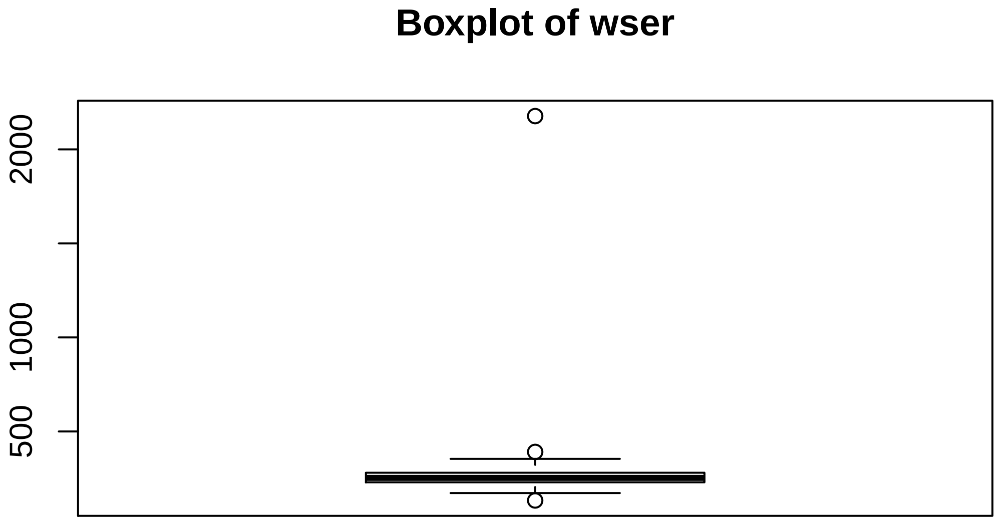
```
boxplot(crime_df$polpc, main = "Boxplot of polpc")
```



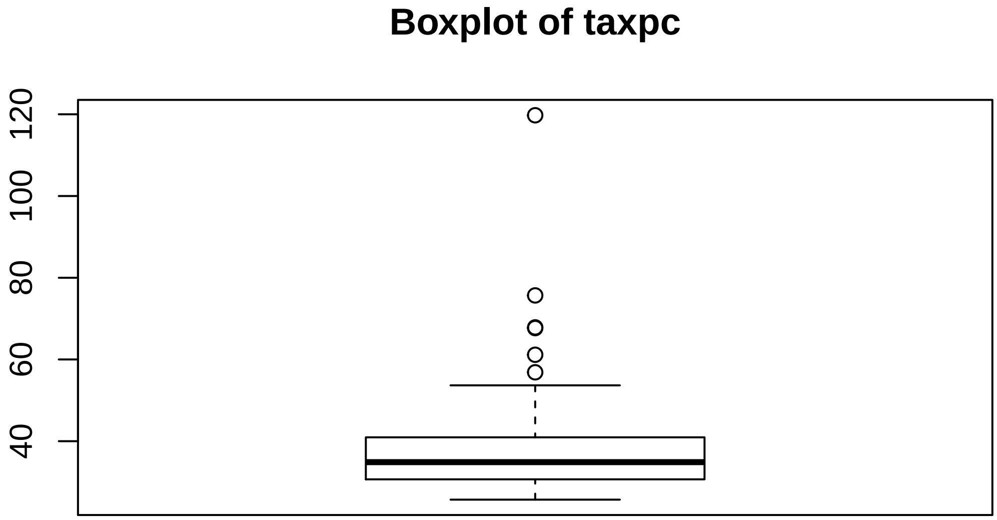
```
boxplot(crime_df$prbarr, main = "Boxplot of prbarr")
```



```
boxplot(crime_df$wser, main = "Boxplot of wser")
```



```
boxplot(crime_df$taxpc, main = "Boxplot of taxpc")
```



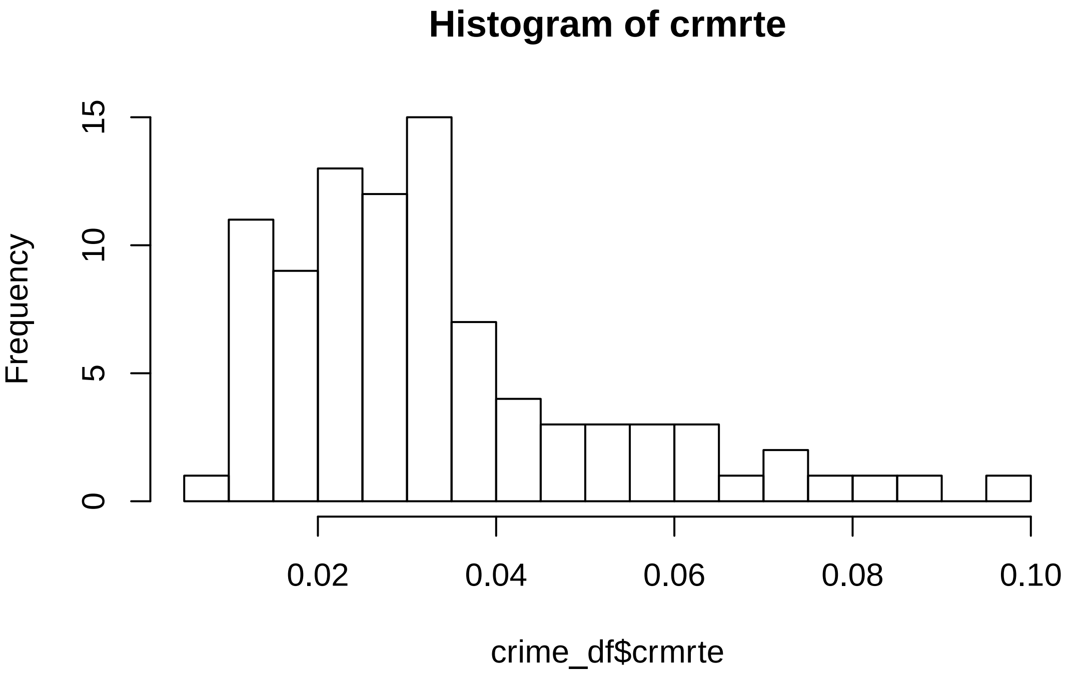
3.0 Model Building Process

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.

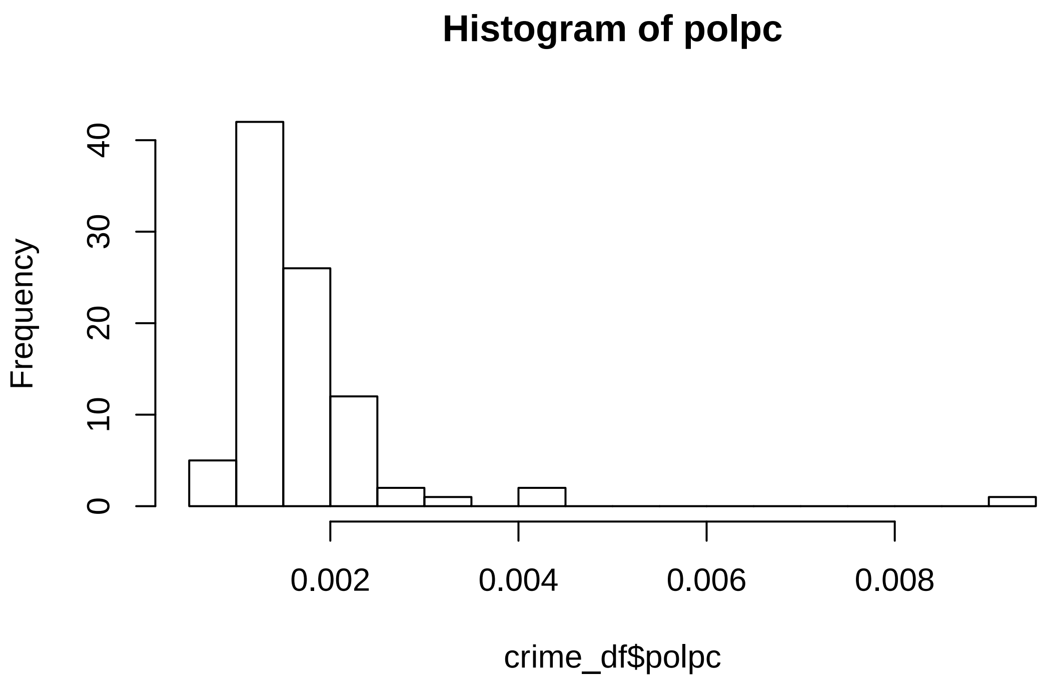
```
hist(crime_df$crmrte, main = "Histogram of crmrte", breaks = 30)
```



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

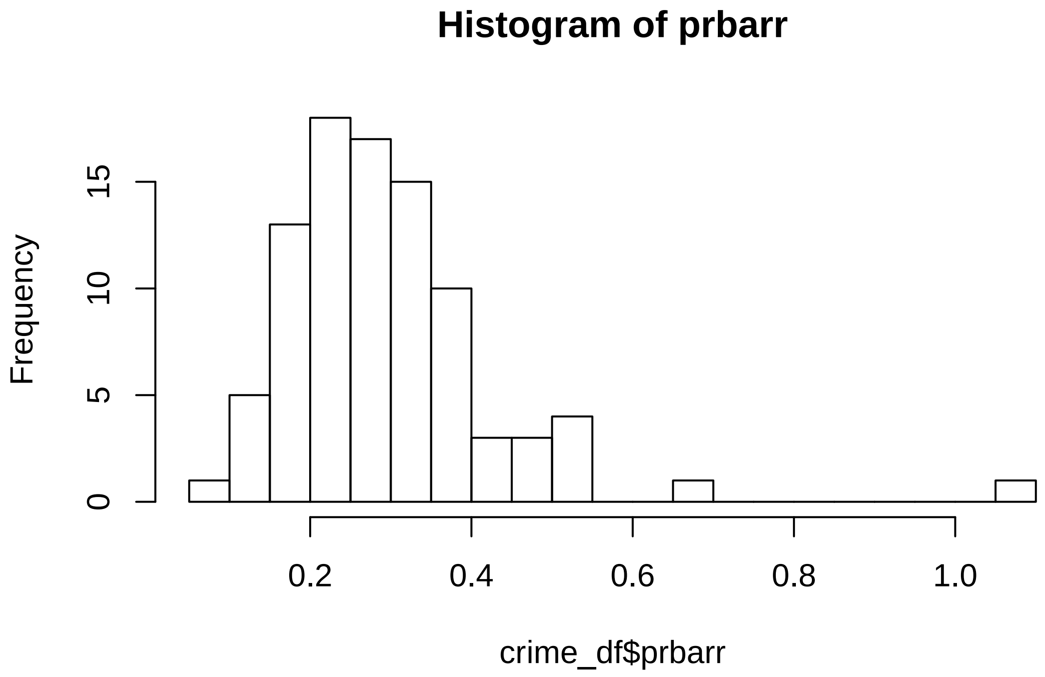
Given that our candidate is considering policies involving increasing the size of the police force, instituting stricter arrest protocols, and addressing issues of minorities in the inner cities, the police per capita (`polpc`), probability of arrest, (`prbarr`), population per square mile (`density`), 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)
```



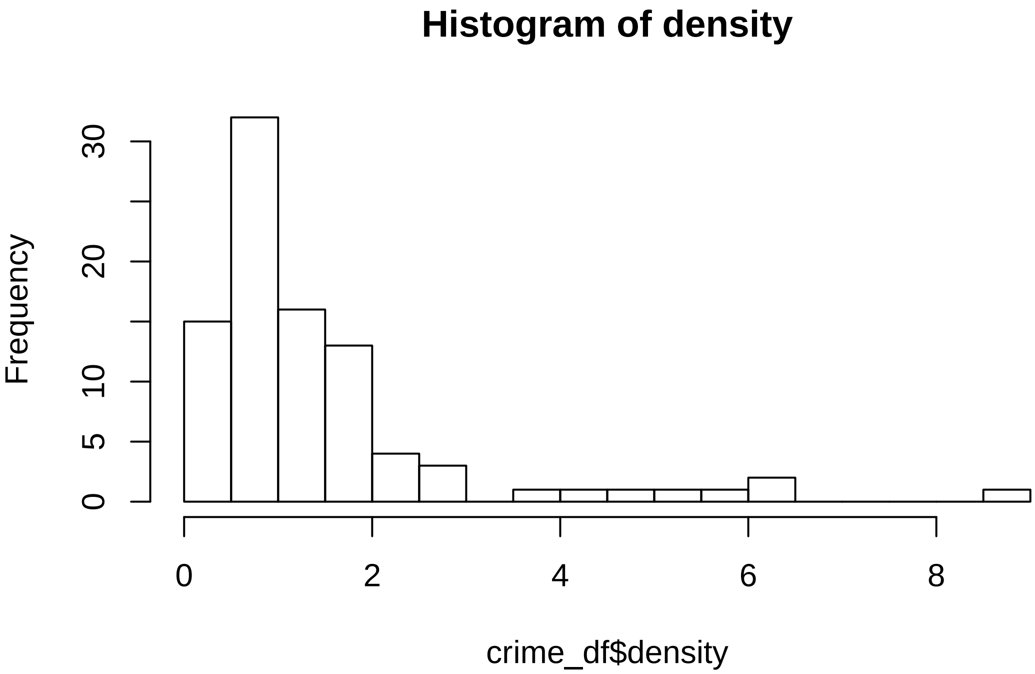
The histogram of `polpc` shows the outlier point mentioned in the outlier identification section.

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



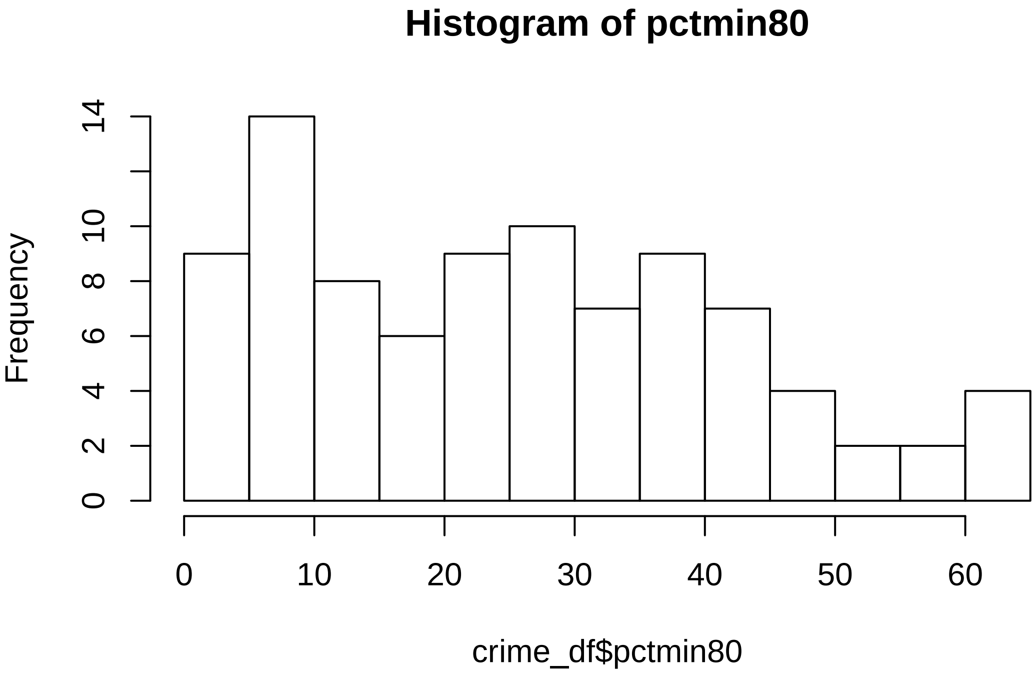
The histogram of `prbarr` also shows an outlier point, which is the same outlier record for `polpc`.

```
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.

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



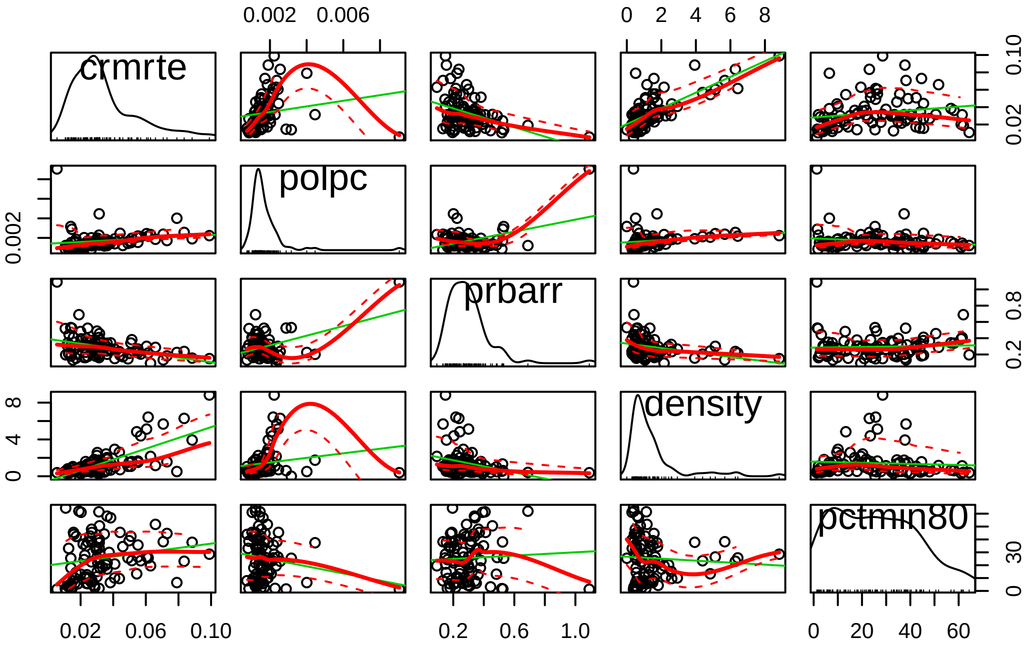
There are no outliers in the histogram of `pctmin80`.

With the exception of the one record that is an outlier 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 outlier record (#51).

3.1 Scatterplot Matrix

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

```r
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

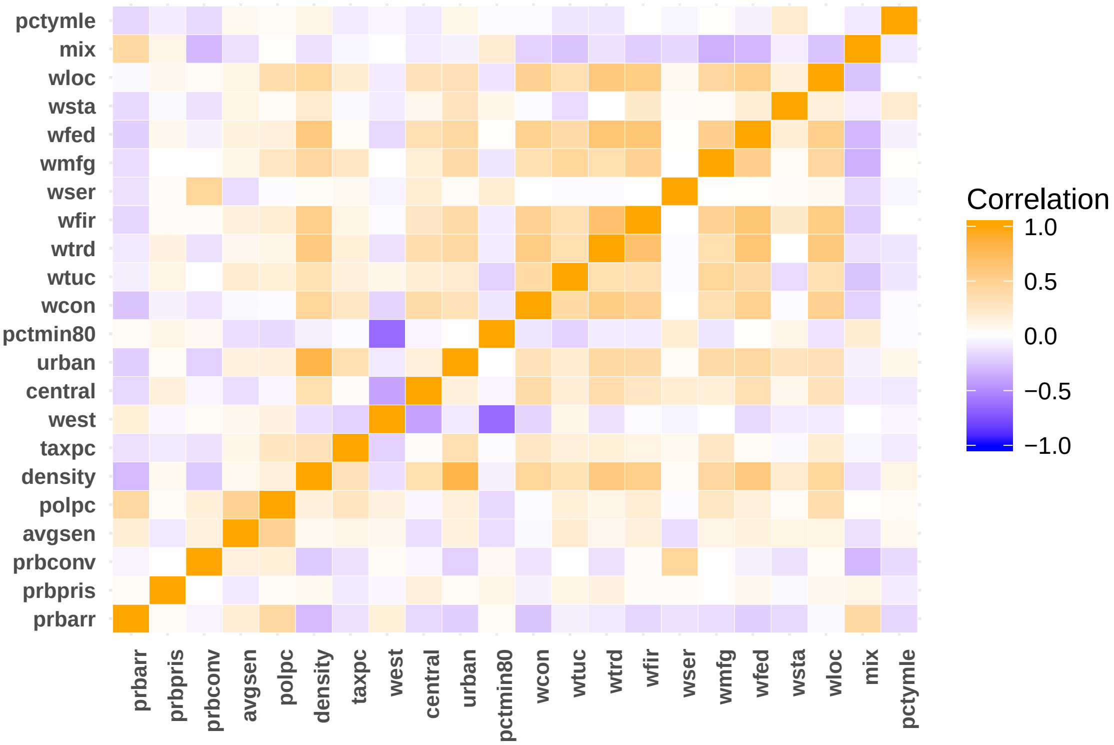
Since additional variables may be included in other models for crime rate, it is important to identify those explanatory variables with a high degree of colinearity. Perfectly colinear variables are prohibited in OLS regression, so in the unlikely event that any such variables are found, only 1 from 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.

```r
# 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())
```

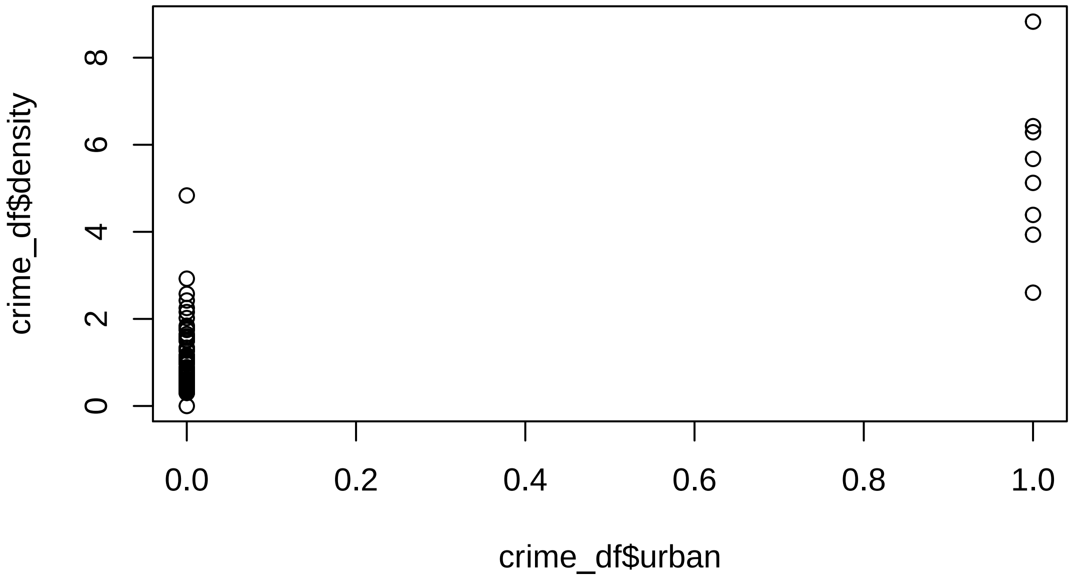


After reviewing the correlation matrix in detail, there were 5 pairs of variables that have a somewhat strong correlation to each other (i.e. has correlation > 0.6), which are plotted below. Based on the plots, the following variables were removed from the final model:

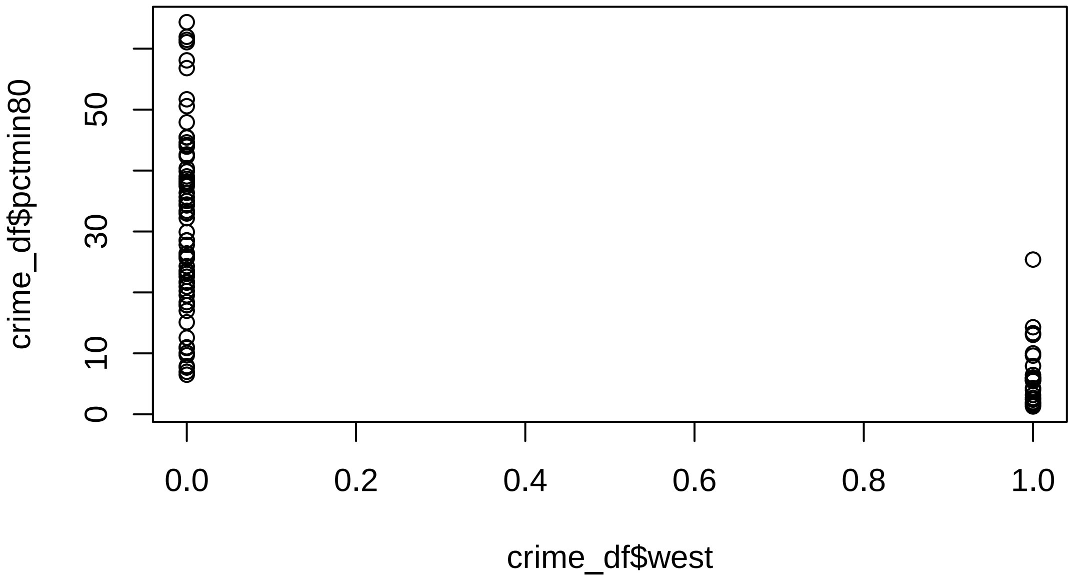
- `urban` (82% correlation with `density`. Kept `density` because it is a continuous variable providing more information than the categorical `urban` variable)
- `west` (-64% correlation with `pctmin80`. Kept `pctmin80` because it is continuous.)
- `wtrd`, `wfed`, `wfir` (each of these had correlations >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.

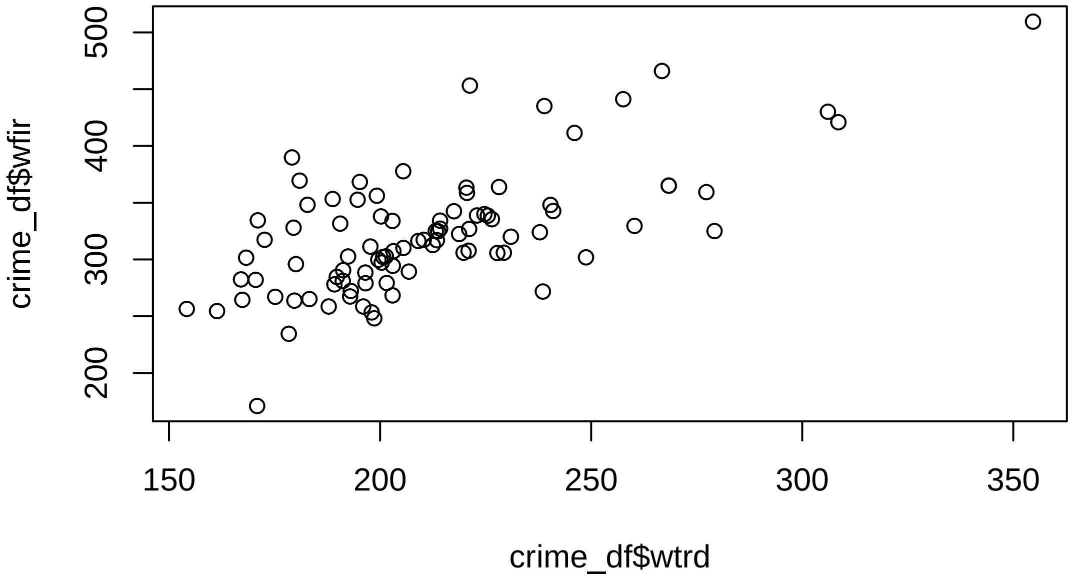
```
plot(crime_df$urban, crime_df$density)
```



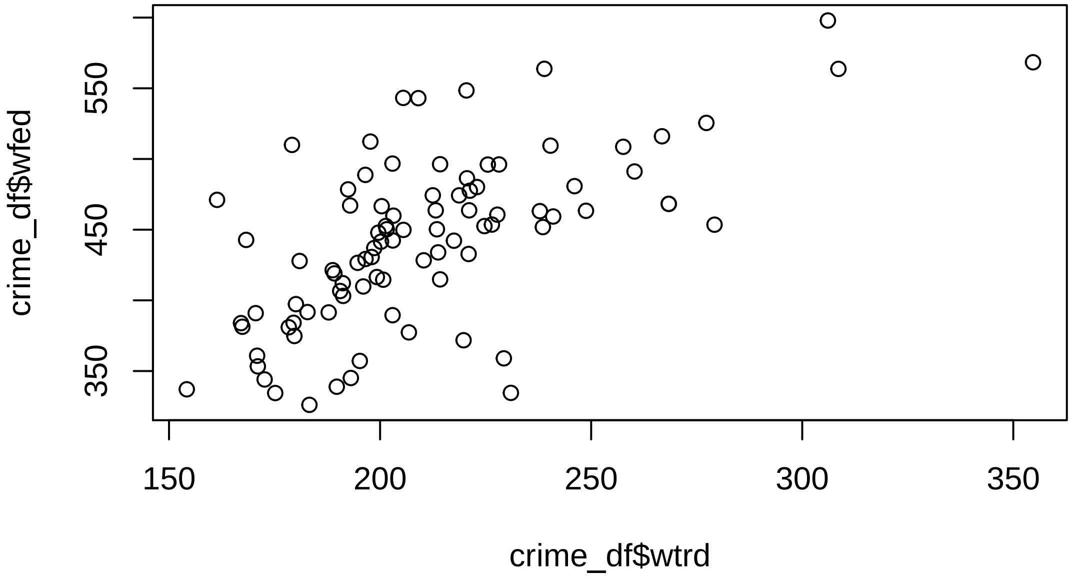
```r
plot(crime_df$west, crime_df$pctmin80)
```



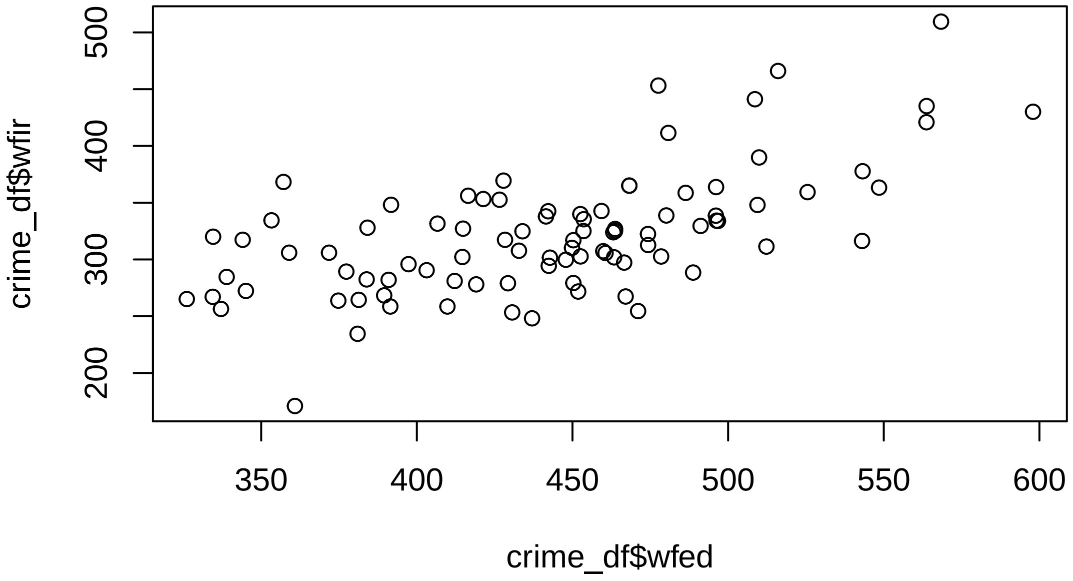
```r
plot(crime_df$wtrd, crime_df$wfir)
```



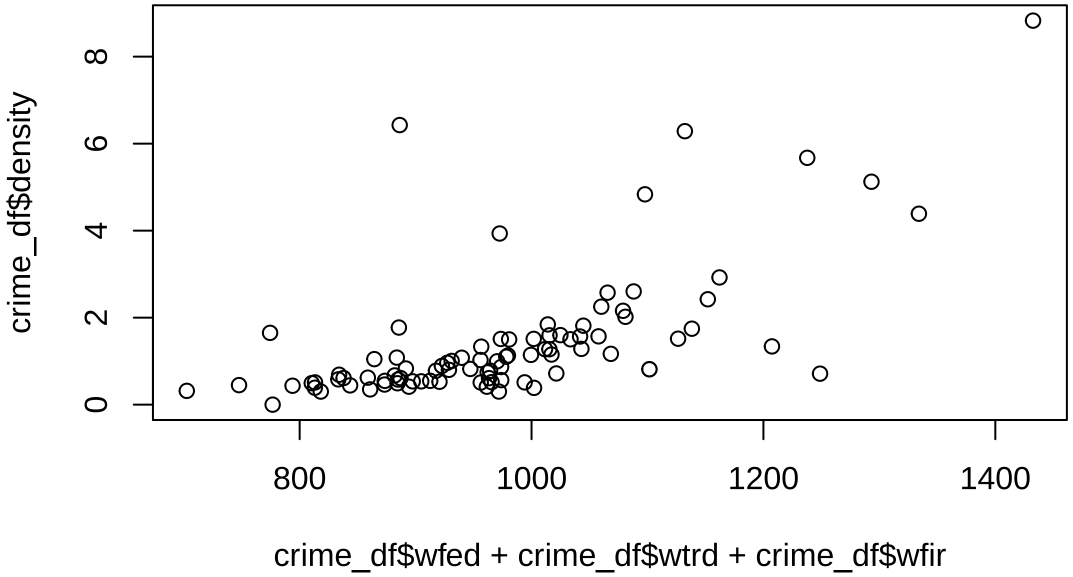
```
plot(crime_df$wtrd, crime_df$wfed)
```



```
plot(crime_df$wfed, crime_df$wfir)
```



```
plot(crime_df$wfed + crime_df$wtrd + crime_df$wfir, crime_df$density)
```



4.0 Regression Models: Base Model

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.

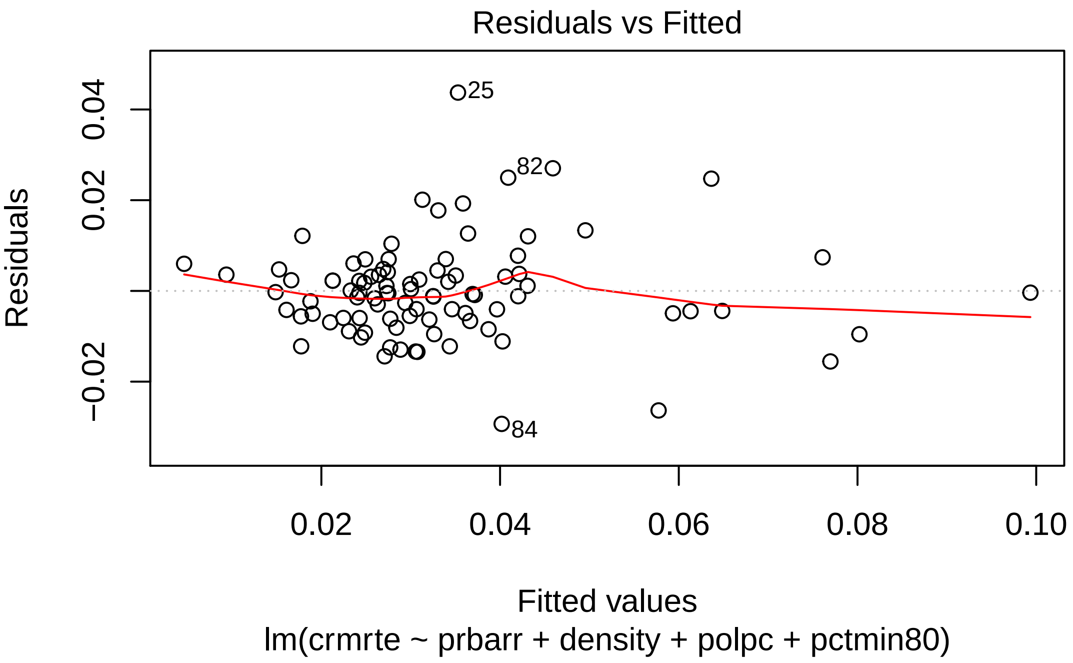
CLM 1. Linear population model: We do not have to worry about this assumption at the moment because we haven't constrained the error term.

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.

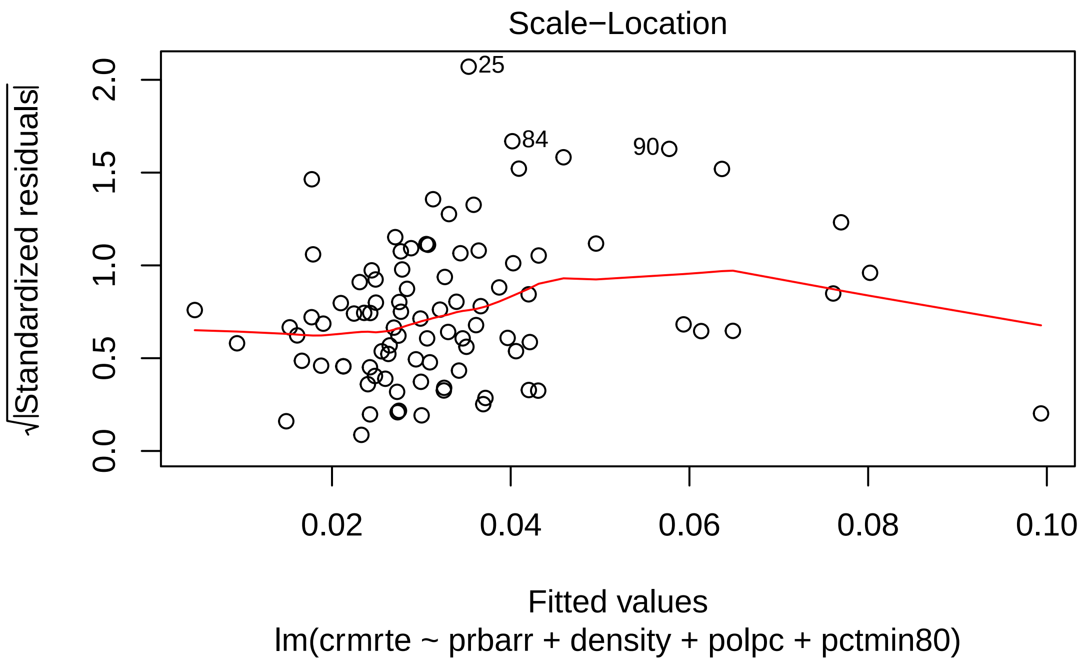
CLM 4. Zero Conditional Mean: $E(u|x) = 0$. For this model, the residuals vs. fitted values plot shown below reveals a relatively flat spline centered around zero. Therefore, there does not seem to be a clear deviation from the zero conditional mean and the assumption holds.

```
# 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)
```

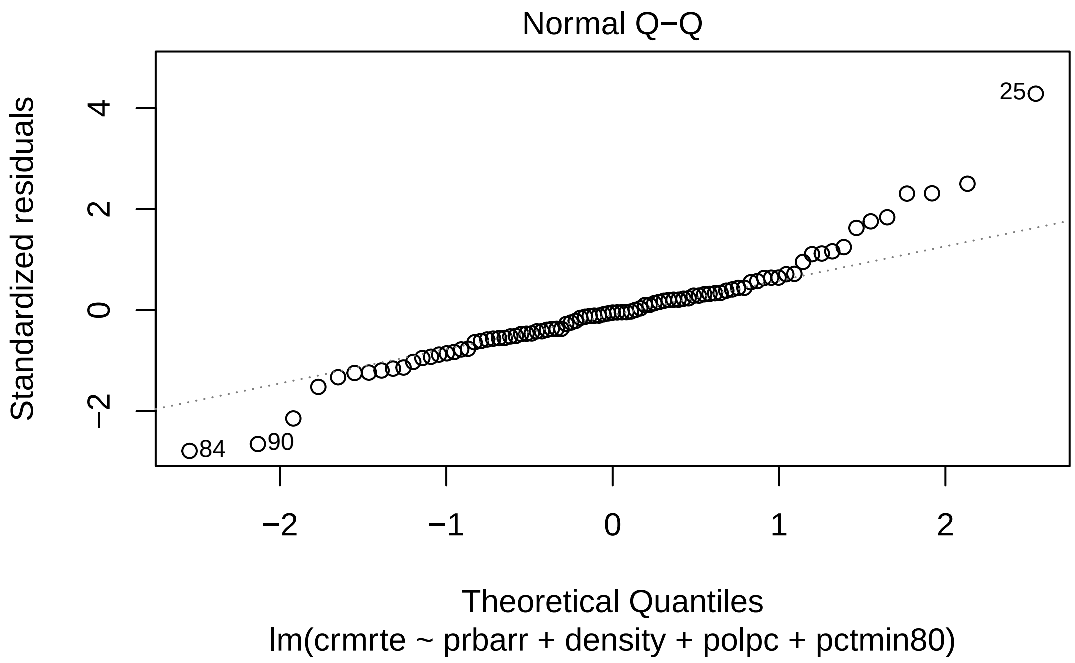


```
# Breusch-Pagan
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 24.521, df = 4, p-value = 6.278e-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)
```

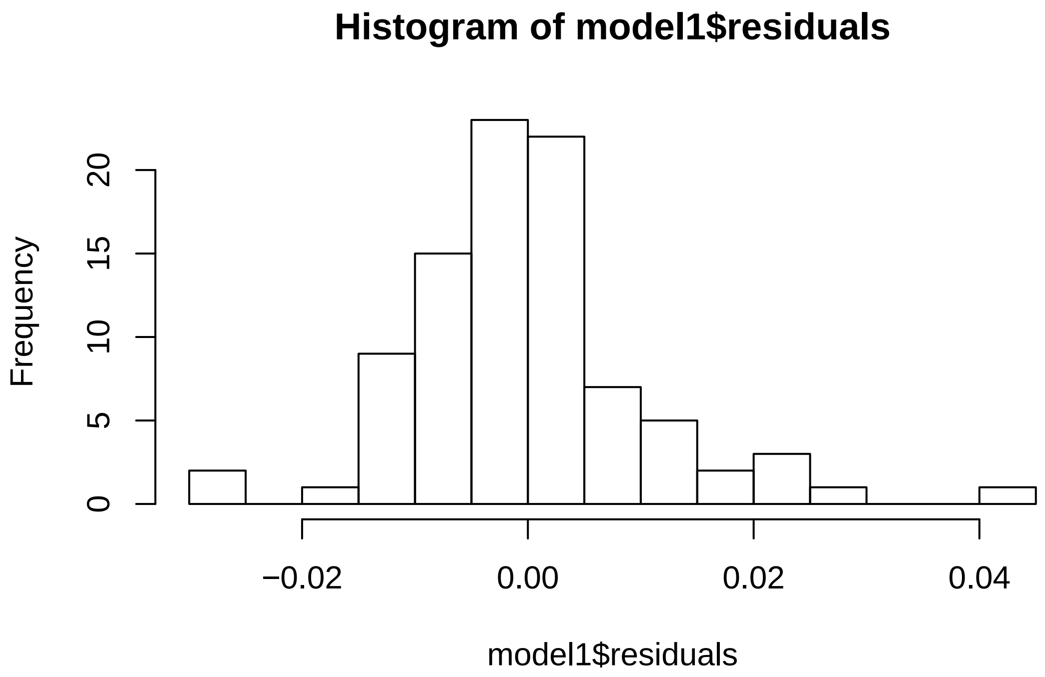


```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.93713, p-value = 0.0002688
```

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:

```
# TO DO: Use this at the end (section 4.3) TO DO: add
# F-statistics and comment on it, code below
# linearHypothesis(model1, c('prbarr = 0', ' density = 0',
# 'polpc = 0', 'pctmin80 = 0'), vcov = vcovHC)

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

# stargazer(model1, title = 'Base Model',type = 'text',
# report = 'vcstp', omit.stat = 'f', se = list(se.model1,
# NULL), star.cutoffs = c(0.05,0.01,0.001))

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

```
## [1] "adj.r.square: 0.656841444317101"
```

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

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.9722e-02  7.9057e-03  2.4946 0.0145210 *
## prbarr      -4.6441e-02  1.9565e-02 -2.3737 0.0198401 *
## density      7.5082e-03  1.1606e-03  6.4690  5.78e-09 ***
## polpc        5.0116e+00  4.2773e+00  1.1717 0.2445552
## pctmin80     3.1834e-04  8.4981e-05  3.7460 0.0003243 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

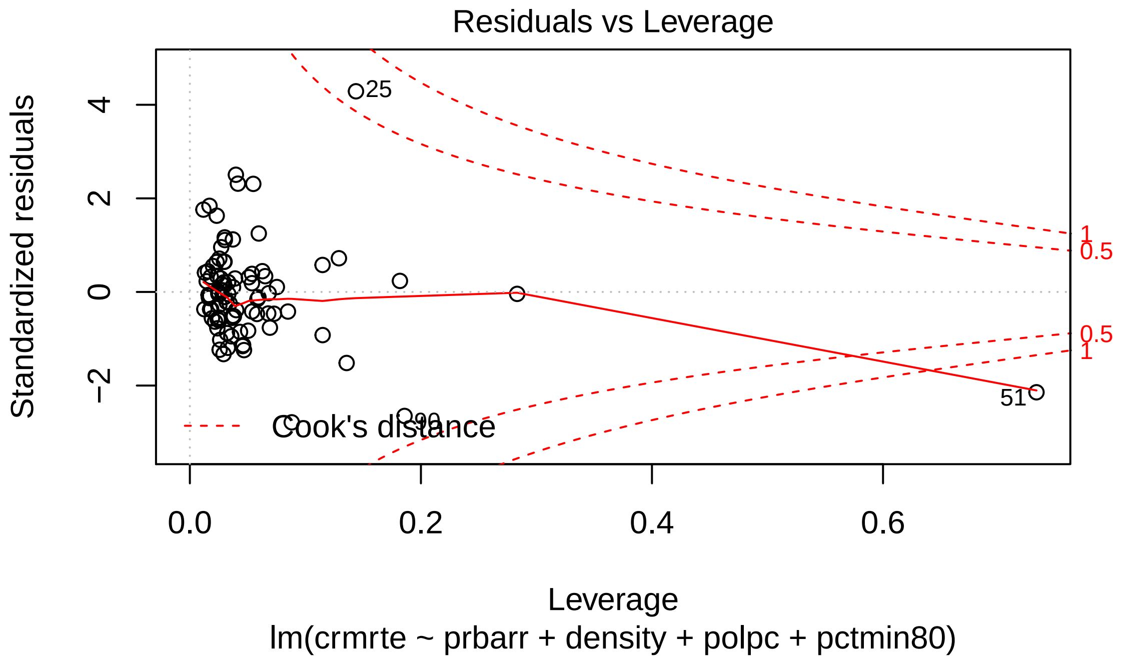
The adjusted r-squared of the model is relatively high at 0.66. This means that 66% of the variation in crime rate is explained by our input variables. Furthermore, the results of our initial model shows that the probability of arrest is statistically significant as a modulator of crime, while the density 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 TO DO** 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 a few 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.

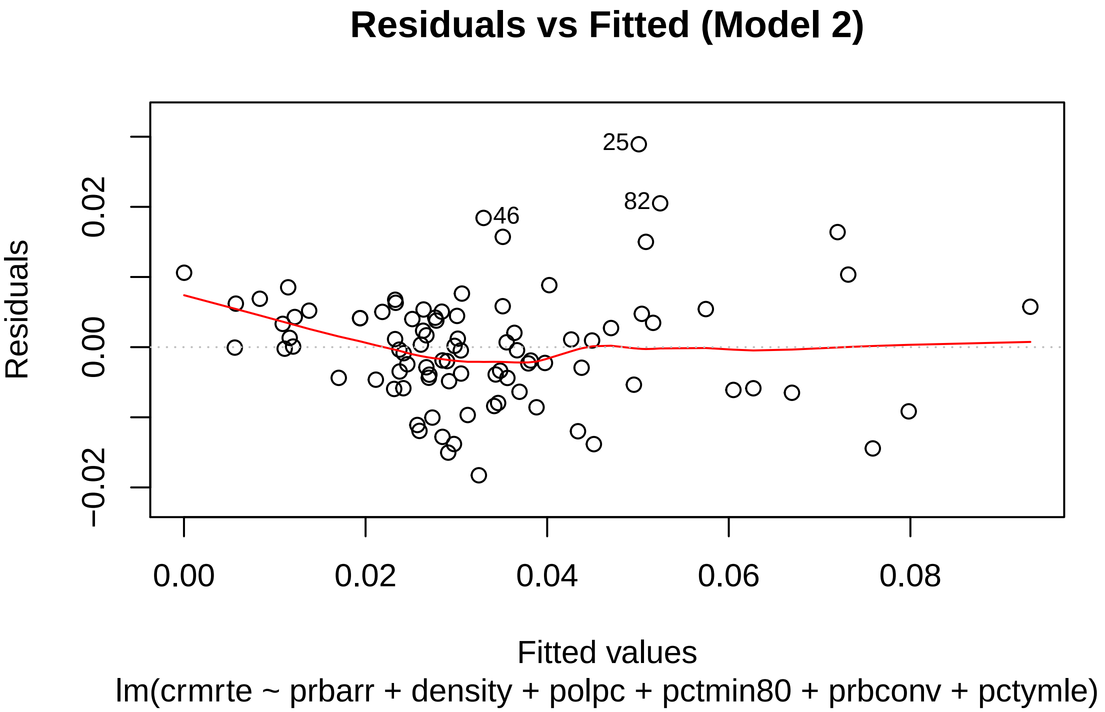
A second model was created which included the three original explanatory variables (probability of arrest, `prbarr`; population per square mile, `density`; and police per capita, `polpc`) plus two additional variables–the "probability" of conviction, `probconv` and percentage of a county's population 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 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)
# plot(model2)

# TO DO: why is this here?  Breusch-Pagan 0.0004972
```

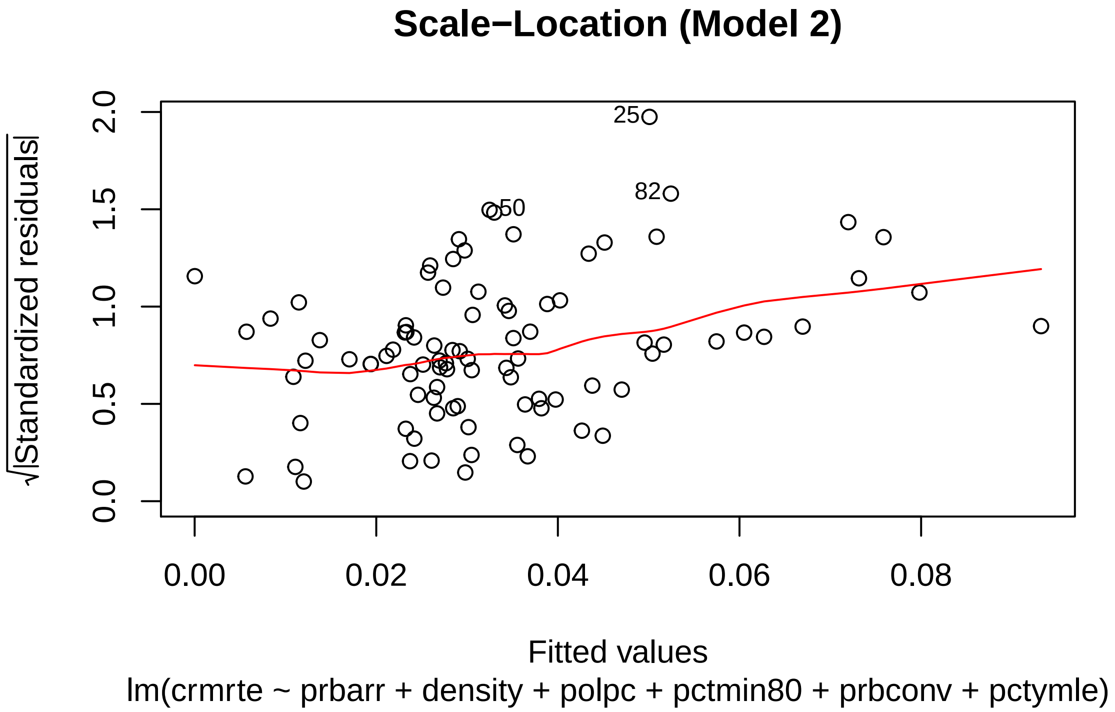
This second model produced some unexpected results with respect to the linear model assumptions. First, the residuals show more deviation from the zero conditional mean assumption than our previous model exhibited. The residuals vs. fitted values plot below shows positive residuals for fitted values 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)")
```



Heteroscedasticity of the second model appears to be greater than that of the first. The Scale-Location plot for the second model is shown below, and it reveals higher standardized residuals for higher fitted values. Using heteroscedastic-robust standard errors when evaluating the model coefficients should prevent this from being a problem.

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



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

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

```
## [1] "Model 2 adj.r.square: 0.797685729596468"
```

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

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)   2.7309e-02  7.7971e-03  3.5024 0.0007413 ***
## prbarr       -6.2411e-02  1.5385e-02 -4.0567 0.0001109 ***
## density       5.6246e-03  1.2283e-03  4.5793 1.603e-05 ***
## polpc         7.7328e+00  2.4873e+00  3.1089 0.0025638 **
## pctmin80      3.6602e-04  5.3742e-05  6.8105 1.363e-09 ***
## prbconv      -2.0579e-02  4.5539e-03 -4.5190 2.016e-05 ***
## pctymle       6.3262e-02  5.3147e-02  1.1903 0.2372745
## ---
## 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.

4.2 Regression Third model

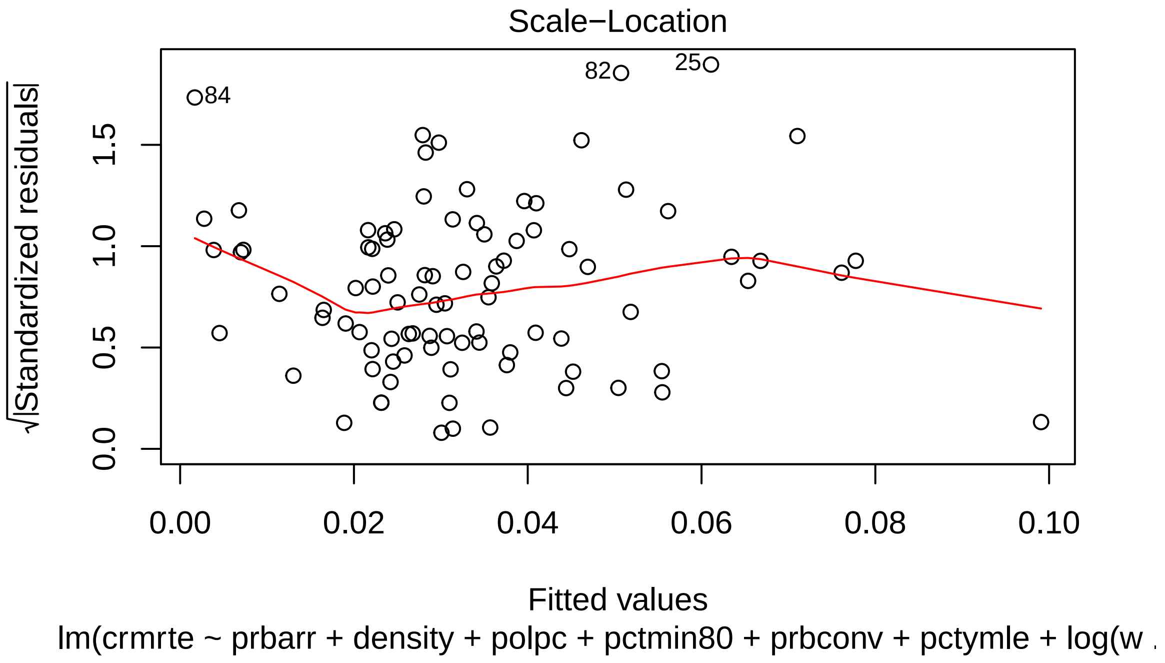
The following is the model that contains almost all available variables as explanatory variables with the exception of variables we excluded due to high level of multi-collinearity.

```
model3 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + prbconv +
    pctymle + log(wcon) + log(wtuc) + log(wtrd) + log(wfir) +
    log(wser) + log(wmfg) + log(wfed) + log(wsta) + log(wloc) +
    log(wser) + taxpc + west + 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, and the Breusch-Pagan test result show that there is heteroscedacity present in the model. Based on this finding, 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 = 34.76, df = 21, p-value = 0.03
```

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

```
## [1] "adj.r.square: 0.656841444317101"
```

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

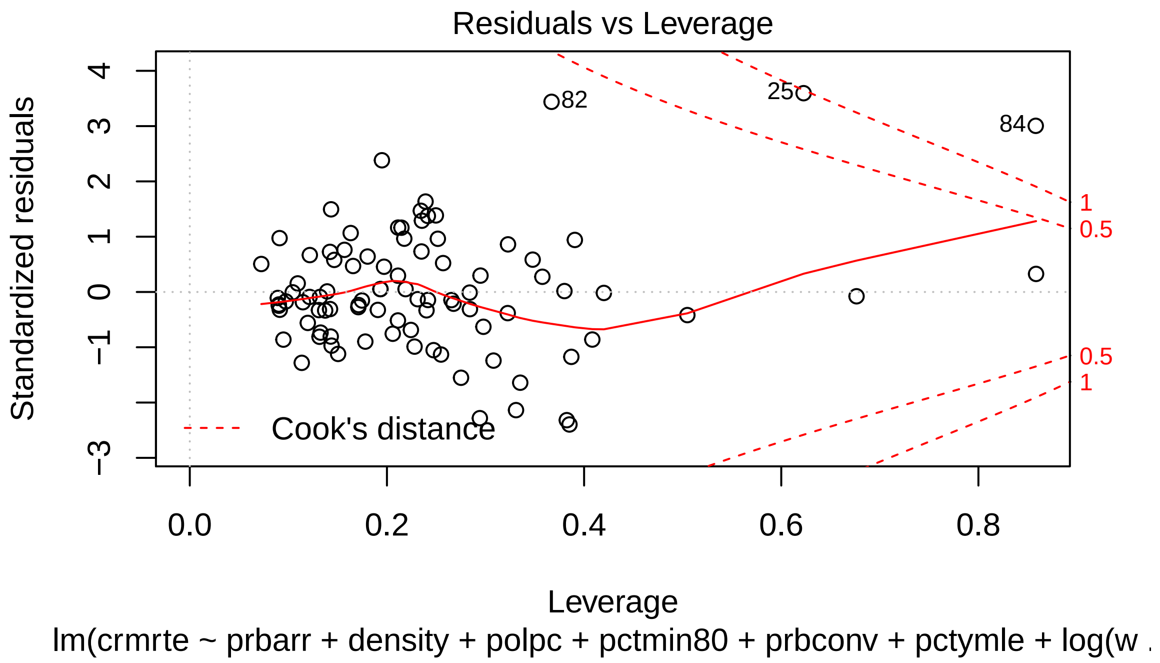
```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -0.04786422  0.15627453 -0.3063 0.760312
## prbarr      -0.05109195  0.01531780 -3.3355 0.001374 **
## density      0.00525874  0.00157734  3.3339 0.001380 **
## polpc        6.87513478  2.90430566  2.3672 0.020732 *
## pctmin80     0.00034245  0.00014495  2.3626 0.020974 *
## prbconv     -0.01743026  0.00650062 -2.6813 0.009166 **
## pctymle      0.11047115  0.04228860  2.6123 0.011028 *
## log(wcon)    0.00642800  0.00969073  0.6633 0.509339
## log(wtuc)    0.00420363  0.00810526  0.5186 0.605678
## log(wtrd)    0.00754483  0.01732896  0.4354 0.664640
## log(wfir)   -0.00905225  0.01231900 -0.7348 0.464939
## log(wser)   -0.00514504  0.02029091 -0.2536 0.800587
## log(wmfg)   -0.00164440  0.00714679 -0.2301 0.818703
## log(wfed)    0.01164777  0.01471911  0.7913 0.431460
## log(wsta)   -0.00644867  0.01226333 -0.5259 0.600678
## log(wloc)    0.00440191  0.02611763  0.1685 0.866650
## taxpc        0.00018299  0.00027795  0.6584 0.512495
## west        -0.00173873  0.00443158 -0.3924 0.696008
## central     -0.00368778  0.00401902 -0.9176 0.362033
## mix         -0.01830292  0.02286615 -0.8004 0.426205
## prbpris      0.00252300  0.01323413  0.1906 0.849364
## avgsen      -0.00051442  0.00047749 -1.0773 0.285077
## ---
## 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 inclueded as part of model 3. In addition it looks like now `pctmyle`‘s slope coefficient is considered signficant compared to model 2; the slope coefficient can be interpreted that a point increase in 'pctymle' corresponds to .11 increase in `crmrte`.

Other Analysis The residuals vs. leverage plot shows that there are two data points (25 and 84), that have Cook's distance greater or equal to 1, indicating that they highly influence the regression. 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 7.7% of our data points have standardized residuals over 2. Therefore, our model has too much error and may represent our data poorly.

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



4.3 Regression Table

TO DO: Be sure to convert SE's to robust before displaying.

Joint Signficance

In addition to the table above, we also tested the joint signficance of the model, to see if the variables that were added for model 2 and model 3 improved the fit 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     86
## 2     84  2 17.134 5.743e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see above, the addition of 'probconv' and 'pctymle' are jointly significant, thus improving the fit of the model. Next we will further test to see if the remainder of the variables that were not included as part of model 2 can further improve the fit of the model.

```
# joint significance between model1 and model2
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(wtrd) + log(wfir) + log(wser) +
##     log(wmfg) + log(wfed) + log(wsta) + log(wloc) + log(wser) +
##     taxpc + west + central + mix + prbpris + avgsen
##   Res.Df Df      F Pr(>F)
## 1     84
## 2     69 15 0.9381 0.5277
```

```
summary(model2)$adj.r.squared
```

```
## [1] 0.7976857
```

```
summary(model3)$adj.r.squared
```

```
## [1] 0.813829
```

It does not look like the additional variable that were included as part of model 3 are jointly significant , did not improve the model in a statistically significant way.

```
# TO DO: do we still need the code below or delete it?
crime_df2 <- crime_df[-c(84, 25), ]

model1 <- lm(crmrte ~ . - county - year - crmrte - urban - west -
    wtrd - wfed - wfir, data = crime_df2)

summary(model1)$r.squared
```

```
## [1] 0.8688977
```

```
summary(model1)$coefficients
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  3.097640e-02 1.561081e-02  1.9842914 5.108937e-02
## prbarr      -5.078247e-02 8.704418e-03 -5.8341022 1.478721e-07
## prbconv     -1.962352e-02 3.293124e-03 -5.9589361 8.903946e-08
## prbpris      4.754774e-03 1.048442e-02  0.4535087 6.515656e-01
## avgsen      -3.961756e-04 3.497705e-04 -1.1326727 2.611624e-01
## polpc        6.460940e+00 1.346144e+00  4.7995886 8.534766e-06
## density      6.845704e-03 7.973078e-04  8.5860246 1.369694e-12
## taxpc       -7.103721e-05 1.021711e-04 -0.6952767 4.891513e-01
## central     -3.517708e-03 1.926533e-03 -1.8259265 7.206619e-02
## pctmin80     3.898315e-04 5.176540e-05  7.5307361 1.236934e-10
## wcon         4.081808e-05 2.373695e-05  1.7196007 8.986184e-02
## wtuc         4.373842e-06 1.324754e-05  0.3301626 7.422493e-01
## wser        -6.293562e-05 2.794740e-05 -2.2519307 2.742112e-02
## wmfg         4.568252e-06 1.208881e-05  0.3778909 7.066390e-01
## wsta        -4.273992e-05 2.105298e-05 -2.0301130 4.609284e-02
## wloc         4.531803e-05 4.143979e-05  1.0935875 2.778324e-01
## mix         -2.294321e-02 1.269191e-02 -1.8077035 7.488831e-02
## pctymle      9.580106e-02 3.779334e-02  2.5348663 1.345432e-02
```

The following is the model that contains a transformed explanatory variable.

```
# TO DO: do we still need the code below or delete it?
model_transform <- lm(crmrte ~ prbarr + log(prbconv) + density,
    data = crime_df2)

summary(model_transform)$r.squared
```

```
## [1] 0.6570935
```

```r
summary(model_transform)$coefficients
```

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)   0.025420503 0.0035106022  7.241066 1.857260e-10
## prbarr       -0.028710438 0.0089889944 -3.193954 1.969045e-03
## log(prbconv) -0.006276946 0.0022761837 -2.757662 7.125235e-03
## density       0.007903815 0.0008331222  9.486981 5.580124e-15
```

The following is the model that contains only variables that were identified to be most relevant to crmrte based on their marginal R-squared and standardized slope coefficient values.

```r
# TO DO: do we still need the code below or delete it?
model_key <- lm(crmrte ~ prbarr + prbconv + polpc + density +
    pctmin80, data = crime_df2)

summary(model_key)$r.squared
```

```
## [1] 0.8204393
```

```r
summary(model_key)$coefficients
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  0.0300488820 3.494735e-03  8.598328 4.156915e-13
## prbarr      -0.0555832603 8.317408e-03 -6.682763 2.515871e-09
## prbconv     -0.0179293179 3.139371e-03 -5.711118 1.698543e-07
## polpc        6.1601721055 1.204450e+00  5.114512 1.989594e-06
## density      0.0063705861 6.966292e-04  9.144873 3.349488e-14
## pctmin80     0.0003808799 5.212093e-05  7.307620 1.527153e-10
```

Stargazer Regression Table for Model Specifications

```r
library(stargazer)
stargazer(model_transform, model_key, model1, title = "Linear Models Parameters Predicting Crime Rate",
    type = "text", report = "vc", keep.stat = c("rsq", "n"),
    omit.table.layout = "n")
```

Linear Models Parameters Predicting Crime Rate

```
 Dependent variable:
----------------------
        crmrte
 (1)    (2)     (3)
```

prbarr -0.029 -0.056 -0.051

log(prbconv) -0.006

prbconv -0.018 -0.020

prbpris 0.005

avgsen -0.0004

polpc 6.160 6.461

density 0.008 0.006 0.007

taxpc -0.0001

central -0.004

pctmin80 0.0004 0.0004

```
wcon 0.00004
wtuc 0.00000
wser -0.0001
wmfg 0.00000
wsta -0.00004
wloc 0.00005
mix -0.023
pctymle 0.096
Constant 0.025 0.030 0.031
```

```
Observations 89 89 89
R2 0.657 0.820 0.869
==================================
```

Recommendation

For interpretability purposes, the model was re-done using non-standardized variables: -prbarr -prbconv -polpc -density -pctmin80

Recommendation for political campaign: - police per capita has a positive slope coefficient with crmrte, and this may be due to more police are present in areas with high crmrte. This suggests that purely hiring more police officers may not be an impactful solution. - However probability of arrest and conviction both have a negative slope coefficients. The model suggests that perhaps a zero tolerance policy towards crime is needed to increase arrests and convictions and thus deter crimes from happening. - In terms areas with large minority population and high density, since these variable cannot be changed that much, perhaps a community outreach (e.g. job training program, afterschool programs, tutor/mentor program) to educate areas with a lot of minority can be done, so that crimes can be reduced in those areas.

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 has, 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 `prbarr` on crime rate may also be overestimated by omitting the effect of education in the model.

TO DO: Density?

Potential Omitted Variable #5: Rate of Alchohol and Drug Abuse

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * substance\_abuse$$

$$substance\_abuse = \alpha_0 + \alpha_1 * pctmin80$$

The last omitted variable we considered is the Rate of substance abuse as an indicator of crime. We assume that counties with a higher rate of substance abuse would also have a higher crime rate because the usage of illegal drugs is a crime and the abuse of alchohol can lead to 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 (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 `prbarr` on crime rate may also be overestimated by omitting the effect of substance abuse in the model.

https://drugabuse.com/country-vs-city-addictions-differ-says-samhsa/ https://www.samhsa.gov/specific-populations/racial-ethnic-minority TO DO: Cities have more drug abusers, fewer alchohol abusers TO DO: In Cities, larger percentage of minorities in cities abuse drugs, larger percentage of caucasians in rural areas abuse druges.

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.

TO BE SORTED LATER

TO BE SORTED LATER

TO BE SORTED LATER

TO BE SORTED LATER

Standardized Regression Model

TO DO: Eliminate. Save the comments on the diagnostic plots for use in the non-standardized model analysis.

A multi variable regression model was created using the data set that has been standardized above.

Then the model was evaluated for potential high leverage/influence data points as well as potential biases.

In review the following findings were noted: - row 84 and 25 have a high Cook's distance and high standardized residuals, which means the data point can be problematic for the regression model. - row 25 and 84 were also noted earlier to be an extreme outier for the wser variable. Thus based on this finding the point will be removed and the regression will be redone. - Judging from the residuals vs. fitted plot the model may have some bias when the predicted value crmrte is between 0 to 0.04. Particularly the model tend to underpredict lower crmrates, and overpredict medium crmrte. - From the Normal Q-Q line, it looks like that majority of predictions follow the line, indicating a normal and independent distribution.

```
# TODO clean out the warning std_model <- lm(crmrte ~ . -
# county-year-crmrte-urban-west-wtrd-wfed-wfir, data =
# std_crime_df)

# plot(std_model,1) plot(std_model,5) plot(std_model,2)

# summary(std_model)$r.squared
```

```
std_crime_df2 <- std_crime_df[-c(84,25),]

std_model2 <- lm(crmrte ~ . - county-year-crmrte-urban-west-wtrd-wfed-wfir, data =  std_crime_df2)

plot(std_model2,1)
plot(std_model2,5)
plot(std_model2,2)
```

TO DO: eliminate.

In order to find which variables are most impactful to crmrte, the marginal R-squared against the standardized coefficients were reviewed. Based on the plots, the following variables were found to have the highest marginal R-squared and absolute slope coefficient: -prbarr -prbconv -polpc -density -pctmin80

```
coeff_df = data.frame(summary(std_model)$coefficients)
#summary(std_model)$r.squared

#base R-Squared
base_model <- lm(crmrte~.-county-year-crmrte, data=std_crime_df)
base_r2 <- summary(base_model)$r.squared

#create list of variables for the for-loop
```

```
var_names <- colnames(std_crime_df)
remove <- c('county',
            'year',
            'crmrte',
            'urban',
            'west',
            'wtrd',
            'wfed',
            'wfir')
var_names <- var_names[! var_names %in% remove]

#initiate an empty vector to store the marginal R-Squared
var_r2_delta = c()

#loop through the variable names and store the marginal R-Squared
for (i in var_names) {
    fmla <- as.formula(paste("crmrte ~ - crmrte +", paste(var_names[! var_names %in% i], collapse= "+")
    delta_model <- lm(fmla, data=crime_df)
    r2_delta <- base_r2-summary(delta_model)$r.squared
    var_r2_delta <- c(var_r2_delta, r2_delta)
}

#put the variable and marginal R-squared in a dataframe
mar_r2_df <- data.frame(v1=var_names, v2=var_r2_delta)
colnames(mar_r2_df) <- c('variable', 'marginalr2')

#sort dataframe by marginal R-squared in a descending order
#mar_r2_df <- mar_r2_df[rev(order(mar_r2_df$marginalr2)),]

plot(abs(coeff_df[-c(1),]$Estimate),mar_r2_df$marginalr2)

subset(mar_r2_df, marginalr2 > .04)
```