W203: Lab 3 Strategies for Crime Reduction in N.C.

Mark Barnett, Siobhan Harrington, and I-Wae Niu

04/15/2018

1. Introduction - Income Inequality and Crime Rates

We are a team of data scientists working for a gubernatorial candidate in North Carolina. The candidate and our boss, Earnie Anders has a long-standing platform of challenging income inequality. Policies to combat income inequality are woven throughout his platform. While it is generally accepted that income inequality is a negative force, our candidate is looking for hard evidence in data across the board (impacts on economic growth, population health, corruption, and crime rates). Our team was tasked with understanding the relationship between income inequality and crime rates in North Carolina. The opposing candidate supports a strong punitive process for criminals and has long advocated for harsher sentences and police enforcement, so we will also be researching the impacts of this alternative policy on crime.

To that end, we have secured data gathered during a preceding investigation by Cornwell and Trumball (1994) for most counties across the state of North Carolina. The data set used for this analysis is modified from the original study, most impactfully by the removal of time series data limiting our ability to observe changes over time to crime rates in relation to the other captured variables. We seek to understand the answers to the following questions:

1. Does income inequality effect crime rates? We will use available weekly wage variables to calcuate a wage gap variable as a proxy.
$$crmrte = \beta_o + \beta_1 wagegap + u$$

2. If so, how does that compare with the traditional method of reducing crime rates with increased certainty and severity of punishment?

$$crmrte = \beta_o + \beta_1\ wagegap + \beta_2\ probarrest + \beta_3\ probconviction + \beta_4\ probprison + \beta_5\ avgsentence + u$$

Effectively, in this race there are two approaches to crime reduction policies: the "carrot" and the "stick". The carrot approach is to provide incentives to avoid committing crime through policy by increasing individual quality of life measurements (e.g, employment rates, average education, increasing disposable income) versus the stick approach with a focus on the controlling activities of the state through increased police presence and rigorous application of the law measured through crime identification and prosecution (Probability of Arrest and Probability of Conviction) criminal process effectiveness (e.g., Probability of Prison and Average Prison Sentence).

2. Initial EDA

2.1 Load the data

```
library(car)
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```r
library(MASS)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```r
library(sandwich)
library(ggcorrplot)
# setwd("~/Desktop")
crime = read.csv("crime_v2.csv")
str(crime)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

```r
stargazer(crime, type = "latex", nobs = FALSE, mean.sd = TRUE, median = TRUE,
          iqr = TRUE, float = FALSE)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 15, 2018 - 12:05:26

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
county	101.615	58.794	1	52	105	152	197
year	87.000	0.000	87	87	87	87	87
crmrte	0.033	0.019	0.006	0.021	0.030	0.040	0.099
prbarr	0.295	0.137	0.093	0.206	0.271	0.344	1.091
prbpris	0.411	0.080	0.150	0.365	0.423	0.457	0.600
avgsen	9.647	2.847	5.380	7.340	9.100	11.420	20.700
polpc	0.002	0.001	0.001	0.001	0.001	0.002	0.009
density	1.429	1.514	0.00002	0.547	0.962	1.568	8.828
taxpc	38.055	13.078	25.693	30.662	34.870	40.948	119.761
west	0.253	0.437	0	0	0	0.5	1
central	0.374	0.486	0	0	0	1	1
urban	0.088	0.285	0	0	0	0	1
pctmin80	25.495	17.017	1.284	9.845	24.312	38.142	64.348
wcon	285.358	47.487	193.643	250.782	281.426	314.795	436.767
wtuc	411.668	77.266	187.617	374.632	406.504	443.436	613.226
wtrd	211.553	34.216	154.209	190.864	203.016	225.126	354.676
wfir	322.098	53.890	170.940	286.527	317.308	345.354	509.466
wser	275.564	206.251	133.043	229.662	253.228	280.541	2,177.068
wmfg	335.589	87.841	157.410	288.875	320.200	359.580	646.850
wfed	442.901	59.678	326.100	400.240	449.840	478.030	597.950
wsta	357.522	43.103	258.330	329.325	357.690	382.590	499.590
wloc	312.681	28.235	239.170	297.265	308.050	329.250	388.090
mix	0.129	0.081	0.020	0.081	0.102	0.152	0.465
pctymle	0.084	0.023	0.062	0.074	0.078	0.083	0.249

2.2 Remove missing values and identify anomalous values

In this section we outline what changes we made to the baseline data set.

2.2.1 Data Set Housecleaning

We notice that *prbconv* is coded as a factor vs. a numeric variable like the other probability variables and transform *prbconv* back to a numeric variable. Given *crmrte* is our dependent variable, we will remove all observations that have missing values on this variable. We also convert regional variables *west*, *central* and *urban* into factor variables and convert *prbarr*, *prbconv*, *prbpris*, and *pctmle* into percentages instead of proportions.

```
# Transform prbconv from factor to numeric
crime$prbconv = as.numeric(levels(crime$prbconv)[crime$prbconv])
```

```
## Warning: NAs introduced by coercion
```

```
# create clean dataset without missing values on crime rate
crime2 = crime[!is.na(crime$crmrte), ]

# convert regional variables into factors
crime2$west = as.factor(crime2$west)
crime2$central = as.factor(crime2$central)
crime2$urban = as.factor(crime2$urban)
```

```
# convert to percent % values
crime2$prbarr_pct = crime2$prbarr * 100
crime2$prbconv_pct = crime2$prbconv * 100
crime2$prbpris_pct = crime2$prbpris * 100
crime2$pctymle_pct = crime2$pctymle * 100
```

2.2.2 Deterrent Variables Anomalies

At a high level, looking at the histogram for our core deterrent varaiables *prbarr*, *prbconv*, *avgsen* and *polpc* we see some consistency across the variables.

```
par(mfrow=c(1,4))
hist(crime2$prbarr, breaks = 30, main = "Histogram of Probability of Arrest",
     xlab = "Probability of Arrest")
hist(crime2$prbconv, breaks = 30, main = "Histogram of Probability of Conviction",
     xlab = "Probability of Conviction")
hist(crime2$avgsen, breaks = 30, main = "Histogram of Average Sentence",
     xlab = "Probability of Conviction")
hist(crime2$polpc, breaks = 30, main = "Histogram of Police per Capita",
     xlab = "Probability of Conviction")
```

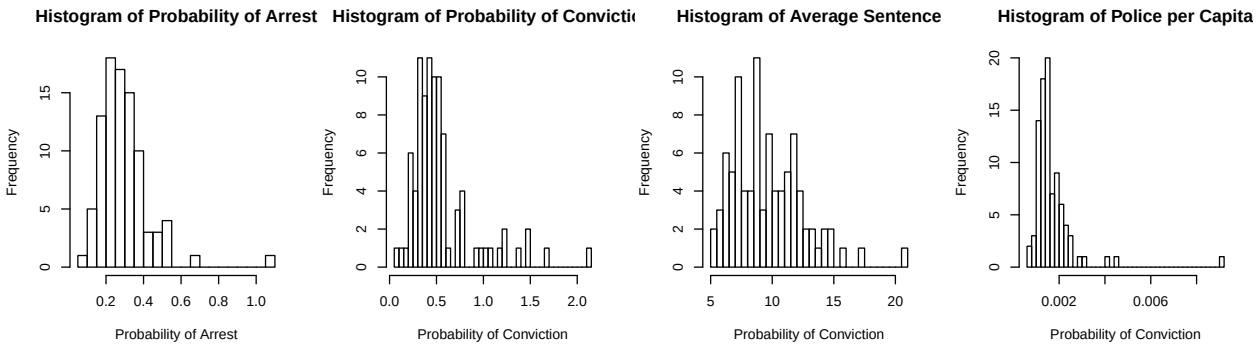

Figure 1: Histograms of prbarr & prbconv variables

There appears to be a normal distribution but with a right tail skew due to outliers in the data set (figure 1). We will investigate to understand if these outlier values are in any way associated with on another.

Upon further examination we find there is 1 record where *prbarr* is greater than 1 which is possibly an input error. There are also 10 records where *prbconv* is greater than 1, which could be attributed to multiple convictions i.e. where the offender has committed more than one crime. The only concern we have so far is with the 1 record (**county 115**) which appears to be an outlier in several variables, namely *prbarr* is 1.090910, *avgsen* is the maximum value of 20.70 days (median is 9.1) and *polpc* is also the maximum range value 0.00905433 (median is 0.00148532).

```
crime2[crime2$county == 115, c(1, 3:8)]
```

```
##    county    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
```

As a further sanity check on our reasoning, we performed a Cook's distance check against the the 4 outlier figures from county 115 (figure 2):

```
model2231 = (lm(crmrte ~ prbarr, data = crime2))
model2232 = (lm(crmrte ~ prbconv, data = crime2))
model2233 = (lm(crmrte ~ avgsen, data = crime2))
model2234 =(lm(crmrte ~ polpc, data = crime2))

par(mfrow=c(2,4))
plot(jitter(crime2$crmrte), jitter(crime2$prbarr), xlab = "Probability of Arrest", ylab = "Crime Rate")
plot(jitter(crime2$crmrte), jitter(crime2$prbconv), xlab = "Probability of Conviction", ylab = "Crime Ra
plot(jitter(crime2$crmrte), jitter(crime2$avgsen), xlab = "Average Sentence Days", ylab = "Crime Rate")
plot(jitter(crime2$crmrte), jitter(crime2$polpc), xlab = "Police per Capita", ylab = "Crime Rate")
plot(model2231, which = 5)
plot(model2232, which = 5)
plot(model2233, which = 5)
plot(model2234, which = 5)
```

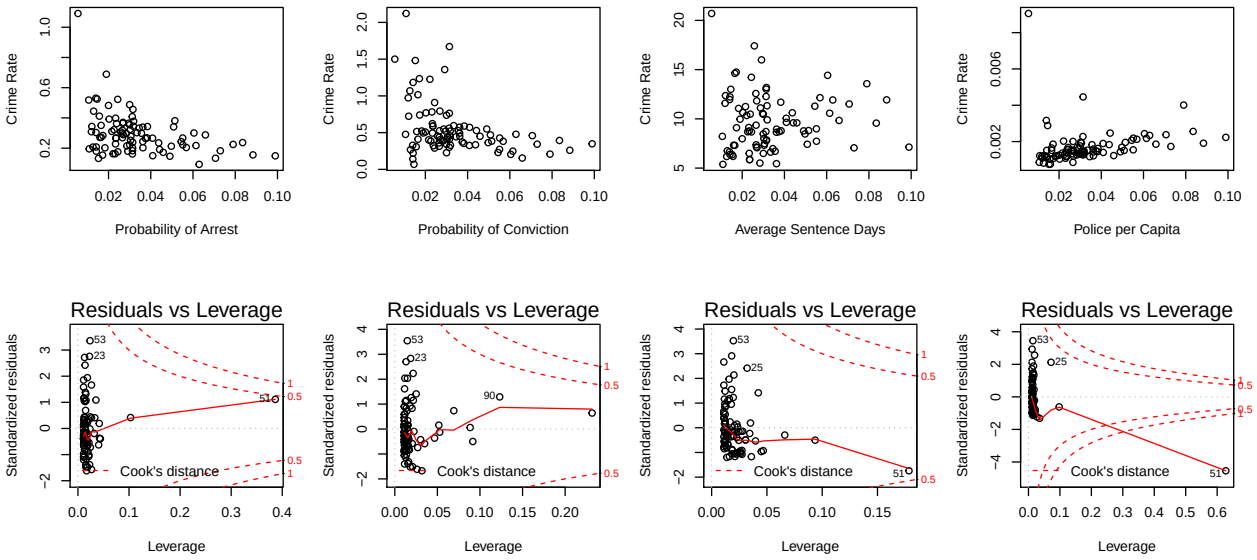

Figure 2: Cook's Distance of prbarr, prbconv, avgsen & polpc variables

From the county 115 data outliers, only the police per capita lies outside the cook's distance '1' threshold. With that in mind, and the consistency of this county showing the maximum value in the data set exerting strong leverage on our expected models it puts in question the entire record. We will remove county 115 assuming that the data is somehow flawed through the combination of extreme data points relative to the other counties.

```
crime2 <- crime2[!crime2$county == 115,]
```

2.2.3 Duplicate Records

We will now check for duplicate records in the data set.

```
anyDuplicated(crime2)
```

```
## [1] 88
```

Apparently record 88 is a duplicate of record 89.

```
crime2[87:88,]
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 88    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
## 89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 88 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 88 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle prbarr_pct prbconv_pct prbpris_pct pctymle_pct
## 88 0.07819394    26.6055     58.8859     42.3423    7.819394
## 89 0.07819394    26.6055     58.8859     42.3423    7.819394
```

As we can see, all the entries, from county to the variable data are identical and we'll therefore remove one of the duplicate county 193 entries.

```
# remove duplicate record from dataset
crime2 <- crime2[-(88), ]
```

2.2.4 Region variables (west, central, urban)

The region variables of West and Central are indicators that represent where the county is in North Carolina. Each record should have only one indicator switched on (marked with a 1). We found 1 record with both West & Central coded as "1". We believe this is likely a coding error and after examining that particular county's location (county 71 is Gaston county located in the Piedmont region of central North Carolina) on a map have decided to manually recode this county to Central.

```
crime2[(crime2$west == "1" & crime2$central == "1"), c(1,9,11:13)]
```

```
##    county  density west central urban
## 33     71 4.834734    1       1     0
```

A quick scan through the *west*, *central* and *urban* variables show that most counties are coded either as west or central or none, with an overlapping classification of whether it is urban or not. By looking up a county map of North Carolina, we can see that the state can be classified into West, Central and East regions. We will therefore, recode these 3 separate variables into 2 region variables i.e., *region* classifying counties as East/West/Central and *region_urbrul* with an overlay of urban/rural.

```
# function to tell us whether a county is west, central or east
f1 = function(w,c){
  if(w=="1" & c=="0") r="West"
  else if(w=="0" & c=="1") r="Central"
  else if(w=="1" & c=="1") r="Central"
  else r="East"
}

# function to tell us whether a county is west-urban, west-rural, central-urban, central-rural, east-ur
f2 = function(w,c,u){
  if(w=="1" & c=="0" & u=="1") r="W_Urban"
  else if(w=="1" & c=="0" & u=="0") r="W_Rural"
  else if(w=="0" & c=="1" & u=="1") r="C_Urban"
  else if(w=="0" & c=="1" & u=="0") r="C_Rural"
  else if(w=="0" & c=="0" & u=="1") r="E_Urban"
  else if(w=="1" & c=="1" & u=="0") r="C_Urban"
```

```
  else r="E_Rural"
}

# recode ONE record where West==1 & Central==1 as Central
f3 = function(w,c){
  if(w=="0" & c=="0") i="1"
  else i="0"
}

# create new "region"" variable
crime2$region = as.factor(mapply(f1,crime2$west,crime2$central))
crime2$region_urbrul = as.factor(mapply(f2,crime2$west,crime2$central, crime2$urban))

# arranging "region" factor levels
crime2$region = factor(crime2$region, levels = c("West", "Central", "East"))
crime2$region_urbrul = factor(crime2$region_urbrul, levels = c("W_Rural", "W_Urban", "C_Rural", "C_Urba

# create new factor variable "east"
crime2$east = as.factor(mapply(f3,crime2$west,crime2$central))
```

Next, we compare boxplots of *crmrte* by *region* to better understand regional effects. From figure 3A, we observe that the *west* region has lower crime rates than *central* and *east*. By segmenting each region into urban and rural (a proxy for density levels), we confirm once again from figure 3B below that *west* has lower crime rates across both urban and rural counties. We explore the relationships between *region* and our key variables of interest in more detail in the following sections.

```
g1 <- ggplot(crime2, aes(x=region, y=crmrte, fill=region)) + geom_boxplot() + ggtitle("Crime Rates by R

g2 <- ggplot(crime2, aes(x=region_urbrul, y=crmrte, fill=region)) + geom_boxplot() + ggtitle("Crime Rat

plot_grid(g1, g2, labels = "AUTO")
```

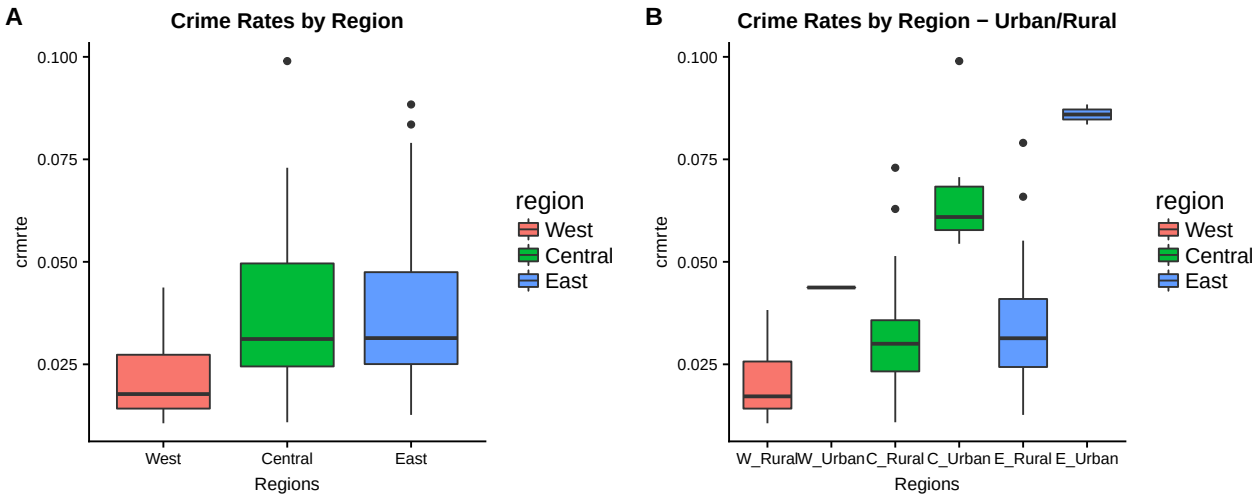

Figure 3: Boxplots of Crime Rates by Region

2.2.5 Weekly Wage Service variables (wser)

We examined all 9 weekly wage variables (figure 4) and noticed an extreme outlier in the *wser* variable

(2177.068). Upon further examination, we attribute this to likely input error i.e. value might have been entered as 2177.068 instead of 217.7068 by mistake. If this was truly a high wage value for the county, we would expect to see similar above average weekly wage levels in other industries which does not seem to be the case.

```
summary(crime2$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.0   253.2   275.7   278.1  2177.1
```

```
crime2[crime2$wser > 1000, c(1, 15:23)]
```

```
##    county     wcon    wtuc     wtrd     wfir     wser   wmfg   wfed   wsta
## 84    185 226.8245 331.565 167.3726 264.4231 2177.068 247.72 381.33 367.25
##      wloc
## 84 300.13
```

```
par(mfrow=c(2,5))
hist(crime2$wser, col="red")
hist(crime2$wcon)
hist(crime2$wtuc)
hist(crime2$wtrd)
hist(crime2$wfir)
hist(crime2$wmfg)
hist(crime2$wfed)
hist(crime2$wsta)
hist(crime2$wloc)
```

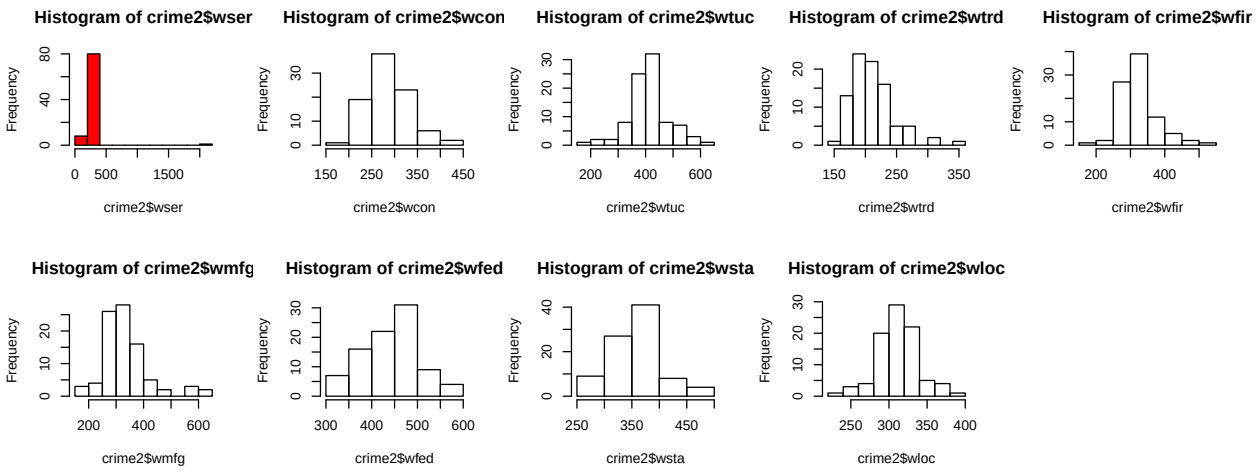

Figure 4: Histograms of weekly wage variables

While we are quite certain that 2177.068 is incorrect, we wanted to understand the impact on our potential models. We ran the Cook's distance this field is well outside the 1 boundary (figure 5).

```
# Check for Cook's distance of error value
par(mfrow=c(1,2))
plot(jitter(crime2$crmrte), jitter(crime2$wser), xlab = "Wage in service industry", ylab = "Crime Rate"
model1 = lm(crmrte ~ wser, data = crime2) # fit the linear model
abline(model1) # Add regression line to scatterplot
plot(model1, which=5)
```

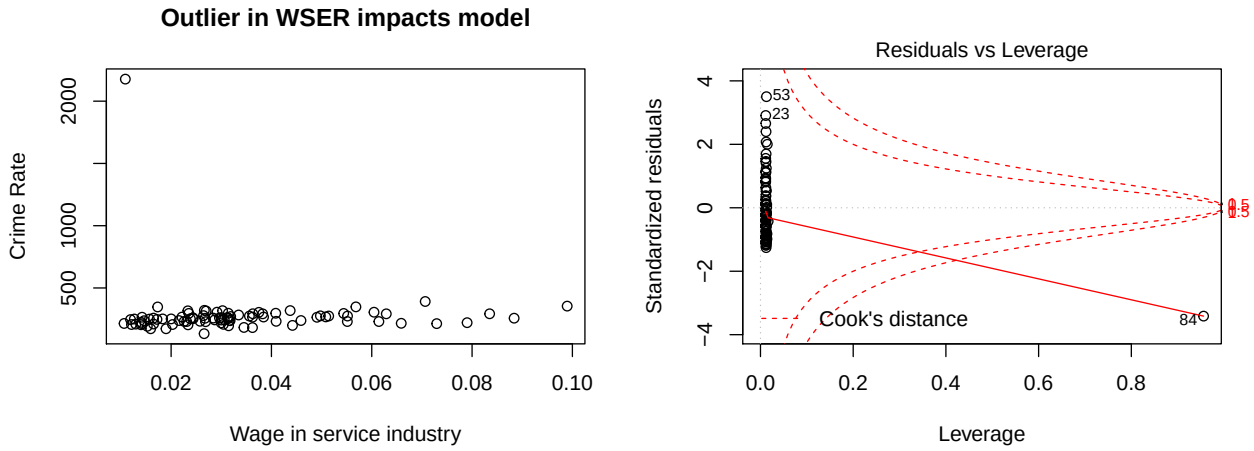

Figure 5: Cook's Distance of wser Outlier

Given the high leverage and influence this outlier will have on our models, we decided to remove this outliner from our dataset.

```
# remove outlier from dataset
crime2 <- crime2[!crime2$wser > 2000,]

hist(crime2$wser, main="Crime2$wser (excluding outlier)")
```

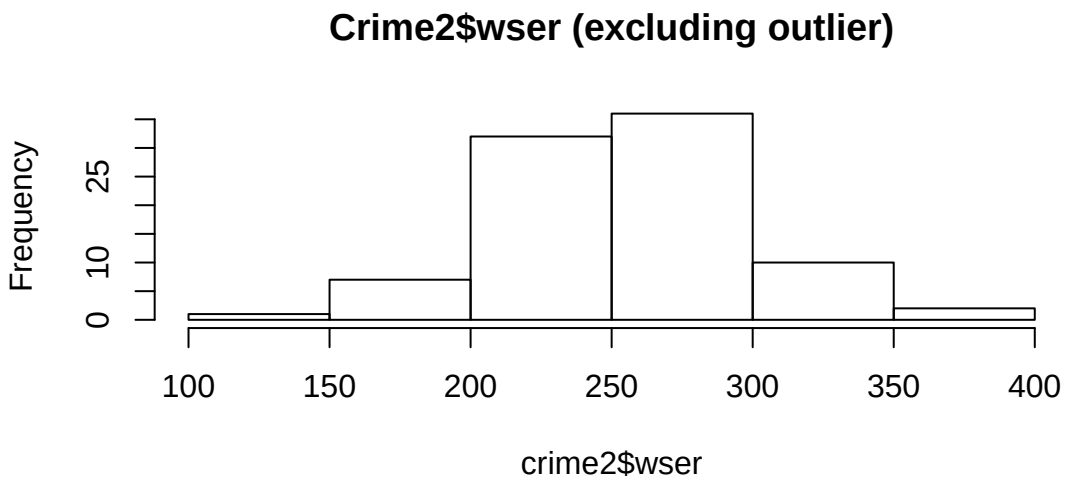

Figure 6: Histogram of Weekly Wage Service Industry

As we motivate to understand the relationship between income inequality and crime rate, we will begin to look initially at the crime rate variable in isolation and build from there to introduce more key variables for our proposed model.

2.3 Crime Rate Overview and Bivariate Analysis

```
summary(crime2$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## 0.01062 0.02201 0.03002 0.03409 0.04088 0.09897
```

```
hist(crime2$crmrte, xlim=c(0,0.1), main = "Histogram of Crime Rate",
     xlab = "crimes committed per person")
```

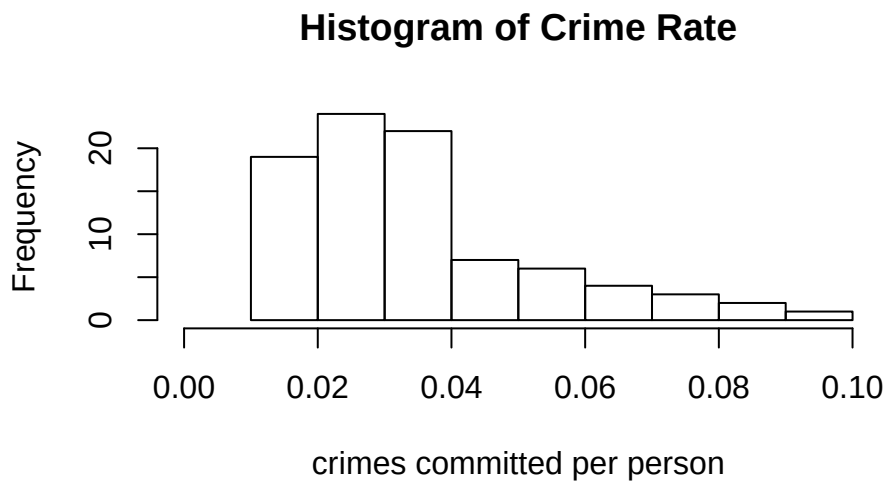

Figure 7: Histogram of Crime Rate

The histogram of crime rate in figure 7 shows a slightly right-skewed distribution with a median of 0.03002 crimes committed per person. Next, we examine the relationship between crime rate and the rest of our variables through a correlation matrix.

```
# reorder the data frame
matrix_data <- crime2[, c(3, 9, 15:23, 4:8, 24, 10, 14, 25)]

# build the correlation model
correlation_all <- round(cor(matrix_data), 1)

# print the correlation matrix
ggcorrplot(correlation_all,  outline.col="white", ggtheme = ggplot2::theme_dark, type = "lower", lab = T
```

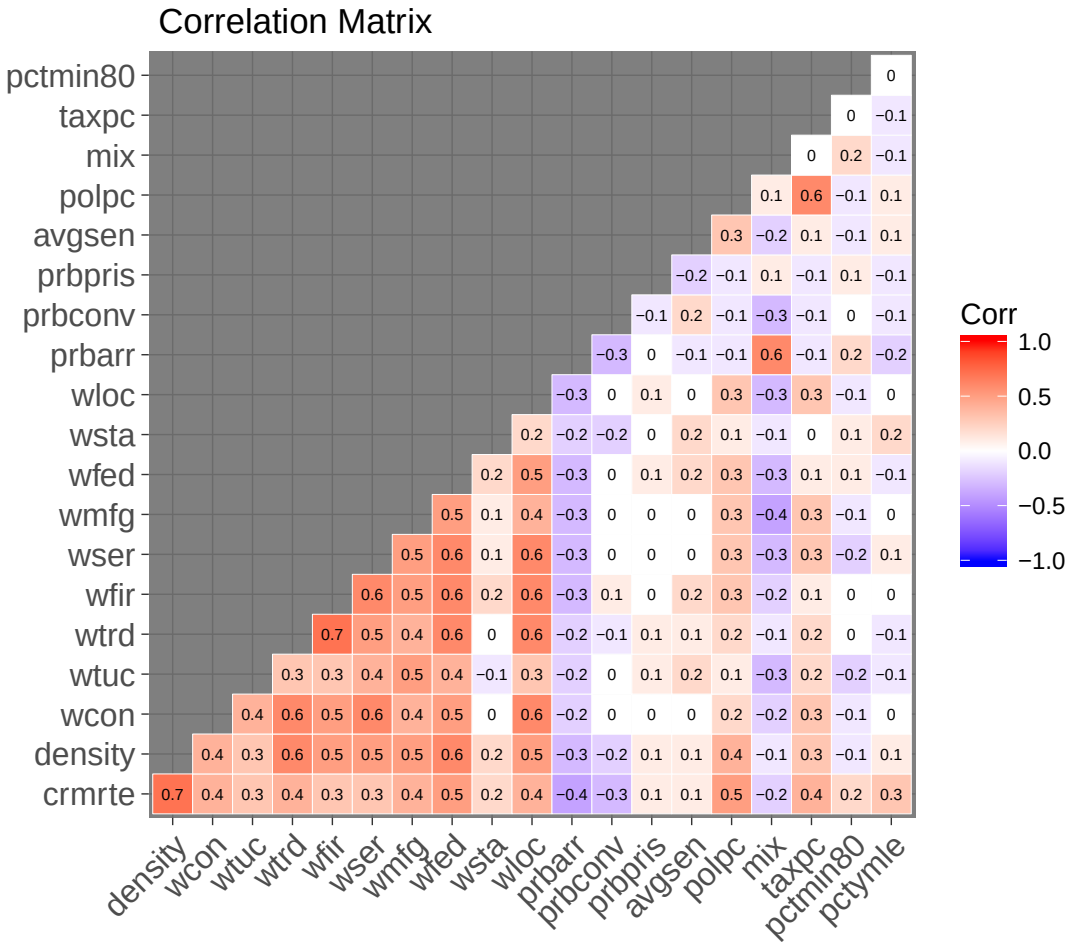

Figure 8: Correlation Matricies

The correlation heat map (figure 8) enables us to understand relationships between variables within the data, and is an excellent starting point for bi-variate investigation. Our dependent variable *crmrte* is on the bottom row, and each square represents the correlation across the x axis. The first variable is density, the largest correlation with *crmrte* followed by logical groupings carrot (wage) variables, stick (deterrent) variables and finally with the remaining demographic variables.

Looking at density, we notice it has a very strong association with crime rate (correlation: 0.7263491). That seems logical when we consider that a large population of people living in close proximity with different incomes provides motivation and access to crime. From this diagram, density seems like an important control variable, and will research it further in section 2.3.2. Our region indicators could not be included in this diagram, but we wonder how region and density interact and will investigate this further in section 2.3.3.

We first focus in on crime rate and wages since they directly relate to our task. The correlation matrix indicates a positive relationship between wages and crime rate. This makes sense when we notice that wages on the whole are correlated with density. The correlation with crime rate varies among the 9 wage variables between .02 and .05. As we design our models we will need to consider multi-collinearity and our research question. In section 2.3.1. we will explore the relationship between crime rate and income inequality, using wage variables as proxies for income inequality.

On the oppositional policy side, conventional wisdom would dictate each variable representing likelihood and severity of punishment to result in a lower crime rate in a negative relationship. This is true in some cases. *prbarr* (correlation: -0.3999229) and *prbconv* (correlation: -0.3428826). That would confirm an intuitive sense that the higher the probability of arrest and conviction, the lower the crime rates. However, the other two

"stick" variables we have identified, *prbpris* and *avgsen*, are slightly *positively* correlated - indicating that these variables have very little impact on crime rates. This implies that capturing and prosecuting criminal activity might reduce crime, but the actual punishment (prison) and severity (length of sentence) are not influencing *crmrte*.

Pctmin80 and *pctymle* variables show some correlation with crime rate, (correlation: 0.2010615) and (correlation: 0.2806584), respectively. From our initial EDA, these fields look like high quality data points, but are not particularly relevant.

Finally, this matrix also raises some surprising results that raise questions for our studies motivation or their inclusion in the model. *Polpc* has a strong POSITIVE correlation (.5) to *crmrte*; however it seems counter-intuitive to say "police determine crime". More likely, police levels have increased due to crime. Creating a causality loop between two variables introduces endogeneity into a model, and it cannot be considered an independent variable. While *polpc* is a variable that is intended to decrease future crime, time series models that support complex relationships between dependent and independent variables are out of scope for this project.

The *taxpc* also has an unexpected result, as it does not seem to correlate to higher incomes. In fact, the majority of counties have a taxpc rate between $30 and $41 with a mean of 38.055. This rate seems surprisingly flat and low, and we wonder which taxes are included. Does this value include a real estate tax - which would be a wealth indicator or a sales tax - which is born by all income levels. Given the low dollar amount per capita of *taxpc* (mean of 38.2440399 dollars per capita) it is safe to say this variable doesn't represent the total taxes collected per person and is therefore not a variable that would help us understand anything about the general population.

2.3.1 Wages

Given we have 9 different weekly wage variables, we will transform these into one **wage gap variable** that will serve as a **proxy for wage inequality**. We will use the difference in maximum and minimum weekly wage variables across industries for each county to calculate a weekly *wage_gap* proxy variable for each county. Ideally, we would have more granular data on the max and min income levels (e.g., average firm wage gap across firms by county), but given our limited data set we hope to use the simple extremes to regress *crmrte* on our underlying motivation for this analysis.

```
# Transform into proxy for income inequality. wage_gap is the difference in Max/Min weekly wages per co
f4 = function(a,b,c,d,e,f,g,h,i){
  diff = max(a,b,c,d,e,f,g,h,i) - min(a,b,c,d,e,f,g,h,i)
  return(diff)}

crime2$wage_gap = mapply(f4, crime2$wcon, crime2$wtuc, crime2$wtrd, crime2$wfir, crime2$wser, crime2$wm
```

```
par(mfrow=c(1,2))
hist(crime2$wage_gap, breaks = 30, main = "Histogram of Wage Gap",
     xlab = "Difference between max & min weekly wages")
plot(crime2$wage_gap, crime2$crmrte,
     xlab = "Difference between max & min weekly wages", ylab = "crime rate",
     main = "Crime Rate by Wage Gap")
abline(lm(crime2$crmrte ~ crime2$wage_gap))
```

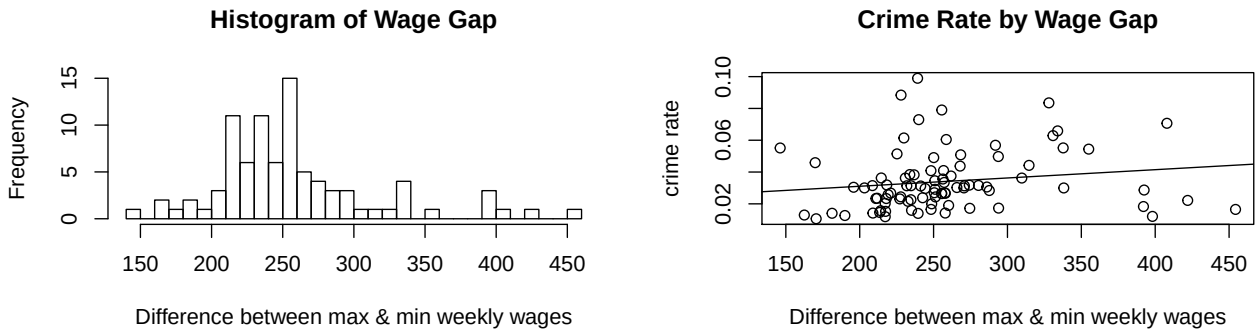

Figure 9: Histogram and Scatterplot of Wage Gap

The histogram of *wage_gap* in figure 9 shows a slightly right-skewed distribution with a median weekly wage differential of 249.4025116 between max and min weekly wages. From the above scatterplot, we observe an overall positive linear relationship between *wage_gap* and *crmrte* in the sample whereby counties with higher wage gap levels also tend to have higher crime rates. We explore this relationship further through a boxplot of *crmrte* by *wage_gap* levels as well as a plot of *crmrte* means by *wage_gap* levels.

```
summary(crime2$wage_gap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   146.2   224.2   249.4   258.1   271.7   454.4
```

```
# Create wage gap bins
wage_gap_bin = cut(crime2$wage_gap, breaks = c(-Inf,summary(crime2$wage_gap)[2], summary(crime2$wage_gap
     c("Low", "Average", "High"))
summary(wage_gap_bin)
```

```
##     Low Average    High
##      22      44      22
```

```
# plot of means by wage gap level
crime_means_wage = by(crime2$crmrte, wage_gap_bin,  mean, na.rm = T)

par(mfrow=c(1,2))
boxplot(crime2$crmrte ~ wage_gap_bin,
      xlab = "wage gap level", ylab = "crime rate",
     main = "Boxplot of Crime Rate by Wage Gap")
plot(sort(unique(wage_gap_bin)), crime_means_wage,
     xlab = "wage gap level", ylab = "mean crime rate",
     main = "Mean Crime Rate by Wage Gap")
```

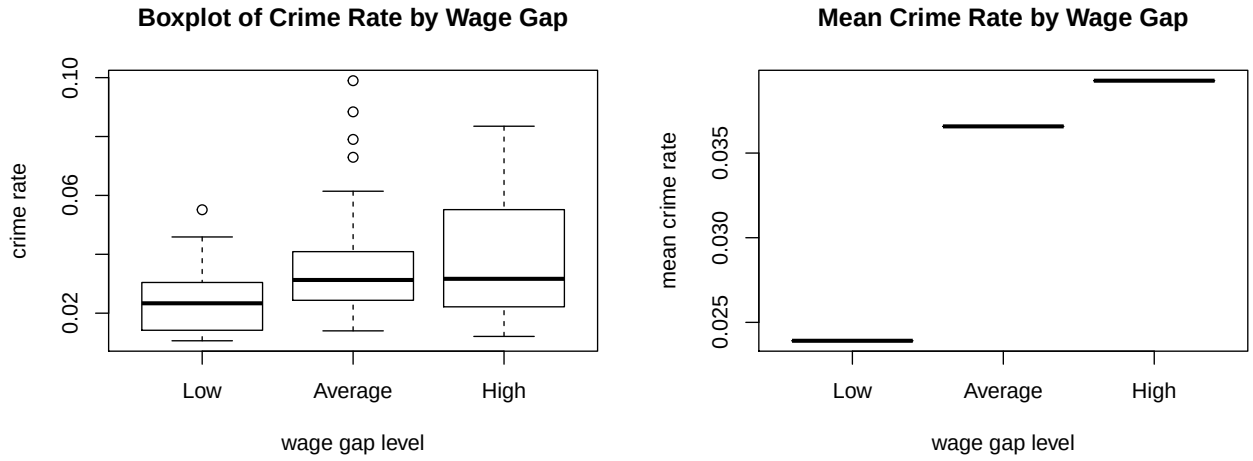

Figure 10: Boxplot and Plot of Means of Crime Rate by Wage Gap

The boxplot and plot of means in figure 10 confirms the positive relationship we observed between *wage_gap* and *crmrte*. By binning the *wage_gap* variable, we can clearly see that as wage gap levels increase, crime rates increase as well. We observe a big jump in *crmrte* between low to average wage gap levels (figure 10).

2.3.2 Density

Given the strong correlation we observed in the scatterplot matrix between *density* and many of the key variables, it is important that we understand the relationship between *density* and *crmrte* (dependent variable) as well as *wage_gap* (independent variable of interest) and control for this during our model building process, if necessary. The histogram of *density* in figure 11 shows a right-skewed distribution with a median density of 1.0007541 people per sq. mile.

```
# Create density bins for the levels of density
density_bin = cut(crime2$density, breaks = c(-Inf,summary(crime2$density)[2], summary(crime2$density)[5]
     c("Low", "Average", "High"))

# plot of means by density level
crime_means = by(crime2$crmrte, density_bin,  mean, na.rm = T)

par(mfrow=c(1,4))
hist(crime2$density, breaks = 30, main = "Histogram of Density",
     xlab = "people per sq mile")
plot(crime2$density, crime2$crmrte,
     xlab = "density", ylab = "crime rate",
     main = "Crime Rate by Density")
abline(lm(crime2$crmrte ~ crime2$density))
boxplot(crime2$crmrte ~ density_bin,
      xlab = "density level", ylab = "crime rate",
     main = "Boxplot of Crime Rate by Density")
plot(sort(unique(density_bin)), crime_means,
     xlab = "density level", ylab = "mean crime rate",
     main = "Mean Crime Rate by Density")
```

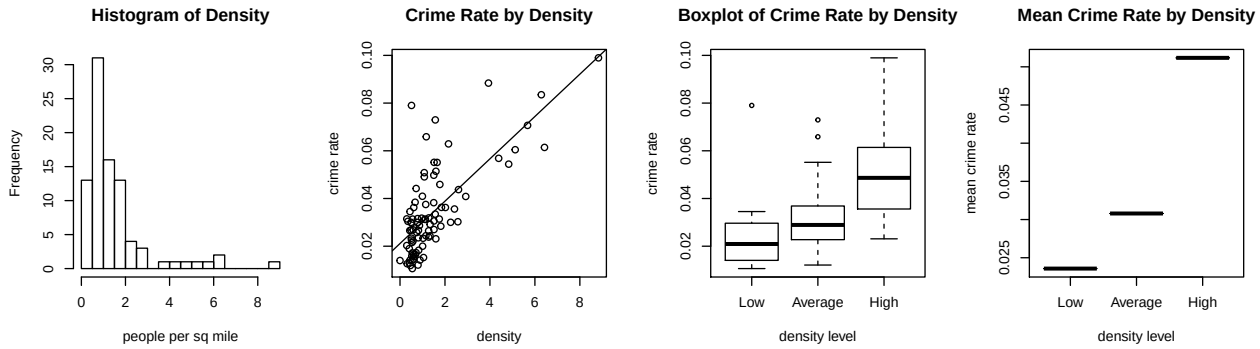

Figure 11: Relationship between Crime Rate & Density

The scatterplot in figure 11 ("Crime Rate by Density") confirms an overall positive relationship between *crmrte* and *density* (correlation: 0.7263491) which we had observed in earlier sections. The boxplot (figure 11, "Boxplot of Crime Rate by Density") and plot of means (figure 11, "Mean Crime Rate by Density") also confirm the positive relationship we observed between *crmrte* and *density* i.e., as density levels increase crime rates increase. We observe a big jump in *crmrte* between average and high density level counties.

```
# plot of means by density level
wagegap_means_density = by(crime2$wage_gap, density_bin, mean, na.rm = T)

par(mfrow=c(1,3))
plot(crime2$density, crime2$wage_gap,
     xlab = "density", ylab = "wage gap",
     main = "Wage Gap by Density")
abline(lm(crime2$wage_gap ~ crime2$density))
boxplot(crime2$wage_gap ~ density_bin,
      xlab = "density level", ylab = "wage gap",
     main = "Boxplot of Wage Gap by Density")
plot(sort(unique(density_bin)), wagegap_means_density,
     xlab = "density level", ylab = "mean wage gap",
     main = "Mean Wage Gap by Density")
```

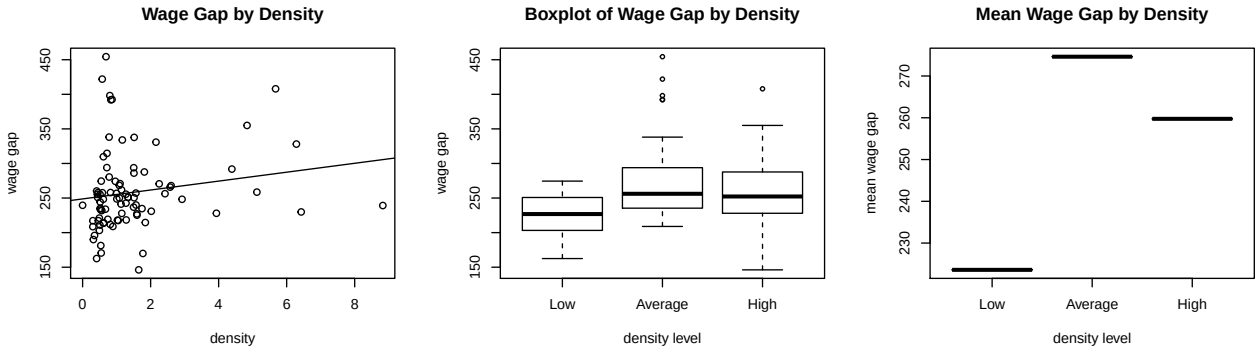

Figure 12: Relationship between Wage Gap & Density

In figure 12, although not as strong a correlation, there appears to be an overall positive relationship between *wage gap* and *density* (correlation: 0.1716091) where wage gap increases as density levels increase. There is greater variance in wage gap at the high density level which makes sense as types of jobs and levels of pay

diverge more in bigger cities. The relationship is more pronounced, i.e. we observe a big jump in *wage gap*, between low and average density levels.

Given the clear interaction of *density* with both our dependent and independent variables, we will make sure to **control for *density* in our models**.

2.3.3 Region

Next, we examine the *region* variable to see if there are any region specific impacts that are not explainable by density we should be controlling for as well. From the boxplot in figure 13, we observe a clear impact of *region* variable on *crmrte* when compared across similar density levels. Therefore Region and Density are not perfectly co-linear and have separate impacts on the data. We will therefore explore incorporating *region* as a dummy variable and control variable into our model building process.

```
# double check for region variable impact controlled for density levels

g1 <- ggplot(crime2, aes(x=density_bin, y=crmrte)) + geom_boxplot(aes(fill=region), position = position_

g1
```

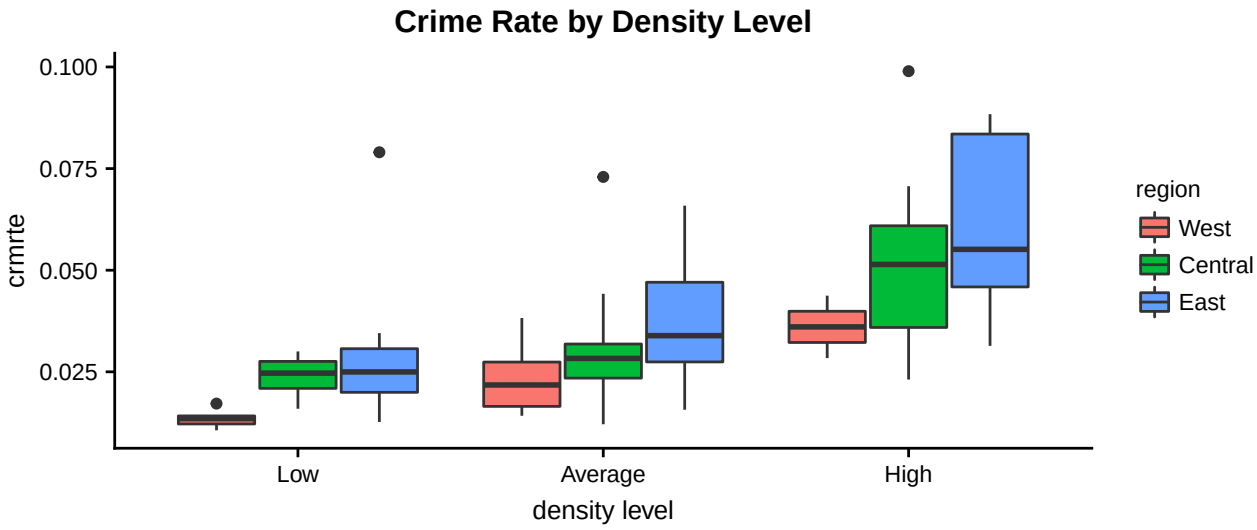

Figure 13: Relationship between Crime Rate, Density and Regions

2.3.4 Non Relevant Variable Related to Our Motivation to Understand Wage Gap on Predicting Crime Rate

As we saw above looking at correlation, there are certain variables we are not exploring in this analysis focused on the potential for wage gap reduction to reduce crime. We reprint earlier comments here for ease of understanding:

The correlation matrix also raises some surprising concerns that raise questions about the validity of data or their inclusion in the model. *Polpc* has a strong POSITIVE correlation (.5) to *crmrte*; however it seems counter-intuitive to say "police determine crime". More likely, police levels have increased due to crime. Creating a causality loop between two variables introduces endogeneity into a model, and it cannot be considered an independent variable. While *polpc* is a variable that is intended to decrease future crime, time series models that support complex relationships between dependent and independent variables are out of scope for this project. The *taxpc* also has an unexpected result, as it does not seem to correlate to higher incomes. In fact, the majority of counties have a taxpc rate between $30 and $41 with a mean of 38.055.

This rate seems surprisingly flat and low, and we wonder which taxes are included. Does this value include a real estate tax - which would be a wealth indicator or a sales tax - which is born by all income levels. Given the low dollar amount per capita of *taxpc* (mean of 38.2440399 dollars per capita) it is safe to say this variable doesn't represent the total taxes collected per person and is therefore not a variable that would help us understand anything about the general population.

2.4 EDA Summary

This data analysis has enabled us to make the following decisions about our data and how we will use it in our models.

Firstly, we have identified that crime rate is a fairly clean field that can operate as our dependent field. We will also leverage the cleansed wage fields provided to act as a proxy for income inequality i.e. the "carrot" variable. Ideally, we would have more information about the spread of wealth (income inequality index and the distribution of wealth per county), but we will attempt to use this dataset instead. *prbarr*, *prbconv*, *prbpris*, and *avgsen* will serve as proxies for the "stick" variable. These variables will all be used to help directly answer our research questions.

Next, we have two variables that we will include in our models as control variables: *density* and our transformed *region* field. These fields seem highly correlated to our dependent variable, but a deep dive into urbanization policy and regional factors is outside the scope of this study. Therefore, we will include these fields only to remove the effects from the variables that we are studying.

We have two variables that we choose not to consider in our models: *polcpc* and *taxpc*. The police force clearly have a strong role to play in deterring, arresting, and convicting criminals. However, the field *polcpc* is not a clear measure of the effectiveness of the police force. Because police per capita is often a reaction to the crime rate, our team felt that including this field would introduce a great deal of endogeneity since there is a loop of causality between police per capita and crime.

We also chose to exclude taxes per capita because we do not have a strong understanding of how North Carolina county taxes are assessed. Is it a property tax that would indicate wealth? Is it a sales tax that would not necessarily reflect wealth, but instead spending? A combination of density and *taxpc* could indicate how much revenue a county has, but we do not, at this time, understand their challenges and budget. Therefore, we chose to leave out these variables that could potentially make our models misleading.

Finally, we did not investigate the demographic variables *pctmin80* and *pctymle* as they are unrelated to our core question of wage gap and police enforcement core to this study.

3 Model Development - Predicting Crime Rates while Controlling for Density and Region

Now that we have a clear understanding of the data and which fields are reliable and related to our research question, we focus on our model building task.

3.1 Building the model

For ease of interpretation, we will look at the elasticity between *crmrte* and *wage_gap* i.e. 1% change in wage gap affect x% change in crime rates by logging both variables. Our models can be represented as:

Model 1 We are interested in exploring the relationship between crime rate and income inequality and will use *wage_gap* variable as a proxy for income inequality. Since *density* is highly correlated with most variables including our outcome variable, we will control for *density* in our model and also include *region* dummy variable to control for regional effects.

$$crmrte = \beta_o + \beta_1\ wagegap + \beta_2\ density + \beta_3\ west + \beta_4\ east + u$$

Model 2 We will introduce *prbarr*, *prbconv*, *prbpris* and *avgsen* into the model as proxies for our "stick" variables (both certainty and severity of punishment). From this model, we should be able to understand and compare the impacts of "carrot" vs. "stick" variables while controlling for *density* and *region* effects.

$$crmrte = \beta_o + \beta_1\, wagegap + \beta_2\, density + \beta_3\, west + \beta_4\, east + \beta_5\, prbarr_pct + \beta_6\, prbconv_pct + beta_7\, prbpris_pct + beta_8\, avgsen + u$$

Model 3 For our final model, we will add in effectively all variables (*mix*, *pctmin80* and *pctymle*) with the exception of *polpc* and *taxpc* due to reasons explained in previous sections. The intention of this last "kitchen sink" model is to demonstrate the robustness of our results to model specifications.

$$crmrte = \beta_o + \beta_1\, wagegap + \beta_2\, density + \beta_3\, west + \beta_4\, east + \beta_5\, prbarr_pct + \beta_6\, prbconv_pct + beta_7\, prbpris_pct + beta_8\, avgsen + \beta_9\, m$$

```
# estimate the models
m1 <- lm(log(crmrte) ~ log(wage_gap) + density + west + east, data = crime2)

m2 <- lm(log(crmrte) ~ log(wage_gap) + density + west + east + prbarr_pct + prbconv_pct + prbpris_pct +

m3 <- lm(log(crmrte) ~ log(wage_gap) + density + west + east + prbarr_pct + prbconv_pct + prbpris_pct +
```

3.2 Inference

In this section, we explore if and how we can apply our linear modeling findings to the general population.

3.4.1 Model 2 - Classical Linear Model Assumptions

MLR.1 Linear population model: Since we did not constrain the error term, there is no need to check the linear population model.

MLR.2 Random Sampling: 97 observations were collected out of the 100 North Carolina counties. As outline in our EDA in section 2, we assume that the 6 records that were excluded due to empty *crmrte* data were random. We removed an additional 3 records, 1 given the consistent extreme outlier of data in county 114 (section 2.2.2), 1 record due to duplicate entries (section 2.2.3) and finally 1 record that had what we suspect is a typo on *wser* which given our motivation on measuring *wage_gap* could have high leverage and influence on our model (section 2.2.5). Our final data set includes 88 observations which will in general allow us to leverage OLS asymptotics and the Central Limit Theorem as we explore the further CLS assumptions.

MLR.3 No perfect multicollinearity: R produced results of our models where each variable had a unique coefficient. R would not be able to do this if we had perfect multi-collinearity. We can also use the VIF function to check. As we can see below, the variance inflation factors are all way below 10 which indicates we do not have a multicollinearity issue.

```
vif(m2)
```

```
## log(wage_gap)       density          west          east    prbarr_pct
##      1.078437      1.427924      1.387584      1.457338      1.376707
##   prbconv_pct   prbpris_pct        avgsen
##      1.278746      1.056782      1.107158
```

MLR.4 Zero-conditional mean E(u|x) = 0: In order to determine if we have a zero conditional mean, we can use the residuals vs fitted plot. The red line provided by R does not appear straight particularly on the right tail, so we are in violation of this assumption. We will explore variable transformations to limit any potential violations.

```
plot(m2, which=1)
```

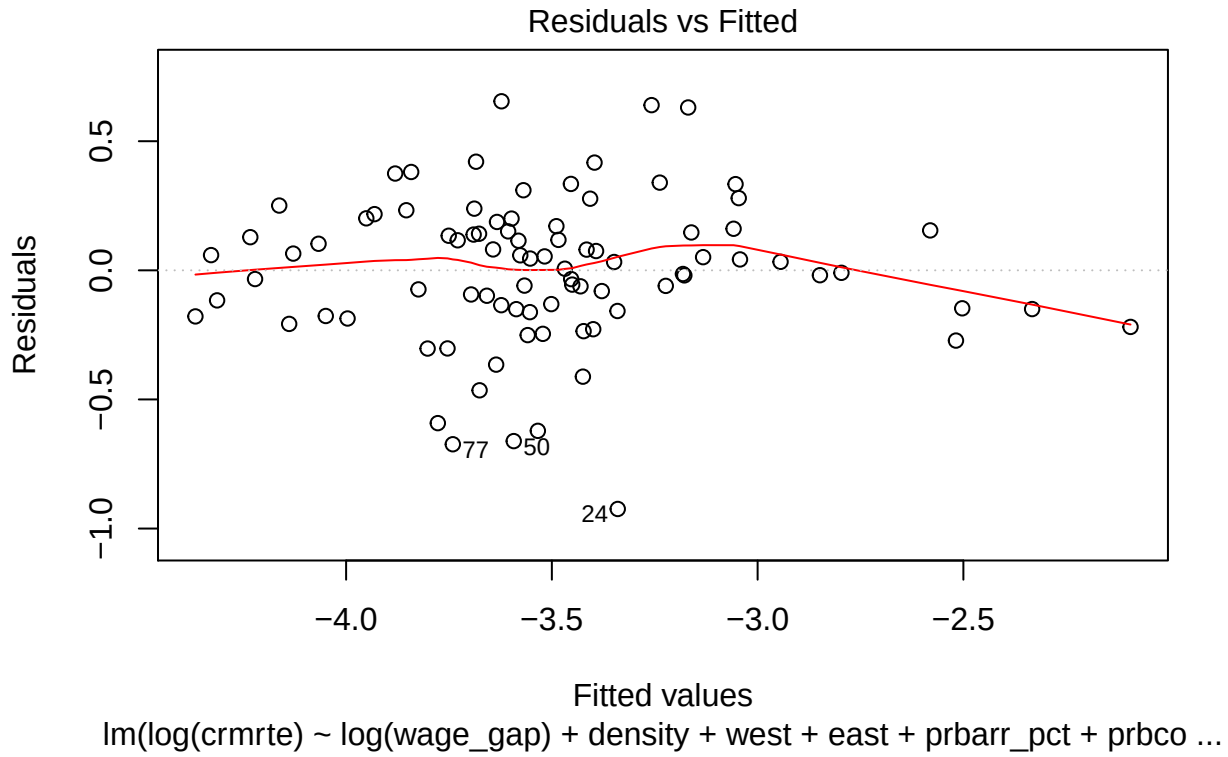

Figure 14: Model 2 - Residuals vs. Fitted Plot

MLR.5 Homoskedasticity: From the Residuals vs Fitted plot in figure 14, the residuals do not appear to consistently surround the red line. This indicates heteroskedacity. To further confirm assumption 5, we will run the Breush-Pagan test since we have a sample over 30 observations. Because the p-value is large - we fail to reject the null hypothesis of "The hypothesis is homoscedastic"

```
bptest(m2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m2
## BP = 7.9985, df = 8, p-value = 0.4336
```

To further confirm this model's homoskedasticity, we run the scale location plot. It is clear from figure 15, that the line is not straight and therefore heteroskedastic. We will proceed to use heteroskedasticity-robust standard errors.

```
plot(m2, which =3)
```

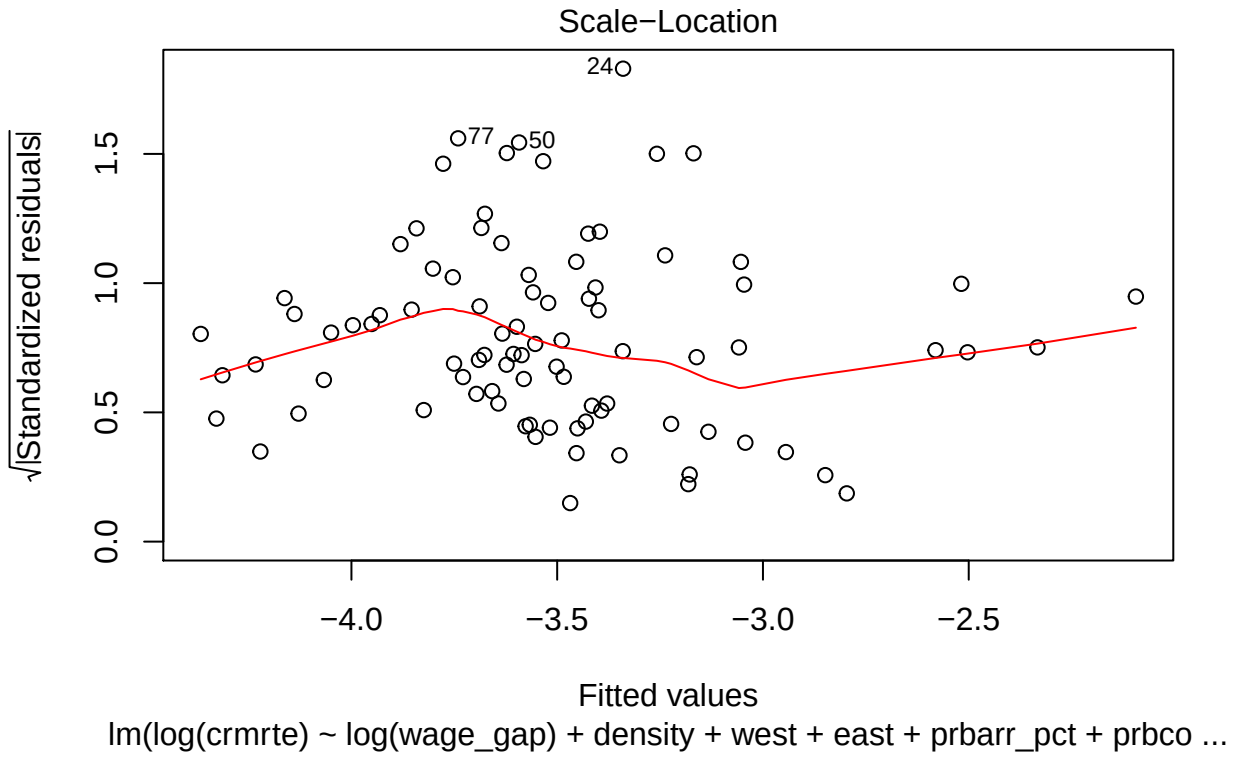

Figure 15: Model 2 - Scale-Location Plot

MLR.6 Normality of Errors: Since our data set is more than 30 records, we could assume the CLT. However, we would like to ensure we do not have a severe skew. In order for assumption 6 to be valid, the error (u) must be normally distributed with a 0 mean and a normal curve. We can evaluate it using a histogram of residuals (it should follow a normal curve) or looking at the qq plot (it should be linear). We can see by these plots in figure 16 that our data is not perfectly normal, but it is fairly close. We can also evaluate this using the shapiro test. Since our data is very close to normal by three diagnostics and we have enough records to invoke the CLT, we believe we meet this assumption.

```
par(mfrow=c(1,2))
hist(m2$residuals, main="Histogram of Model 2 Residuals")
plot(m2, which=2)
```

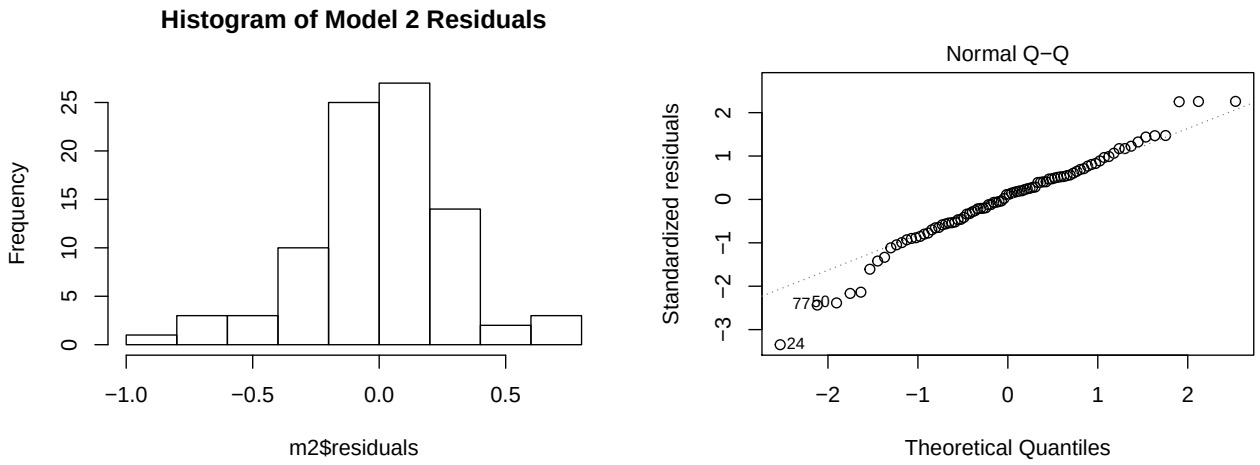

Figure 16: Model 2 - Normality of Errors Diagnostic Plots

```
shapiro.test(m2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.97326, p-value = 0.0654
```

3.4.2 Changes to Model 2 Specifications

From the above assessment, we see that our Model 2 violates two of the six MLR assumptions: zero-condition mean and homoskedasticity. To address homoskedasticity violation, we will proceed to use heteroskedasticity-robust standard errors for the rest of our analysis. To address zero-conditional mean violation, we may be able to change the functional form by checking for higher order powers using the RESET specification test.

```
# conduct RESET test on our model 2 to check for higher order powers

m2 <- lm(log(crmrte) ~ log(wage_gap) + density + west + east + prbarr_pct + prbconv_pct + prbpris_pct +

resettest(m2, power=2, type="regressor", data=crime2)
```

```
##
##  RESET test
##
## data:  m2
## RESET = 2.3465, df1 = 6, df2 = 73, p-value = 0.0397
```

In this case, p-value is significant which means our model should benefit from adding higher order polynomials to one or more variables. We will increase the order for the *density* variable by adding $density^2$.

Note that by adding higher power on *density* we have resolved the issues of zero-conditional mean as observed by the relatively flat smoothing curve on the residuals vs. fitted plot in figure 17.A. The changed model 2 also improved on homoskedasticity (flatter smoothing curve on figure 17.B) with relatively normally distributed residuals (figure 17.C). We understand that adding $density^2$ makes the *density* term more complicated to interpret; however, given density is a control variable in our model the benefits of resolving the zero-conditional mean violation was worth the change in functional form. We also take note of one data point (record 53) that has a cook's distance >1 in this new model 2 specification (figure 17.D).

```
# changes to model specification (adding powers to density variable)
m2a <- lm(log(crmrte) ~ log(wage_gap) + density + I(density^2) + west + east + prbarr_pct + prbconv_pct
```

```
par(mfrow=c(2,2))
plot(m2a, which=1, main="A", adj=0)
plot(m2a, which=3, main="B", adj=0)
plot(m2a, which=2, main="C", adj=0)
plot(m2a, which=5, main="D", adj=0)
```

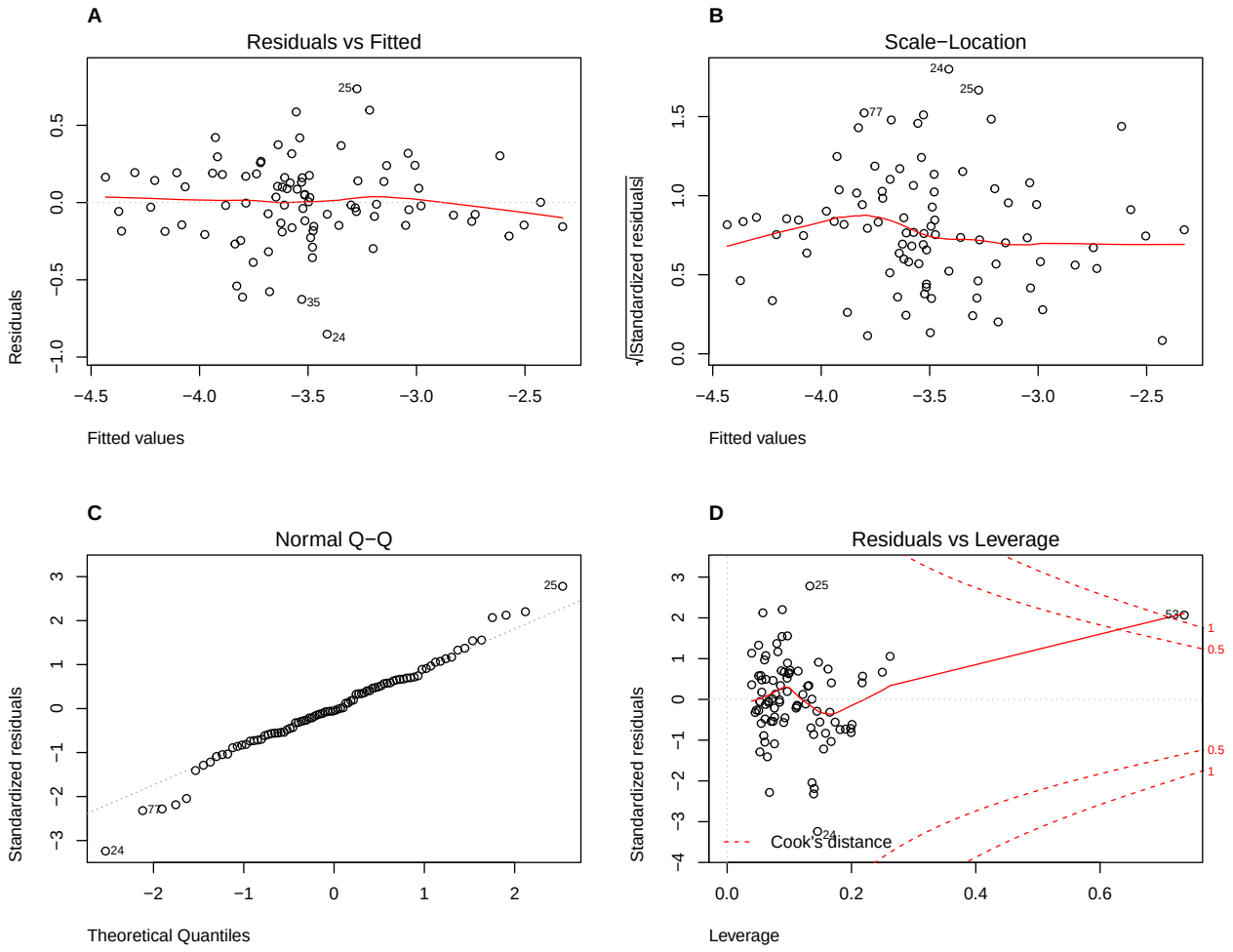

Figure 17: Diagnostic Plots for Changed Model 2

We examine the outlier identified (data record 53) and note that this county has the maximum values within the dataset for *crmrte* (0.0989659) and *density* (8.827652) which explains the high leverage and cook's distance. We decide to leave this record in the dataset given this is a valid record.

```
crime2[52,]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 53    119   87 0.0989659 0.149094  0.3478 0.486183   7.13 0.00223135
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 53 8.827652 75.67243    0       1     1   28.546 436.7666 548.3239
##        wtrd     wfir     wser  wmfg  wfed   wsta   wloc      mix
```

```
## 53 354.6761 509.4655 354.3007 494.3 568.4 329.22 379.77 0.168699
##       pctymle prbarr_pct prbconv_pct prbpris_pct pctymle_pct  region
## 53 0.07916495    14.9094       34.78     48.6183    7.916495 Central
##    region_urbrul east wage_gap
## 53       C_Urban    0   239.18
```

3.4.3 Models 1 & 3 - Summary Findings from Classical Linear Model Findings

For completeness, we ran analyses on the classical linear model assumptions on all our models. In this section we summarize at a high level our findings for our less relevant models. Charts to support our observations are included in appendix 1.

Model 1 was linear, randomly sampled, and contained no perfect collinearity. Unfortunately, Model 1 did not meet zero-conditional mean or homoskedasticity assumptions and did not quite fit the normal errors assumption. The errors were also only borderline normal. These are serious violations of the classical linear model assumptions, so this model would need to be refined. Based on our EDA and contextual understanding, we understand that there are likely many more independent variables that should be included on this list. Adding more variables would help create a more realistic model because of all the factors that contribute to crime rates. We purposely created this model as a baseline with the expectation that predicting crime rate is more complicated than a single variable and simple regression analysis.

Model 3 was an attempt to use all the data we had to understand if there was something important missing in our second model. We leveraged every field except *polpc* and *taxpc* which would be misleading and confusing for the model. Even with those up-front exclusions, this model still does not meet all six classical linear model assumptions. Similar to Model 1, Model 3 fulfills the objectives of linearity, absence of perfect collinearity, and randomly sampled features. However, it also fails on zero conditional mean, homoskedacity, and the plots indicate a low deviance from normality. Instead of adding more variables to this model, we would likely work on trimming some of the collinearity and consider some changes to the functional form of the variables (logging, powers, etc).

4. Regression Table

We estimated the above model equations by ordinary least squares under the assumption that no important variables have been omitted from the equation, random sampling and that the OLS estimator β_1 is unbiased. Below regression table contains the results from our 3 models:

```
se.m1 = sqrt(diag(vcovHC(m1)))
se.m2a = sqrt(diag(vcovHC(m2a)))
se.m3 = sqrt(diag(vcovHC(m3)))

stargazer(m1, m2a, m3, type = "latex", float = FALSE, font.size = "small",
          omit.stat = "f",
          se = list(se.m1, se.m2a, se.m3),
          title = "Linear Models Predicting Crime Rate",
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(m1),1), round(AIC(m2a),1),
                           round(AIC(m3),1))))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 15, 2018 - 12:05:40

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
log(wage_gap)	0.313	0.221	0.203
	(0.242)	(0.186)	(0.206)
density	0.214***	0.340*	0.151***
	(0.028)	(0.137)	(0.027)
I(density^2)		−0.025	
		(0.024)	
west1	−0.272**	−0.216*	−0.138
	(0.098)	(0.086)	(0.118)
east1	0.229*	0.287***	0.120
	(0.096)	(0.079)	(0.121)
prbarr_pct		−0.014***	−0.016***
		(0.004)	(0.004)
prbconv_pct		−0.006***	−0.006**
		(0.002)	(0.002)
prbpris_pct		0.002	−0.0001
		(0.004)	(0.004)
avgsen		−0.005	0.002
		(0.015)	(0.014)
mix			−0.340
			(0.436)
pctmin80			0.007
			(0.005)
pctymle_pct			0.020
			(0.013)
Constant	−5.580***	−4.504***	−4.419**
	(1.311)	(1.211)	(1.373)
AIC	70.1	39.9	44.6
Observations	88	88	88
R^2	0.567	0.726	0.724
Adjusted R^2	0.546	0.694	0.684
Residual Std. Error	0.347 (df = 83)	0.285 (df = 78)	0.289 (df = 76)
Note:			*p<0.05; **p<0.01; ***p<0.001

Model 1

$$\widehat{log(crmrte)} = -5.580 + 0.313\ log(wagegap) + 0.214\ density - 0.272\ west + 0.229\ east$$

The base model estimates that a 1% change in *wage gap* affects a 0.313% change in *crime rates* controlling for *density* and *region* impacts. We included *region* dummy variables using *central* as the base group. This would mean that a 10% reduction in wage gap would reduce crime rates by 3.13% ceteris paribus. Unfortunately, our *wage_gap* term has a t-statistic of about 1.29 which is statistically insignificant and means we fail to

reject the null hypothesis $H_0 : \beta_{wage\_gap} = 0$. The R-Squared statistic shows that the variables in our first model explain 56.7% of the variation in crime rate.

Model 2

$$\widehat{log(crmrte)} = -4.504 + 0.221\, log(wagegap) + 0.340\, density - 0.025\, density^2 - 0.216\, west + 0.287\, east - 0.014\, prbarr\_pct - 0.006$$

As we look to include "carrot" and "stick" variables per our research objectives, we add in the "stick" variables of *prbarr_pct*, *prbconv_pct*, *prbpris_pct* and *avgsen*. The point estimate of 0.221 on *wage_gap* means that, holding all other variables fixed, a 10% decrease in *wage_gap* would reduce *crmrte* by 2.21%. Unfortunately, the t-statistic is very small and we must conclude that from this cross-sectional analysis *wage_gap* has no effect on *crmrte*.

Looking at our "stick" variables, our model estimates that a 10 percentage point increase in probability of arrests reduces crime rates by 14% ceteris paribus. Similarly, a 10 percentage point increase in probability of conviction reduces crime rates by 6% which is a practically significant effect. We can see that among the four "stick" variables, only certainty of punishment variables (*prbarr_pct* and *prbconv_pct*) are statistically significant while severity of punishment variables (*prbpris_pct* and *avgsen*) are insignificant. Looking at regional impacts, holding all other variables constant, we can see from that *east* has 28.7% higher crime rates while *west* has 21.6% lower crime rates compared to Central.

Overall, the AIC for model 2 improves relative to model 1 and, by definition with more coefficients, the increased R-Squared shows that the variables in our second model now explains 72.6% of the variation in crime rate.

Model 3

$$\widehat{log(crmrte)} = -4.419 + 0.203\, log(wagegap) + 0.151\, density - 0.138\, west + 0.120\, east - 0.016\, prbarr\_pct - 0.006\, prbconv\_pct - 0.$$

Finally, when throwing the kitchen sink at the model, we achieve a slightly lower R-Squared and higher AIC value vs. model 2. By including more variables, we had in fact lowered the accuracy and parsimony of our model which confirms the robustness of our modeling choices in model 2.

5. Omitted Variables

We recognize that we had a limited set of data to use in our models. Here we document additional variables that would aid in our model accuracy. While we do not understand the exact impact these variables would have, we can predict the general direction of their impact. The wage gap coefficient was positive in our model, so it is always reflected as positive in the chart below.

#	Omitted Variable	Wage Gap Impact	Crime Rate Impact	Product of Impacts	B1 Coefficient	Direction
1	Average Level of Education	Negative	Negative	Positive	Positive	Away from Zero (Overestimating)
2	Indexed Strength of Unions	Negative	Negative	Positive	Positive	Away from Zero (Overestimating)
3	Government Investment Capital Projects	Negative	Negative	Positive	Positive	Away from Zero(Overestimating)
4	Unemployment Rate	Positive	Positive	Positive	Positive	Away from Zero (Overestimating)
5	Safety Net Metrics by County	Negative	Negative	Positive	Positive	Away from Zero (Overestimating)
6	Rate of Job Automation	Positive	Positive	Positive	Positive	Away from Zero (Overestimating)

Our models indicate a positive B1 coefficient.

1. *"Education is the great equalizer of our time."* said Kofi Annan. A higher average level of education for the general population would indicate a more skilled workforce capable of higher efficiency and increased wages. Our assumption is that as the general population's education level increases, the wage gap decreases (negative relationship). Likewise, a higher educated population usually indicates there are more honest opportunities and the crime rate would also decrease (negative). The product of both these negative impacts would be positive. This product combined with a positive B1 coefficient means that our omitted bias is pulling our estimate away from zero and our model is over-estimating the impact without education included in the model.

2. *"Unions reduce wage inequality because they raise wages more for low- and middle-wage workers than for higher-wage workers, more for blue-collar than for white-collar workers, and more for workers who do not have a college degree."* https://www.epi.org/publication/briefingpapers_bp143/ As lower-waged individuals come together to collectively bargain, they can demand higher wages more effectively. As the percentage of unionized workers increases, the wage gap would decrease (negative). With more blue-collar individuals earning higher wages, there will be more job opportunity alternatives for potential criminals. Therefore this impact would also be negative. When the negative/negative combination combines with the positive B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.

3. When governments invest in large capital projects (building roads and bridges), lower wage construction workers are in increased demand. As the demand grows, there will be pressure to increase wages- thereby decreasing the wage gap (negative relationship). As more people are employed and earning a good wage, criminal activity will be less tempting and crime rates will reduce (negative relationship). When the negative/negative combination combines with the positive B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.

4. A low unemployment rate creates economic pressure to raise wages in order to attract employees. Likewise, a high unemployment rate means that companies can attract employees (especially less skilled employees) at lower wages. Thus, as the unemployment rate goes up, so does the wage gap. More unemployed individuals will likely lead to higher crime rates (positive) out of necessity and a feeling on injustice. When the positive/positive combination combines with the positive B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.

5. Regions with strong safety net measures (health care, unemployment, job retraining centers) tend to have a lower wage gap since they enable lower wage workers to overcome challenges and get back to work. As safety net measures increase, the wage gap would decrease (negative). Giving these individuals the chance to get back on their feet would reduce the chances that these individuals end up committing crimes (negative). When the negative/negative combination combines with the positive B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.

6. Job Automation will prove to be a great challenge and change agent in the twenty-first century. As the rate of job automation increases, it will first impact the lower wage jobs as they can be most easily automated and impact the highest quantity. Increased job automation rates will likely lead to less demand for lower end workers, lower wages and a higher wage gap (positive). As job automation rates rise, it is reasonable to conclude that crime rates will increase (positive) as many people consider the economic situation unfair and have no other options.

Since we have also included *prbarr* and *prbconv*, we should also consider omitted variables that would impact our model 2.

#	Omitted Variable	Conviction Impact	Crime Rate Impact	Product of Impacts	B1 Coefficient	Direction
1	Percent of Arrests found not guilty	Negative	Positive	Negative	Negative	Away from Zero (Overestimating)
2	Percent of Arrests not prosecuted in court	Negative	Positive	Negative	Negative	Away from Zero (Overestimating)
3	Percent of Arrests dismissed based on a technicality	Negative	Negative	Positive	Negative	Towards Zero (Underestimating)

1. "Not Guilty" Arrest percentages would help us understand if the rate of conviction is highly influenced by the police investigative work or some sort of court backlog. This information would be critical for policy decisions (police job retraining or allocating more resources to the justice system). As arrests are found not guilty, the impact on conviction would decrease (negative). As the percentage of arrests found not guilty increases, the crime rate would also increase since conviction would seem less likely and have a less deterring impact (positive). When the negative/positive combination combines with the negative B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.
2. Understanding how many arrest charges are not prosecuted in court, presumably because of a backlog, would enable policy makers to effectively resolve systematic issues. As arrests are not prosecuted, the conviction rate will go down (negative). Lack of prosecution will also likely increase crime rates (positive). When the negative/positive combination combines with the negative B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.
3. The number of cases dismissed would illustrate criminal justice system deficiencies. As the number of cases dismissed increases, the impact on convictions would decrease (negative). Similarly, as the number of cases dismissed increases, the crime rate would likely increase (positive) since potential criminals realize there is a decreased chance of conviction. When the negative/positive combination combines with the negative B1 coefficient, the net result is that this omitted variable is likely going to pull the model away from zero and that the current model is overestimated.

7. Conclusion

It important to remember some of the decisions that we, as a research team, made in building this model. At a high level, we removed counties with missing crime rate figures or clearly invalid data, then transformed some regional data to identify the East region and finally overlaid density and region to assess the impact of geography and population distribution. By choosing the wage gap, instead of trying to build a model across individual wage categories, we are attempting to reduce the multicollinearity with our regional control variable using this calculated field as a proxy for income distribution.

Unfortunately for the Earnie Anders' campaign, the wage gap variables did not prove to be statistically significant based on their high p values. From a practical perspective, most voters understand that income inequality is a tough challenge in the 21st century. Achieving a 10% decrease in wage gap seems like an insurmountable task - even if the reward is a 2.21% decrease in crime. The omitted variables overall would only further decrease the impact on crime. However, this does not mean that the candidate is wrong. Our proxy for wage-gap is not perfect and may have had misleading results. This research question merits further investigation, especially since many other studies have linked income inequality and crime rates.

The opposition candidate seems to be correct with the stick variables of arrest and conviction. These variables were both highly statistically significant with p-values less than .001. From a practical significance, a 1% increase in probability of arrest or conviction will result in a decrease in crime rate by (1.4% or .6%,

respectively). These numbers sound more achievable through government controlled efficiency improvements. More study is required to understand how to best support the police and the court system in fairly arresting and convicting the criminals. This dataset suggests that increasing the police force in response to crime should be evaluated in a time-series analysis to understand if it is effective. It is interesting to note that the West region seems to have lower crime rates despite its density. We suggest commissioning a study to look into police and justice system best practices in that region and consider applying them across the state. Earnie should consider adopting some policies that would address increasing the arrest/conviction rate in a fair and just way such as increased police support and/or addressing any judicial overload issues.

Based on this study, the opposition is wrong to encourage harsher punishment. Spending money on more and longer prison sentences does not appear to have statistical significance. The practical significance is negligible with a 1% increase in the probability of prison resulting in .2% lower crime rates, and a 1% longer sentencing actually increasing crime rate by .5%. Earnie could effectively argue the minimal deterrent effect of severity using some of the data we have analyzed. Earnie should continue to consider policies that work to shift punishments away from prisons and towards fees, community service, and rehabilitative programs.

APPENDIX

Appendix 1 - Models 1 & 3 Classical Linear Model Assumptions

For completeness, we ran analyses on the classical linear model assumptions on all our models. In this section we summarize our findings for our less relevant models with charts to support our observations below:

Model 1 was linear, randomly sampled, and contained no perfect collinearity. Unfortunately, Model 1 did not meet zero-conditional mean (figure 18.A) or homoskedasticity assumptions (figure 18.B) and did not quite fit the normal errors assumption. The errors were also only borderline normal (figure 18.C). These are serious violations of the classical linear model assumptions, so this model would need to be refined. Based on our EDA and contextual understanding, we understand that there are likely many more independent variables that should be included on this list. Adding more variables would help create a more realistic model because of all the factors that contribute to crime rates. We purposely created this model as a baseline with the expectation that predicting crime rate is more complicated than a single variable and simple regression analysis.

$$crmrte = \beta_o + \beta_1\ wagegap + \beta_2\ density + \beta_3\ west + \beta_4\ east + u$$

```
par(mfrow=c(2,2))
plot(m1, which=1, main="A", adj=0)
plot(m1, which=3, main="B", adj=0)
plot(m1, which=2, main="C", adj=0)
plot(m1, which=5, main="D", adj=0)
```

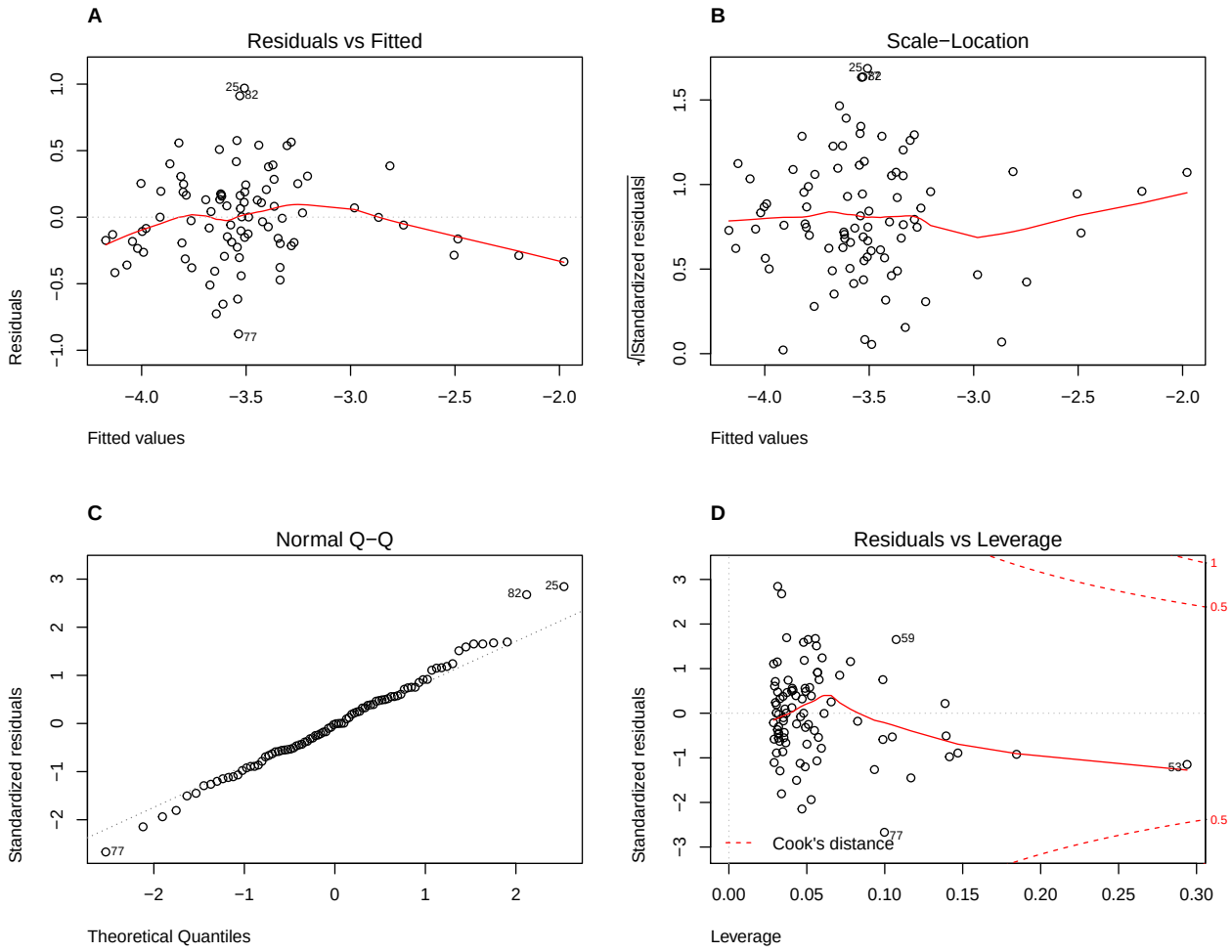

Figure 18: Model 1 - Diagnostic Plots

Model 3 was an attempt to use all the data we had to understand if there was something important missing in our second model. We leveraged every field except *polpc* and *taxpc* which would be misleading and confusing for the model. Even with those up-front exclusions, this model still does not meet all six classical linear model assumptions. Similar to Model 1, Model 3 fulfills the objectives of linearity, absence of perfect collinearity, and randomly sampled features. However, it also fails on zero conditional mean (figure 19.A), homoskedacity (figure 19.B), and the plots indicate a low deviance from normality (figure 19.C). Instead of adding more variables to this model, we would likely work on trimming some of the collinearity and consider some changes to the functional form of the variables (logging, powers, etc).

$crmrte = \beta_o + \beta_1\, wagegap + \beta_2\, density + \beta_3\, west + \beta_4\, east + \beta_5\, prbarr_pct + \beta_6\, prbconv_pct + beta_7\, prbpris_pct + beta_8\, avgsen + \beta_9\, m$

```
par(mfrow=c(2,2))
plot(m3, which=1, main="A", adj=0)
plot(m3, which=3, main="B", adj=0)
plot(m3, which=2, main="C", adj=0)
plot(m3, which=5, main="D", adj=0)
```

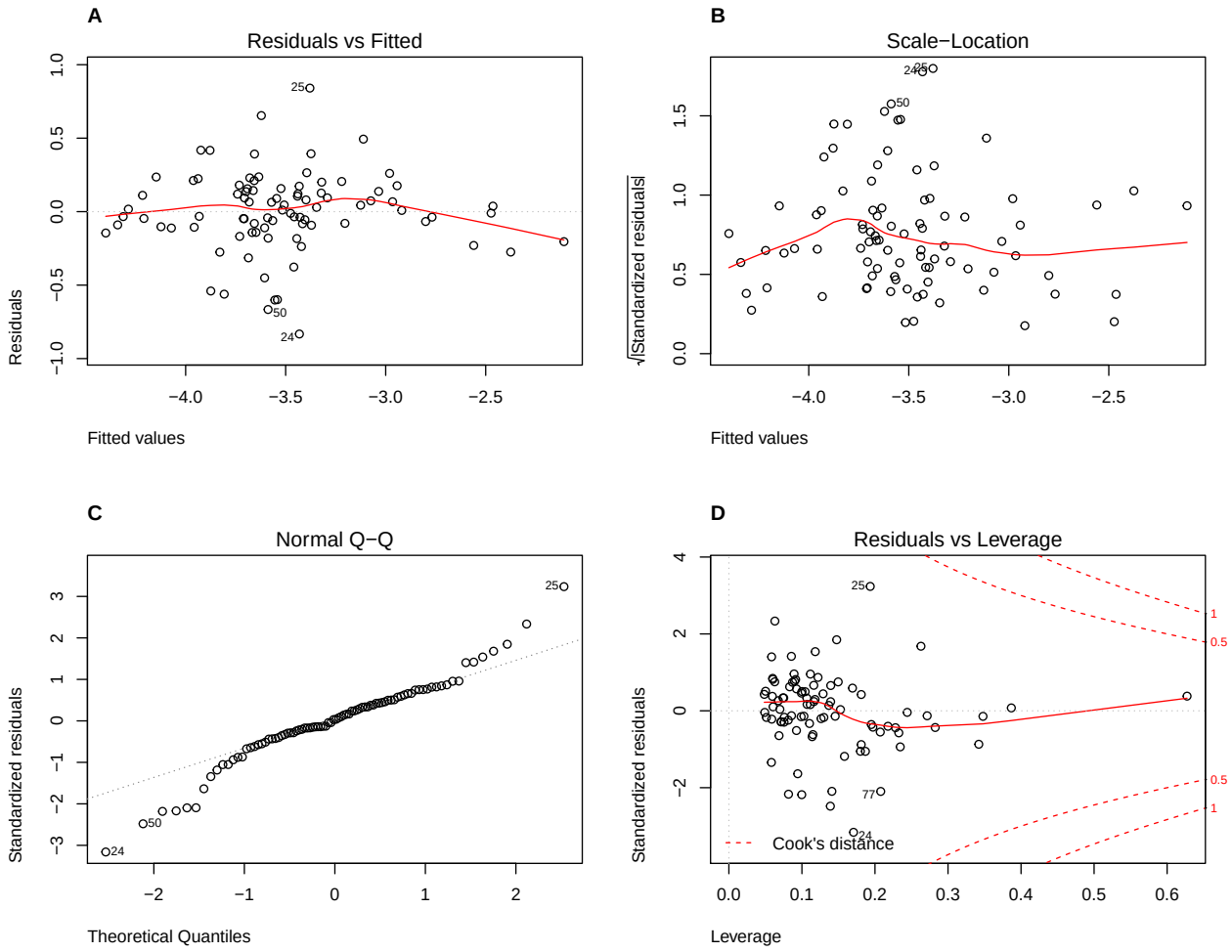

Figure 19: Model 3 - Diagnostic Plots