W203: Statistics for Data Science

LAB 3: Reducing Crime

Robert Louka Ryan Sawasaki Joshua Noble Praveen Joseph

1. An Introduction

As crime has seen an increase in the 1980's, citizens of North Carolina have been looking to local government politicians to address this growing problem. In preparation for the upcoming election, our team of political consultants has been tasked with providing insight to drive policy directed at reducing crime levels. Before pushing a political campaign aimed at crime reduction, we must first identify the key determinants of crime and their significance in order to properly focus resources to target these issues.

Many studies have examined numerous potential determinants of crimes and it remains a complex and evolving issue. Traditionally, criminal activity is often linked to issues of inequality and poverty. In addition, factors revolving around the criminal justice system are often viewed as having a significant impact, both positive and negative, on crime rate. While there is little debate that these variables affect crime, a one size fits all policy on crime does not properly address the unique issues at the state and county levels. This report aims to identify the complex interactions of crime determinants in North Carolina using recently compiled statistics from FBI and government agencies.

While many studies have been conducted on individual crime factors, this report examines multiple factors holistically. The primary research question this report addresses is: Which demographic, economic and deterrent factors significantly affect crime? To answer this question, our team has been provided a dataset of 1987 statistics from select North Carolina counties. There are 100 counties in the state, however, 10 counties have been omitted from the dataset. Given that the omitted counties comprise of a small percentage of the total population of the state (less than 2%), the results of this study are not significantly impacted by the omission of the 10 counties. The data provided for this study has been pulled from multiple credible sources including:
* FBI's Uniform Crime Reports
* FBI's police agency employee counts
* North Carolina Department of Correction
* North Carolina Employment Security Commission
* Census Data

Our dependent variable and the key measure we are focused on is crime rate, which is defined as crimes committed per person. Our independent variables have been grouped into categories of deterrent, demographic, economic, and geographical factors. A comprehensive list of the variables and their respective categories are described in our exploratory data analysis.

While this 1987 dataset provides observational variables that impact crime, the dataset does not provide a comprehensive list of all variables. There are a number of factors that our team has identified that could potentially assist in more accurately measuring a causal effect on crime. These factors are discussed in further detail in the omitted variables section of this report.

The provided dataset only covers a single cross-section of the data from the year 1987. Our team conducted additional outside research to uncover a multi-year panel of corresponding crime data to assist in cross-checking the accuracy of the wage data with the provided dataset. However, the statistical results of this report are limited to the provided 1987 data. A major issue with only using a single cross-section of data is that lag effects cannot be observed. This pertains to the police presence variable, which it is expected that the effects of an increase or decrease in police force may lag for years. In addition, lag effects may also be responsible for the probability of conviction (ratio of convictions to arrests) variable being greater than 1 in some counties. It was surmised that a probability greater than 1 was due to mulitple convictions and convictions potentially occuring years after the arrest. So in a given year, there may be more convictions than arrests.

Using this 1987 dataset, our team has conducted a study on the determinants of crime and prepared recommendations for a political strategy addressing this issue in North Carolina.

2. A Model Building Process

Exploratory Data Analysis

We started by conducting exploratory data analysis. First, we read the original paper [CORNWELL – TRUMBULL (1994)] to get a better understanding of each variable. We defined the variables in the table below and grouped them into five groups (the group construction is discussed in the groupings section below).

Descriptions and Groups of Variables

Variable	Description	Group	Note
county	county identifier	Control	
year	1987		
crmrte	crimes committed per person		ratio of FBI index crimes to county population
prbarr	'probability' of arrest	Deterrent	ratio of arrests to offenses
prbconv	'probability' of conviction	Deterrent	ratio of convictions to arrests
prbpris	'probability' of prison sentence	Deterrent	proportion of total convictions resulting in prison sentences

Variable	Description	Group	Note
avgsen	avg. sentence, days	Deterrent	average sentence in days
polpc	police per capita	Deterrent	
density	people per sq. mile	Demographic	country population divided by county land area
taxpc	tax revenue per capita	Demographic	
west	=1 if in western N.C.	Region	dummy
central	=1 if in central N.C.	Region	dummy
urban	=1 if in SMSA	Urban	dummy
pctmin80	perc. minority, 1980	Demographic	proportion of country population that is minority or nonwhite
wcon	weekly wage, construction	Wages	average weekly wage in that sector
wtuc	wkly wge, trns, util, commun	Wages	average weekly wage in that sector
wtrd	wkly wge, whlesle, retail trade	Wages	average weekly wage in that sector
wfir	wkly wge, fin, ins, real est	Wages	average weekly wage in that sector
wser	wkly wge, service industry	Wages	average weekly wage in that sector
wmfg	wkly wge, manufacturing	Wages	average weekly wage in that sector
wfed	wkly wge, fed employees	Wages	average weekly wage in that sector
wsta	wkly wge, state employees	Wages	average weekly wage in that sector
wloc	wkly wge, local gov emps	Wages	average weekly wage in that sector
mix	offense mix: face-to-face/other	Demographic	ratio of face-to-face crimes (robbery, assault, rape) to non-face-to-face crimes
pctymle	percent young male	Demographic	proportion of country population that is male between 15 and 24

To get a broad sense of the data set the summary function was run.

```
summary(data)
```

This function provides a high level view of each variable. Six rows have missing values for all variables. In addition, there is one duplicate row. Also the variable prbconv is loaded as a factor, so it needs to be converted to numeric. These issues are handled below to create the initial data set.

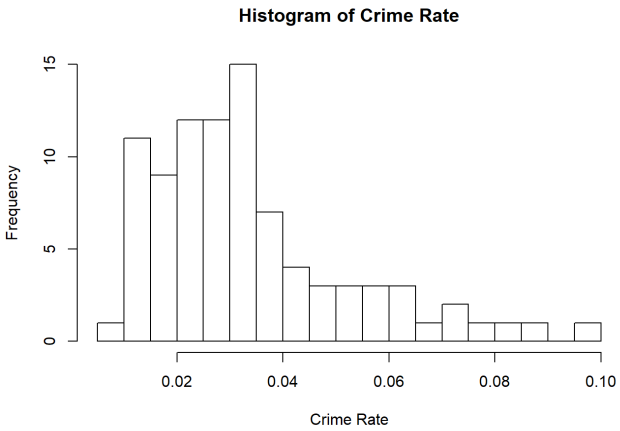
```
#eliminate N/A's (6 rows of NA were removed)
data_crmrte <- data[!is.na(data$crmrte),]

#remove duplicates (1 duplicate record was found)
data_crmrte <- data_crmrte %>% distinct()

#prbconv was defined as factor , we will convert it to numeric
data_crmrte$prbconv <- as.numeric(as.character(data_crmrte$prbconv))
#class(data_crmrte$prbconv)
```

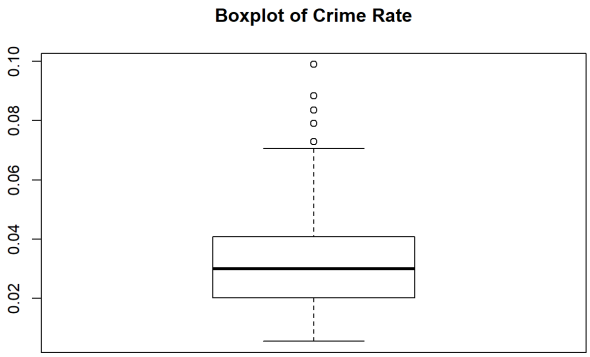
With 25 original variables in the data set the natural place to start is with the dependent variable, crmrte. To get a better sense of this variable, the distribution is graphed below.

```
quantile(data_crmrte$crmrte, c(0, .01, .05, .10, .25, .50,  .75, .90, .95, .99, 1.0))
```

```
##         0%         1%         5%        10%        25%        50%        75%
## 0.00553320 0.01006330 0.01235660 0.01418007 0.02060425 0.03000200 0.04024925
##        90%        95%        99%       100%
## 0.06054659 0.07191830 0.08954881 0.09896590
```

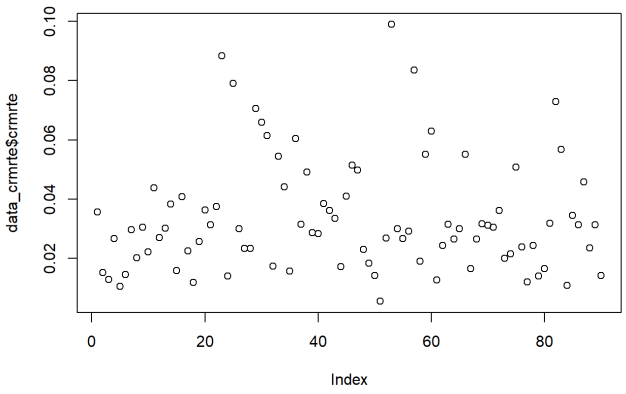
```
hist(data_crmrte$crmrte,breaks=20, xlab="Crime Rate", main="Histogram of Crime Rate")
boxplot(data_crmrte$crmrte, main="Boxplot of Crime Rate")
# The box plot and histogram show signs of right skew in the crime rate
```

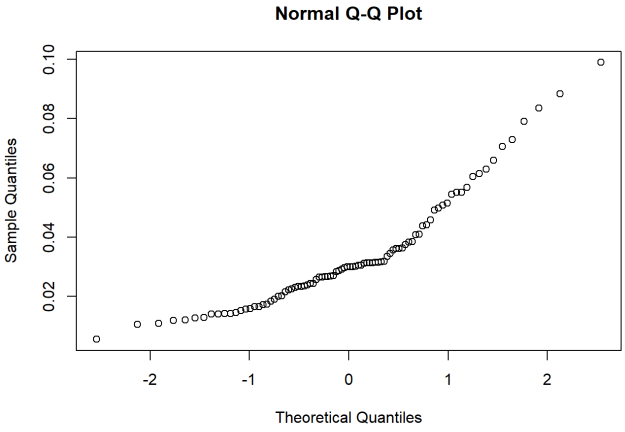





```
plot(data_crmrte$crmrte)
qqnorm(data_crmrte$crmrte) # The q-q plot shows sign of non-normality in crime rate
shapiro.test(data_crmrte$crmrte) # Shapiro-wilk test confirms the non-normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_crmrte$crmrte
## W = 0.89162, p-value = 1.741e-06
```





Outlier Analysis

There are several outliers in the variable crmrte and the distribution is right skewed. With our sample size non-normality is not a top concern but this distribution is not perfectly normal. We analyze outliers for crime rate that are > 2*Std-dev from the mean crime rate (i.e data pts with crime rate > 0.07)

The postiviely skewed outliers (6 counties) on the right side of the distribution are examined to gather some insights:

1. 4 of out of the 6 outliers are in urban areas
2. The average demographic density for the outlier set is greater than 3 times the average density for the overall sample
3. We also observe that data ppt 53 (county 119) which has the highest crime rate, also has the highest density amongst the outliers and is a urban area

This is not very surprising as we expect urban areas with high density of population to have more crimes. we will continue to monitor the impact of the outliers and conisder the treatment of these outlier in a later part of the report.

```
upper <- data_crmrte[data_crmrte$crmrte > 0.07,]
density_table <- data.frame(upper$county, upper$crmrte, upper$density)
kable(round(density_table,2), col.names = c("County", "Crime Rate ", " Density"),
      caption = "Density and Outliers")
```

Density and Outliers

County	Crime Rate	Density
51	0.09	3.93
55	0.08	0.51
63	0.07	5.67

County	Crime Rate	Density
119	0.10	8.83
129	0.08	6.29
181	0.07	1.57

We also look at the lower range of outliers and find only observation 51 (county 115) which has crime rate < 0.01. This outlier data pt (county 115) has some significant outlier characterestics. County 115 has the lowest crime rate in the data set, extremely low density, the highest polpc (police per capita), the highest avg sentence, the third highest prbconv, and the lowest pctmin80.

One possible explanation is that county 115 could be a army/marine base or high security government facility (such as FBI field office, NSA or CIA facility) with national security concerns which explains the highest polpc. This also might explain the very high avgsen (20+ yrs) for crimes convicted in this country tend to have a higher severity/penalty. If this were the case, it would stand to reason that there are very civilan inhabitants and also explain why the population density is very low and crime rate is the lowest with probability of arrest and conviction both >1.

We can argue that this county is certainly an outlier, but not one that should be removed from the data as it represents a plausible county's observation in the dataset.

```
lower <- data_crmrte[data_crmrte$crmrte < 0.01,]
density_table <- data.frame(lower$county, lower$crmrte, lower$density, lower$polpc, lower$avgsen, lower$prbarr, lower$prbcon
v )
kable(round(density_table,4), col.names = c("County", "Crime Rate ", " Density", "Police", "Avg Sentence", "Prob of Arrest",
"Prob of Conv"),
      caption = "Special Outliers")
```

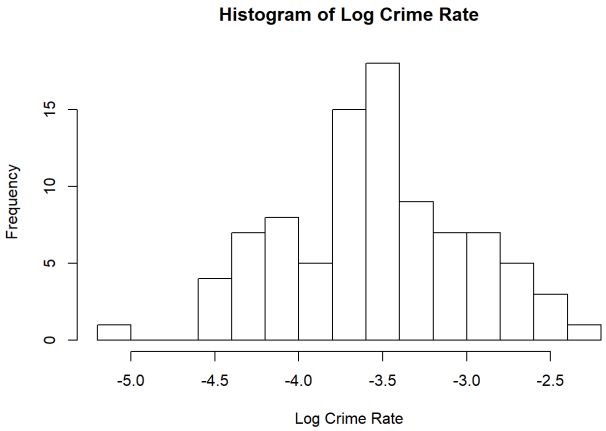
Special Outliers

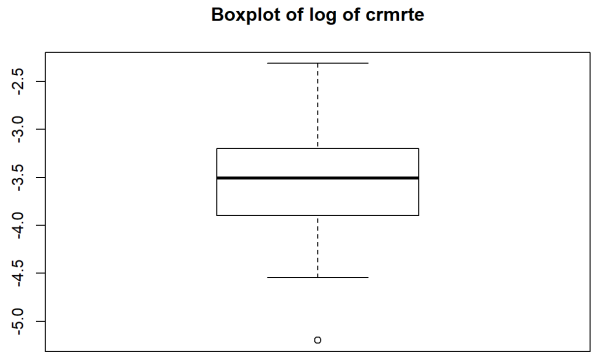
County	Crime Rate	Density	Police	Avg Sentence	Prob of Arrest	Prob of Conv
115	0.0055	0.3858	0.0091	20.7	1.0909	1.5

For campaign purposes, we want to predict crime. We want our candidate to be able to say that he or she can reduce crime in order to win votes. What is the most effective way to convey that? Using crime rate as it appears in the data set is using the level of crime rate and would suggest the following statement as a campaign slogan - "I can reduce crime to this rate by doing x, y, and z".

Transforming crime rate into the log of crime rate allows for the statement "I can reduce crime by n% by doing x, y, and z." We find the latter more powerful and meaningful to voters since most voters have no idea about the level of crime rates. In addition, we will show that the transformation of crime rate improves the normality and distribution of the variable, which will often reduce skew in the errors as well.

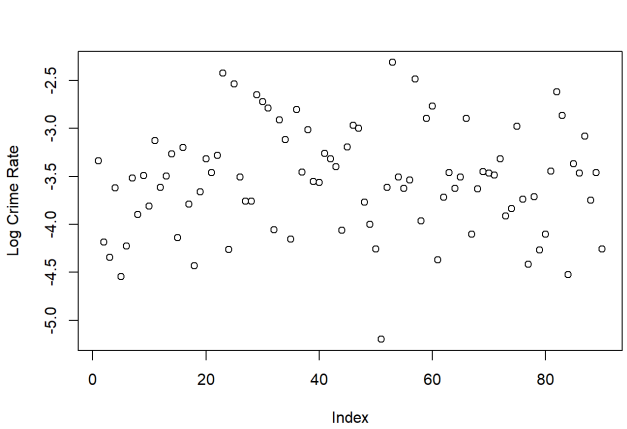
```
data_crmrte$log_crmrte <- log(data_crmrte$crmrte)
hist(data_crmrte$log_crmrte,breaks=20, xlab="Log Crime Rate", main="Histogram of Log Crime Rate")
boxplot(data_crmrte$log_crmrte, main="Boxplot of log of crmrte")
```

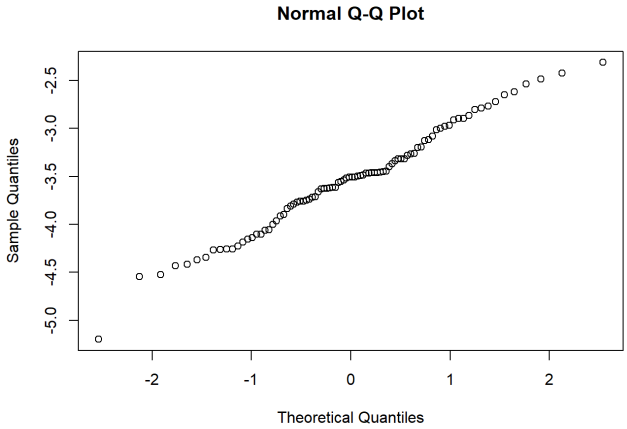





```
plot(data_crmrte$log_crmrte, ylab = "Log Crime Rate")
qqnorm(data_crmrte$log_crmrte)
shapiro.test(data_crmrte$log_crmrte) # Shapiro-wilk test confirms the log transformation was able to eradicate the non-norma
lity in the dependent variable
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_crmrte$log_crmrte
## W = 0.98857, p-value = 0.626
```





The histogram of the transformed crime rate is much more symmetrical and shows much less right skew. The box plot shows all of the outliers on the high end have been removed, though an outlier (county 115) on the low end has been become more prominent.

The scatter plot looks much more normal, and the Q-Q plot is much closer to normal with the data points hugging the 45 degree line much more closely. Given the stronger argument for the political campaign and the benefits to normality we have chosen to model the tranformation of crime rate as opposed to crime rate.

The Shapiro-wilk test fails to reject the Null Hypothesis of Normality, thereby confirming that the log transformation was able establish normal distribution of the data.

Panel Data and Further Data Vetting

We also were able to find the original panel data set. While we won't use this for prediction, we will use it for data vetting. We created three data sets. The first data set is comprised of the entire panel data which includes data from 1981-1987. The second data set is the data from the first 6 years and excludes the data set given in the assignment. Finally, the third data set is just the year 1987 and the columns in our assignment.

```
#Load the new dataset
data_panel <- read.table(file = 'C:/Users/winbase/MIDS/w203/w203_lab3/crime4.txt', sep="", header=FALSE)
#name the columns of the new dataset -->
colnames(data_panel) <- c("county", "year", "crmrte", "prbarr", "prbconv", "prbpris", "avgsen", "polpc","density", "taxpc",
"west", "central", "urban", "pctmin80", "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "wsta","wloc", "mix", "pctym
le", "d82", "d83", "d84", "d85", "d86", "d87", "lcrmrte", "lprbarr", "lprbconv","lprbpris", "lavgsen","lpolpc", "ldensity",
"ltaxpc", "lwcon", "lwtuc", "lwtrd", "lwfir", "lwser", "lwmfg","lwfed", "lwsta", "lwloc", "lmix", "lpctymle", "lpctmin", "cl
crmrte", "clprbarr", "clprbcon", "clprbpri","clavgsen", "clpolpc", "cltaxpc", "clmix")
data_panel$prbconv <- as.numeric(as.character(data_panel$prbconv))
#get all years besides 1987 for comparison
data_panel_additional <- data_panel[data_panel$year != 87,]
#get just the year 1987 for comparison
data_panel_87 <- data_panel[data_panel$year == 87,c("county", "year", "crmrte", "prbarr", "prbconv", "prbpris","avgsen", "po
lpc","density", "taxpc","west", "central", "urban", "pctmin80", "wcon", "wtuc", "wtrd", "wfir","wser", "wmfg", "wfed","wsta"
,"wloc", "mix", "pctymle")]


#compare the assignment data set to the new data set in 1987
data_check <- data_crmrte[, !(colnames(data_crmrte)=="log_crmrte")]
for (col in 1:ncol(data_check)) {
   for (row in 1:nrow(data_check)) {
       val_orig <- data_check[row, col]
       val_new  <- data_panel_87[row, col]
       if(abs(val_orig-val_new)>.001){
         print(paste("county=",data_check[row, 1], "year=",data_panel_87[row, 2],
                     "column=",colnames(data_check)[col],"original=", data_check[row, col],
                     "panel=",data_panel_87[row, col]))

         }
       }
 }
```

```
## [1] "county= 173 year= 87 column= density original= 2.03422e-05 panel= 0.2034221"
## [1] "county= 71 year= 87 column= west original= 1 panel= 0"
```

The code above compares the data set given to us to the same data extracted from the panel data. These data sets should match exactly. The print statements above identify two differences. For county 173 the panel data has a rounded value for density of .20342 while the given data set has a rounded value of .00002. Looking at the distribution for all years besides 1987, the minimum density for any county is .1977186 so the value in the given data set seems to be an extreme outlier, and it is much more likely that the value from the panel data set is correct. This value will be corrected below, and the quantiles are shown for evidence. In addition, county 71 is labeled as west in the given data set but for all years in the panel data set it has a zero value for west. This will also be corrected.

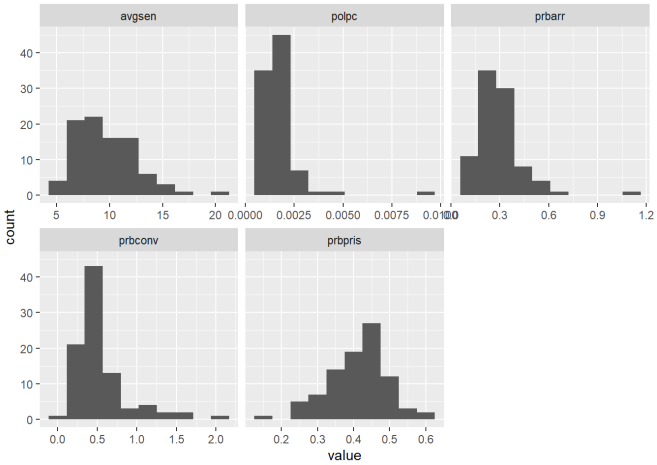
```
#Correct mistakes in original (assigned) data set
data_crmrte[data_crmrte$county==173, "density"] <- .2034221
data_crmrte[data_crmrte$county==71, "west"] <- 0
```

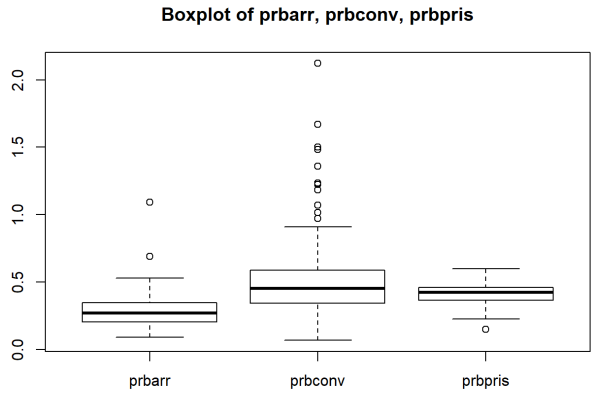
Groupings

Having no background on this paper or in criminal justice in general we looked for ways to make this data more digestable. We decided to group the variables into categories to make it more manageable, and in this process we found five groups that seemed natural: deterrent, wages, demographic, region, and urban. We performed exploratory data analysis on all of these variables.

The first group is deterrent data. As cited in the original paper, these variables were hypothesized to reduce crime rate through disincentivizing crime. Essentially, as the probability of getting caught increases, criminals' desire to commit crimes decreases.

Deterrent Data

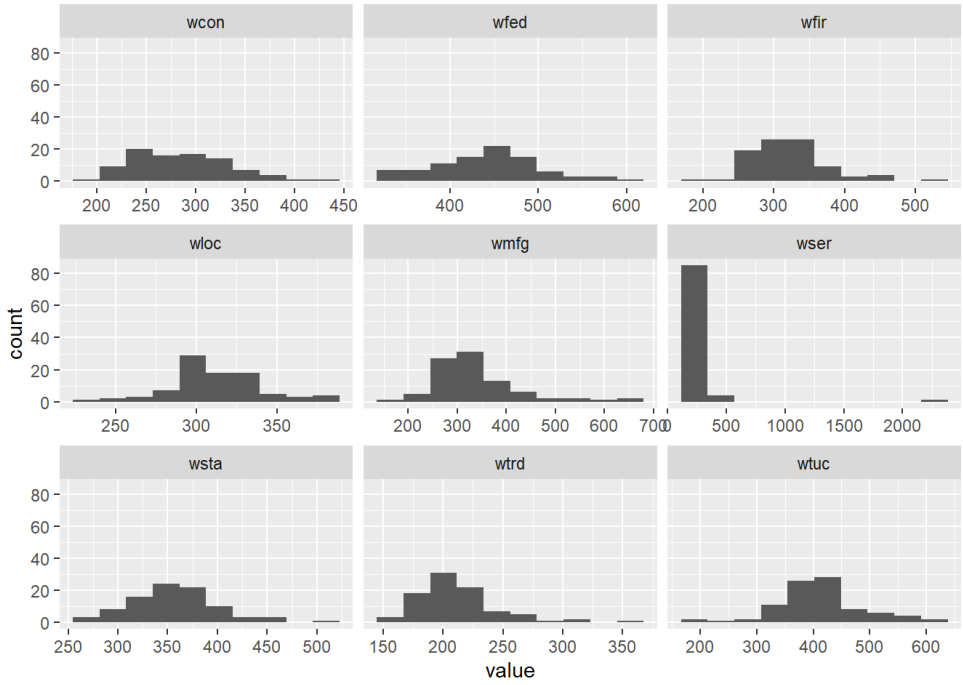




The first four histograms show right skew while prbpris shows left skew. Given their distributions, these variables are candidates to be transformed. In addition, these variables are more easily changed by a politican, which lends itself to a percent change argument similar to crime rate.

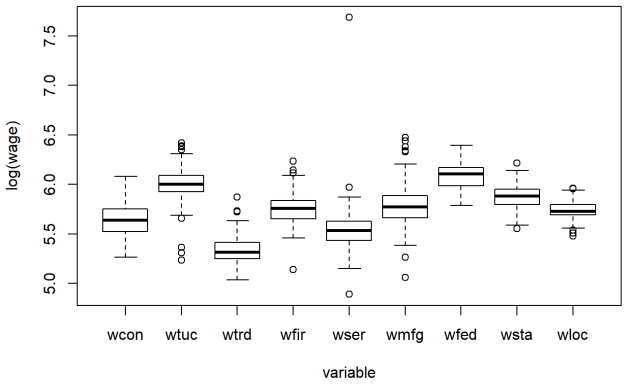
Wages Data

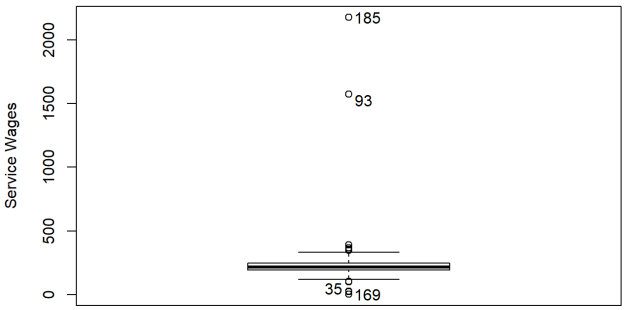
```
#create a dataframe of just the wage variables
wages_data <- data_crmrte[,c('wcon','wtuc','wtrd','wfir', 'wser',
                      'wmfg','wfed', 'wsta', 'wloc')]

#plot histograms of just the wage variables
ggplot(gather(wages_data), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')
```



```
#generate boxplots of just the wage variables
boxplot(log(wages_data), ylab="log(wage)", xlab="variable")
#Show outliers for entire 7 years
#boxplot(data_panel$wser, ylab="Service Wages", id=data_panel$county)
Boxplot(data_panel$wser, ylab="Service Wages", id=list(labels=data_panel$county, n=2, location="avoid"))
```

```
## [1] 169  35 185  93
```



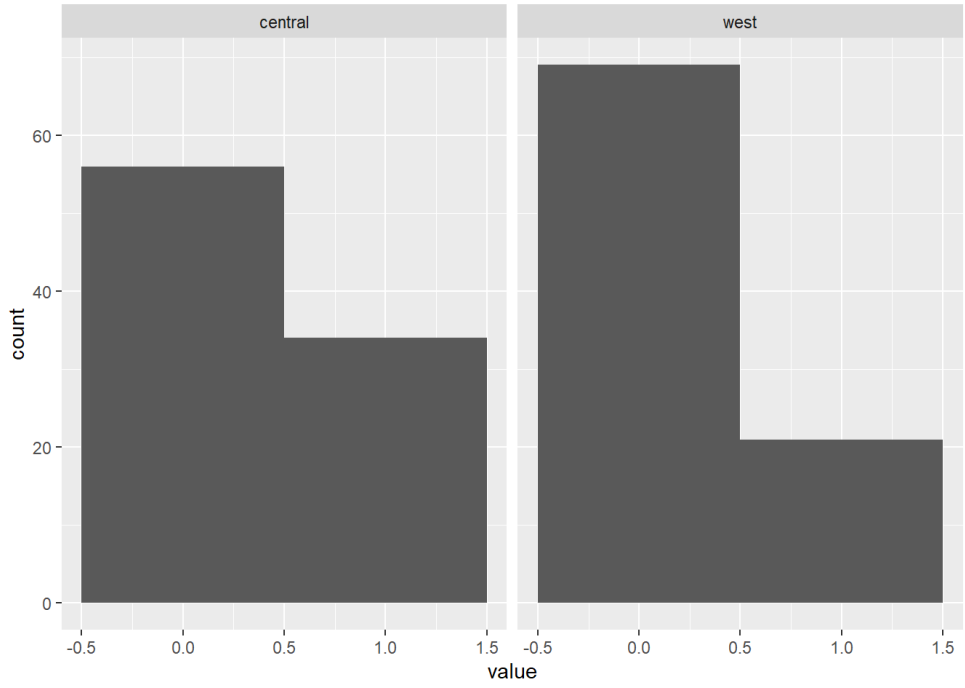


There is an obvious outlier for wser in County 185, (data pt 84 . The mean services wage across all the counties is $275 (with a std dev of 206) and 84 has wser of 2177 (~9sd from mean), which seems like a measurement or typographical error. The next highest average weekly wage in any sector is 646 versus the value of 2177. In addition, the above boxplot shows it is an outlier for any wser for all counties over the entire panel data set; indeed it is an outlier for all sectors. Therefore, this observation will be corrected. In looking at this county for all years, the average wage appears to be increasing. While it is somewhat of a guess, it appears an extra 7 has been inserted into the value, so it is removed below.

```
data_crmrte[data_crmrte$county==185, "wser"] <- 217.068
```

Region Data

```
#create a dataframe of just the wage variables
dummies_data <- data_crmrte[,c('west','central')]

#plot histograms of just the dummy variables
ggplot(gather(dummies_data), aes(value)) +
  geom_histogram(bins = 2) +
  facet_wrap(~key)
```



```
#just a quick check that there is no overlap
region_check <- data_crmrte[which(data_crmrte$west == 1 && data_crmrte$central == 1)]

summary(region_check)
```

```
## < table of extent 0 x 0 >
```

The regions are broken up into central, west, and east. East is left out of the data set and it's effect as the final level of the indicator variable will move to the intercept.

Urban Data

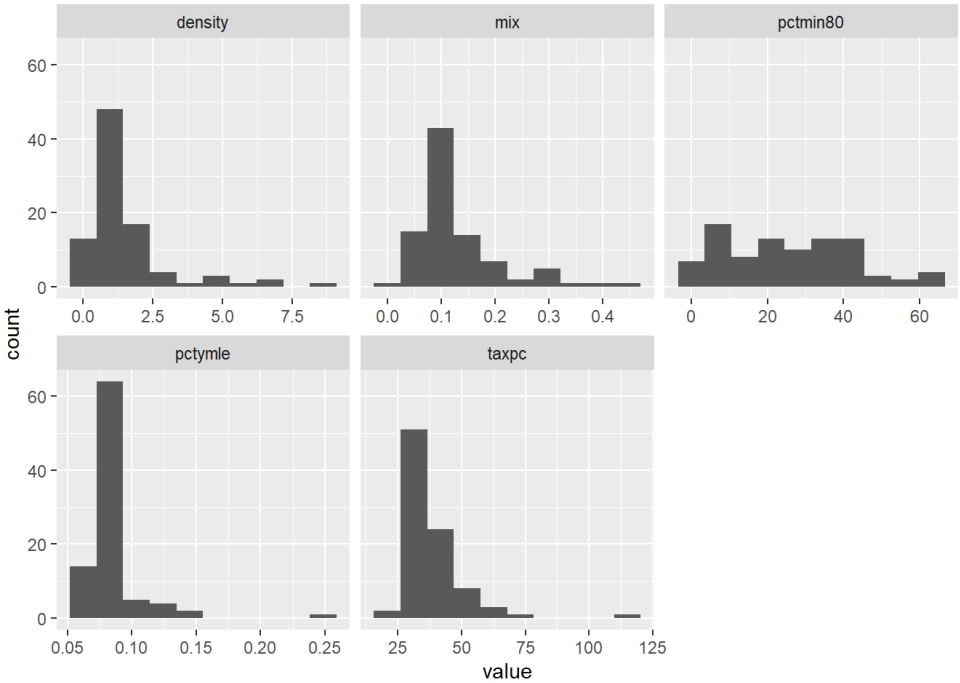
```
#plot histograms of just the wage variables
sum(data_crmrte$urban) # There are only 8 Urban areas out of 90 counties
```

```
## [1] 8
```

Urban did not fit into a great grouping so we left this variable on its own. A histogram shows that the state has relatively few urban counties, something to keep in mind when analyzing other variables such as density.

Demographic Data

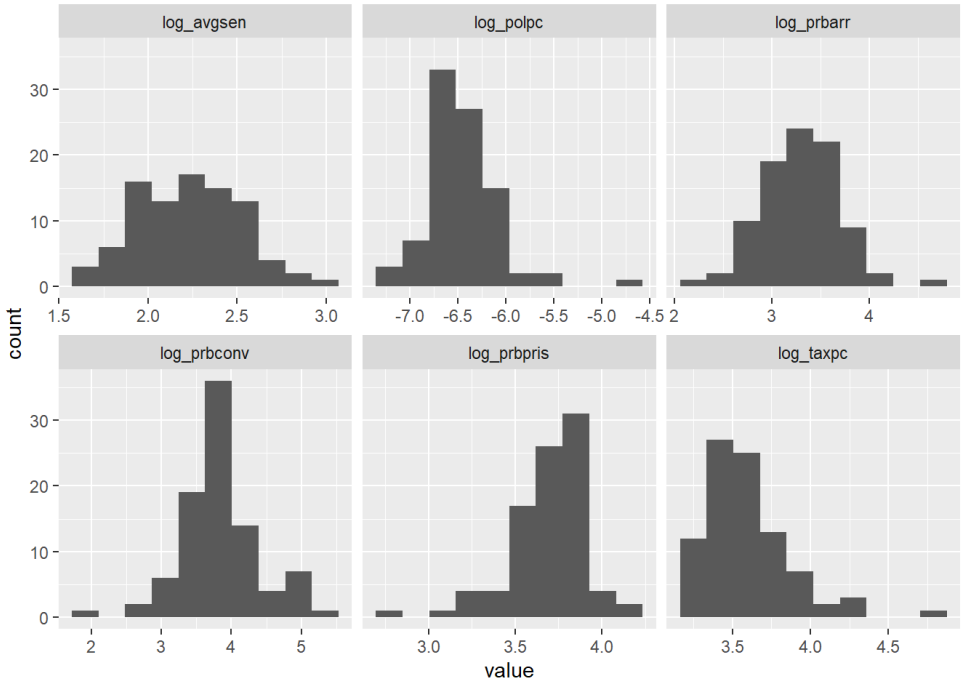
```
#create a dataframe of just the demographic variables
demographic_data <- data_crmrte[,c('density', 'taxpc', 'pctmin80',
                                   'mix', 'pctymle')]

#plot histograms of just the demographic variables
ggplot(gather(demographic_data), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')
```



Once again we see a lot of right skewed distributions in the histograms and in the box plots.

After exploring all of the variables we decided to tranform the other variables that are potentially under a politican's control - the deterrent variables. This gives us our final data set and so we can start running regressions.

```
data_crmrte$prbconv <- as.numeric(as.character(data_crmrte$prbconv))
data_crmrte$log_prbarr <- log(data_crmrte$prbarr*100)
data_crmrte$log_prbconv <- log(data_crmrte$prbconv*100)
data_crmrte$log_prbpris <- log(data_crmrte$prbpris*100)
data_crmrte$log_avgsen <- log(data_crmrte$avgsen)
data_crmrte$log_polpc <- log(data_crmrte$polpc)
data_crmrte$log_taxpc <- log(data_crmrte$taxpc)


#plot histograms of just the demographic variables
ggplot(gather(data_crmrte[,c('log_prbarr', 'log_prbconv', 'log_prbpris', 'log_avgsen', 'log_polpc', 'log_taxpc') ]), aes(val
ue)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')
```



Though the distribution of the variables still exhibits skew, the skew does seem to be reduced.

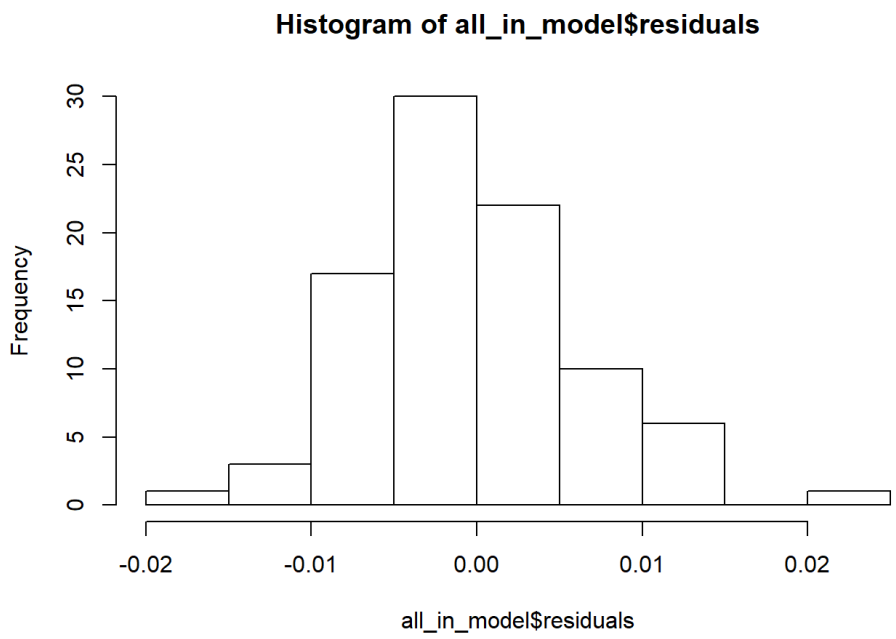
Log Tranformed Dependent Variable Comparison

In order to settle on the final data set we compare an all-in log-log model with an all-in log-linear to see which dependent variables are more suitable. This will explicitly test which dependent variables we should use - the log tranformed variables or the original variables.

```
###### Initial Models #####
all_in_model <- lm(crmrte ~ prbarr + prbconv + prbpris
                   + avgsen + polpc + density
                   + taxpc + west + central + urban + pctmin80 + wcon
                   + wtuc + wtrd + wfir + wser + wmfg
                   + wfed + wsta + wloc
                   + mix + pctymle,
                   data = data_crmrte)
se.all_in_model = sqrt(diag(vcovHC(all_in_model)))

coeftest(all_in_model, vcov = vcovHC) # HC White SE
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.3747e-02  2.8107e-02  0.4891 0.6263748
## prbarr      -5.2708e-02  1.4448e-02 -3.6480 0.0005180 ***
## prbconv     -1.9298e-02  5.0930e-03 -3.7892 0.0003261 ***
## prbpris      1.8359e-03  1.2370e-02  0.1484 0.8824640
## avgsen      -5.5646e-04  4.5512e-04 -1.2227 0.2257379
## polpc        6.8799e+00  2.5762e+00  2.6705 0.0094951 **
## density      5.1841e-03  1.3431e-03  3.8598 0.0002578 ***
## taxpc        1.9602e-04  2.4532e-04  0.7990 0.4270855
## west        -2.0933e-03  4.7536e-03 -0.4404 0.6610908
## central     -4.5658e-03  3.4671e-03 -1.3169 0.1923701
## urban        1.4762e-03  7.5852e-03  0.1946 0.8462861
## pctmin80     2.9217e-04  1.2101e-04  2.4144 0.0184990 *
## wcon         3.2167e-05  2.8650e-05  1.1228 0.2655367
## wtuc         9.4840e-06  1.8550e-05  0.5113 0.6108508
## wtrd         2.9265e-05  8.0695e-05  0.3627 0.7179985
## wfir        -1.9097e-05  2.8124e-05 -0.6790 0.4994667
## wser        -1.0099e-04  4.0356e-05 -2.5024 0.0147816 *
## wmfg        -4.4324e-06  1.3100e-05 -0.3384 0.7361559
## wfed         4.9763e-05  2.7254e-05  1.8259 0.0723241 .
## wsta        -3.2297e-05  3.4136e-05 -0.9461 0.3474910
## wloc         4.5734e-05  7.2717e-05  0.6289 0.5315334
## mix         -2.2032e-02  2.0444e-02 -1.0777 0.2850492
## pctymle      1.2561e-01  4.8437e-02  2.5933 0.0116616 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

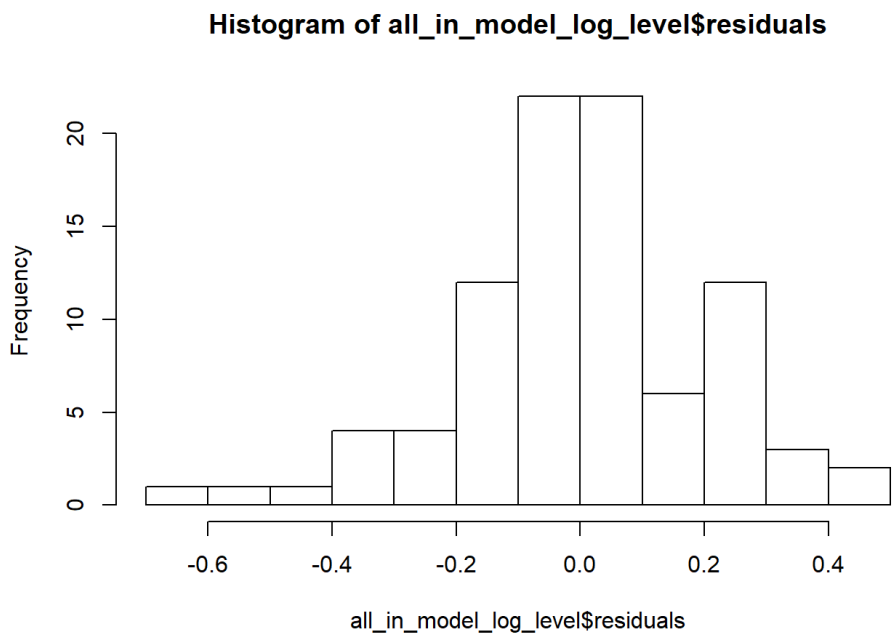
```
# Model residuals look normally distributed
hist(all_in_model$residuals)
```




```
# plot(all_in_model)
#The residuals vs Fitted plot for this model quickly tell us
#that ZCM ( MLR 4) and Random Sampling (MLR2) is violated for this model.

all_in_model_log_level <- lm(log_crmrte ~ prbarr + prbconv + prbpris
                             + avgsen + polpc + density
                             + taxpc + west + central + urban
                             + pctmin80 + wcon
                             + wtuc + wtrd + wfir + wser + wmfg
                             + wfed + wsta + wloc
                             + mix + pctymle,
                             data = data_crmrte)
se.all_in_model_log_level = sqrt(diag(vcovHC(all_in_model_log_level)))
coeftest(all_in_model_log_level, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4.0001e+00  8.0852e-01 -4.9474 5.337e-06 ***
## prbarr      -1.9106e+00  3.8228e-01 -4.9980 4.412e-06 ***
## prbconv     -6.9639e-01  1.3595e-01 -5.1223 2.754e-06 ***
## prbpris     -8.0121e-02  3.7251e-01 -0.2151  0.830353
## avgsen      -9.6865e-03  1.4217e-02 -0.6813  0.498010
## polpc        1.5590e+02  8.4796e+01  1.8386  0.070411 .
## density      1.0615e-01  5.7028e-02  1.8613  0.067095 .
## taxpc        3.4642e-03  6.8480e-03  0.5059  0.614611
## west        -1.3778e-01  1.3894e-01 -0.9917  0.324920
## central     -1.2697e-01  8.9844e-02 -1.4132  0.162213
## urban       -1.1830e-01  2.2651e-01 -0.5223  0.603204
## pctmin80     8.7072e-03  3.2098e-03  2.7127  0.008474 **
## wcon         6.9495e-04  8.3219e-04  0.8351  0.406642
## wtuc         1.1665e-04  6.2233e-04  0.1874  0.851880
## wtrd         2.4260e-04  1.7677e-03  0.1372  0.891255
## wfir        -7.3904e-04  1.0822e-03 -0.6829  0.497009
## wser        -1.6589e-03  1.2455e-03 -1.3319  0.187414
## wmfg        -1.0096e-04  4.4063e-04 -0.2291  0.819466
## wfed         2.6813e-03  1.0321e-03  2.5978  0.011524 *
## wsta        -1.3460e-03  8.7400e-04 -1.5400  0.128265
## wloc         1.0065e-03  2.1753e-03  0.4627  0.645083
## mix         -2.7416e-01  6.0320e-01 -0.4545  0.650934
## pctymle      3.0837e+00  1.1764e+00  2.6212  0.010833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

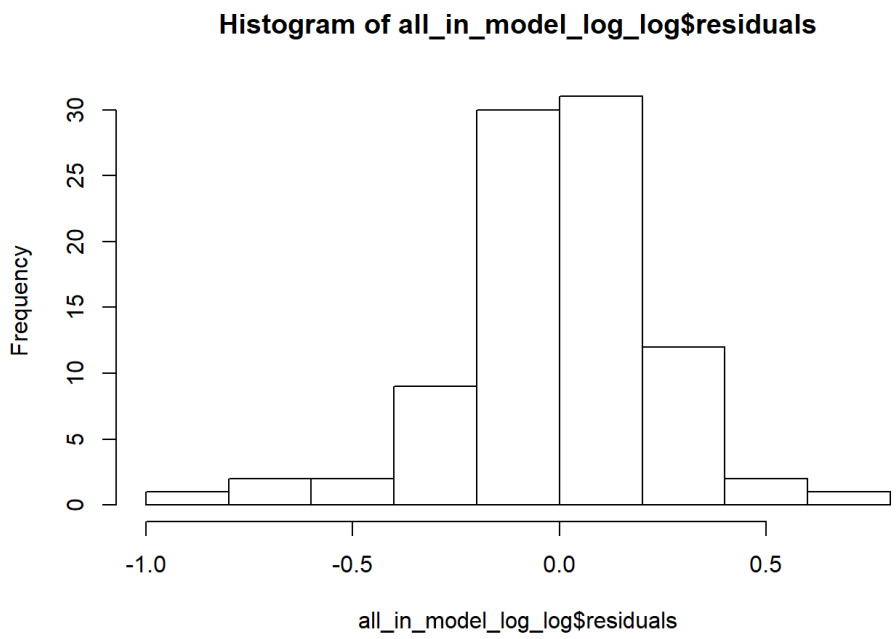
```
# Model residuals look somwewhat normally distributed
hist(all_in_model_log_level$residuals)
```




```
# plot(all_in_model_log_level)
#The residuals vs Fitted plot for this model tell us that
#the log transformation has significantly reduced the skew in the residuals
#and the clustering and exogenity violations have also been ameliorated.



all_in_model_log_log <- lm(log_crmrte ~ log_prbarr + log_prbconv
                           + log_prbpris + log_avgsen + log_polpc
                           + density+ log_taxpc + west + central
                           + urban + pctmin80 + wcon
                           + wtuc + wtrd + wfir
                           + wser + wmfg + wfed + wsta + wloc
                           + mix + pctymle,
                           data = data_crmrte)
se.all_in_model_log_log = sqrt(diag(vcovHC(all_in_model_log_log)))
coeftest(all_in_model_log_log, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  7.3143e-01  2.5438e+00  0.2875  0.77459
## log_prbarr  -5.0875e-01  1.7159e-01 -2.9650  0.00419 **
## log_prbconv -3.6293e-01  1.5625e-01 -2.3228  0.02324 *
## log_prbpris -4.2712e-02  1.8539e-01 -0.2304  0.81849
## log_avgsen  -1.8695e-01  1.7531e-01 -1.0664  0.29006
## log_polpc    2.7436e-01  2.7872e-01  0.9844  0.32848
## density      1.0516e-01  6.5676e-02  1.6012  0.11404
## log_taxpc    3.2344e-02  3.2232e-01  0.1003  0.92037
## west        -2.9345e-01  2.0265e-01 -1.4481  0.15226
## central     -1.9517e-01  1.2634e-01 -1.5447  0.12713
## urban       -8.2439e-02  2.6893e-01 -0.3065  0.76014
## pctmin80     6.0445e-03  4.9502e-03  1.2211  0.22634
## wcon         1.1773e-03  1.0019e-03  1.1750  0.24413
## wtuc         6.8736e-05  8.1591e-04  0.0842  0.93311
## wtrd         2.9887e-04  2.0166e-03  0.1482  0.88262
## wfir        -6.6657e-04  1.1777e-03 -0.5660  0.57329
## wser        -1.4133e-03  1.4954e-03 -0.9451  0.34798
## wmfg         6.6926e-05  5.4721e-04  0.1223  0.90303
## wfed         2.6693e-03  1.2926e-03  2.0651  0.04278 *
## wsta        -1.1509e-03  1.0530e-03 -1.0930  0.27830
## wloc         2.3993e-04  2.5074e-03  0.0957  0.92405
## mix         -2.6989e-01  8.0894e-01 -0.3336  0.73969
## pctymle      2.2576e+00  2.2897e+00  0.9860  0.32770
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model residuals look somwewhat normally distributed with some left skew
hist(all_in_model_log_log$residuals)
```




```
# plot(all_in_model_log_log)
#The residuals vs Fitted plot for this model tells us that
#the log transformation has significantly reduced the skew in the residuals.
#The clustering and exogenity violations have also been ameliorated
#but there are more outliers created from the log transform
#of the dependent variables increasing model error and reducing adj R-squared.



stargazer(all_in_model, all_in_model_log_level,
          all_in_model_log_log,
          type = "text", omit.stat = "f",
          se = list(se.all_in_model, se.all_in_model_log_level,
                    se.all_in_model_log_log),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ============================================================
##                                   Dependent variable:
##                               ------------------------------
##                                 crmrte        log_crmrte
##                                   (1)        (2)        (3)
## ------------------------------------------------------------
## prbarr                          -0.053***  -1.911***
##                                  (0.014)    (0.382)
##
## prbconv                         -0.019***  -0.696***
##                                  (0.005)    (0.136)
##
## prbpris                           0.002      -0.080
##                                  (0.012)    (0.373)
##
## avgsen                           -0.001     -0.010
##                                 (0.0005)   (0.014)
##
## polpc                            6.880**    155.903
##                                  (2.576)    (84.796)
##
## log_prbarr                                            -0.509**
##                                                        (0.172)
##
## log_prbconv                                            -0.363*
##                                                        (0.156)
##
## log_prbpris                                            -0.043
##                                                        (0.185)
##
## log_avgsen                                             -0.187
##                                                        (0.175)
##
## log_polpc                                               0.274
##                                                        (0.279)
##
## density                          0.005***    0.106      0.105
##                                  (0.001)    (0.057)    (0.066)
##
## taxpc                             0.0002     0.003
##                                  (0.0002)   (0.007)
##
## log_taxpc                                               0.032
##                                                        (0.322)
##
## west                             -0.002     -0.138    -0.293
##                                  (0.005)    (0.139)    (0.203)
##
## central                          -0.005     -0.127    -0.195
##                                  (0.003)    (0.090)    (0.126)
##
## urban                             0.001      -0.118    -0.082
##                                  (0.008)    (0.227)    (0.269)
##
## pctmin80                         0.0003*    0.009**     0.006
##                                  (0.0001)   (0.003)    (0.005)
##
## wcon                             0.00003     0.001      0.001
##                                 (0.00003)   (0.001)    (0.001)
##
## wtuc                             0.00001     0.0001    0.0001
##                                 (0.00002)   (0.001)    (0.001)
##
## wtrd                             0.00003     0.0002    0.0003
##                                  (0.0001)   (0.002)    (0.002)
##
## wfir                             -0.00002    -0.001    -0.001
##                                 (0.00003)   (0.001)    (0.001)
##
## wser                             -0.0001*    -0.002    -0.001
##                                 (0.00004)   (0.001)    (0.001)
##
## wmfg                             -0.00000   -0.0001    0.0001
##                                 (0.00001)   (0.0004)   (0.001)
##
## wfed                             0.00005    0.003**    0.003*
##                                 (0.00003)   (0.001)    (0.001)
##
```

```
## wsta                            -0.00003    -0.001    -0.001
##                                (0.00003)   (0.001)    (0.001)
##
## wloc                            0.00005     0.001     0.0002
##                                 (0.0001)   (0.002)    (0.003)
##
## mix                              -0.022     -0.274    -0.270
##                                 (0.020)    (0.603)    (0.809)
##
## pctymle                         0.126**    3.084**     2.258
##                                 (0.048)    (1.176)    (2.290)
##
## Constant                         0.014    -4.000***    0.731
##                                 (0.028)    (0.809)    (2.544)
##
## ------------------------------------------------------------
## Observations                       90         90         90
## R2                              0.875      0.858      0.798
## Adjusted R2                     0.833      0.812      0.732
## Residual Std. Error (df = 67)   0.008      0.238      0.284
## ============================================================
## Note:                           *p<0.05; **p<0.01; ***p<0.001
```

```
#Looking at the residual plots and checking for OLS assumption violations,
#we feel the log-level model is the best specified model
#within our initial class of models. This model also strikes the best balance
#between explanatory power and understandability of the variable.

# We will use the log-level class of models for further refining the
#specification over the course of the model selection process in this study.
```

The three models above are:
1. The level-level model
2. The log-level model
3. The log-log model

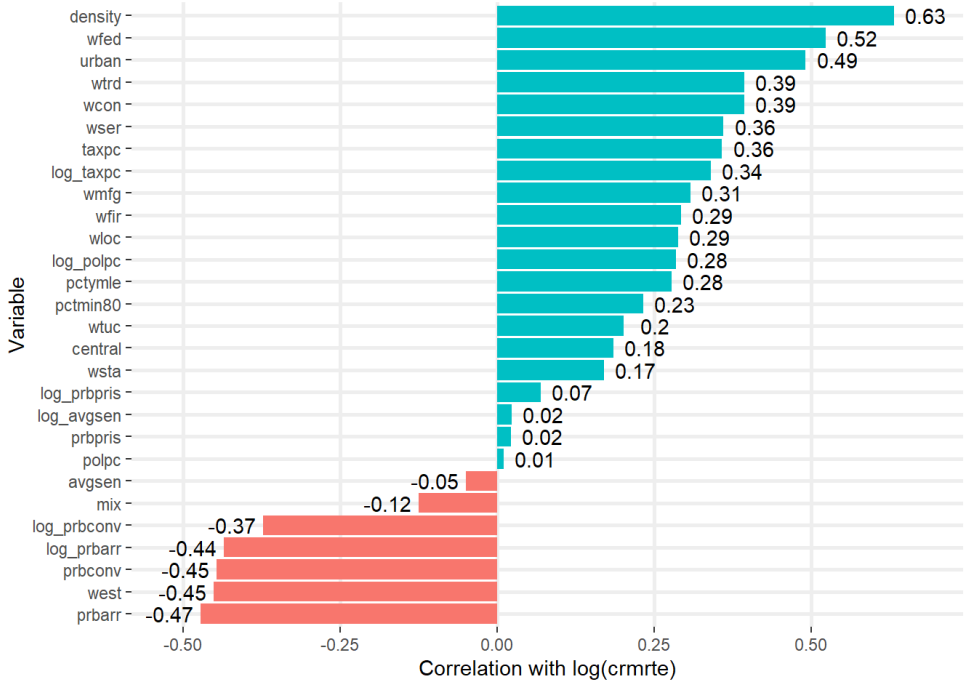
The log transformed dependent variable shows the best specification and most conformity to OLS assumptions within the 3 models. The log transformed deterrent variables (log-log model) explain less variation in log_crmrte as compared the log-level model.

The adjusted r-squared is lower (81.2% vs 73.2%) and the variables with statistical significance (prbarr/log_prbarr, prbconv/log_prbconv) both have higher t-stats using the untransformed variables. Therefore, we will use the log-level model as a base model moving forward.

Model 1: Simple Model

In order to create a simple model we decided to build using a bottom up approach. We started with a correlation matrix.

```
drops <- c("county","year", "crmrte", "log_crmrte")
cm = cor(data_crmrte[,!(names(data_crmrte) %in% drops)], data_crmrte$log_crmrte) #Corr Matrix as % for reading clarity

corrdf = data.frame(cm)
corrdf = corrdf[order(-corrdf$cm),,drop = FALSE]
corrdf$pos = (corrdf$cm) + (ifelse(corrdf$cm>0, 0.05, -0.05))
ggplot(data = corrdf,
       aes(y = cm, x=reorder(rownames(corrdf), cm), fill = cm > 0, width=0.9)) +
       geom_bar(stat = "identity", width=0.9, position = position_dodge(1.4)) +
       geom_text(aes(label=round(cm,2), y=pos), position=position_dodge(width=0.9)) +
       coord_flip() +
       theme(legend.position = "none", panel.grid.minor = element_blank(), panel.grid.major =element_line(size = 1, linetype
= 'solid',
                                colour = "#eeeeee"), panel.background = element_blank()) + ylab("Correlation with log(crmrt
e)") + xlab("Variable")
```



In the above correlation matrix, focusing on the correlations between the independent variable log_crmrte and all other variables in the matrix, density has the highest correlation. This variable makes intuitive sense. As a single variable it might encompass a lot of other factors. More opportunities exist for crime to occur in urban areas (especially when it is unknown what times of crimes are represented (ex: criminal/fraud/etc)). Below is the simple regression.

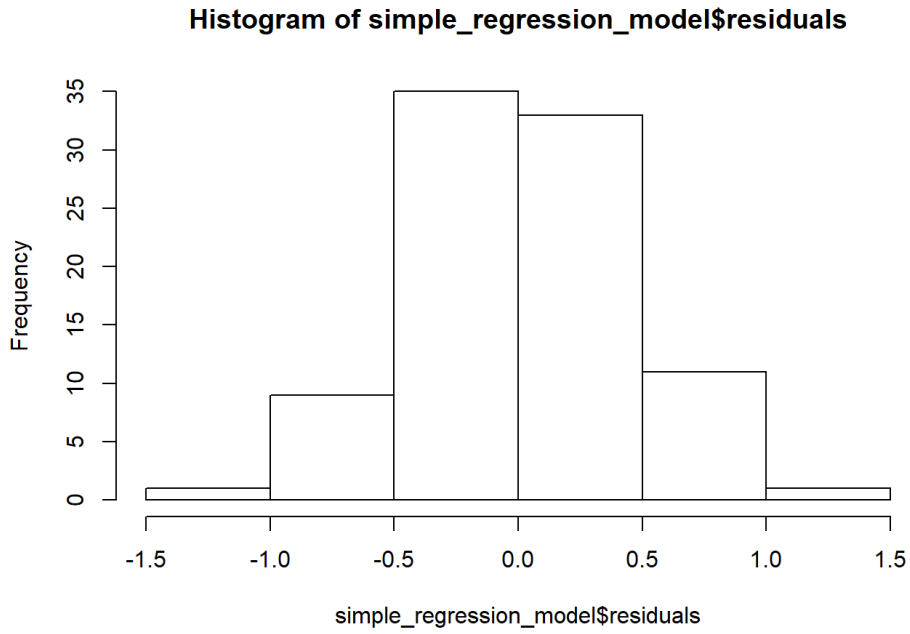
```
simple_regression_model <- lm(log_crmrte ~ density, data = data_crmrte)
se.simple_regression_model = sqrt(diag(vcovHC(simple_regression_model)))
coeftest(simple_regression_model, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.869836   0.068766 -56.2758 < 2.2e-16 ***
## density      0.228181   0.030463   7.4904 5.026e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 # AIC is higher than expected for a single variable model
AIC(simple_regression_model)
```

```
## [1] 106.5187
```

```
#The model residuals look normally distributed
hist(simple_regression_model$residuals)
```




```
stargazer(simple_regression_model,
          type = "text", omit.stat = "f",
          se = list(se.simple_regression_model),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                log_crmrte
## ------------------------------------------------
## density                        0.228***
##                                 (0.030)
##
## Constant                       -3.870***
##                                 (0.069)
##
## ------------------------------------------------
## Observations                      90
## R2                               0.399
## Adjusted R2                      0.392
## Residual Std. Error        0.428 (df = 88)
## ================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

Using r-squared, the variable density explains 39.9% of the variation in the log of crime rate . As density increases by 1 unit (as the county population divided by the county land area increases by 1 unit) crime increases by 0.22%.

Model 2: Kitchen Sink Model

Still, we can do better in predicting the log crime rate than simply using one variable. We now examine a “kitchen sink” model. This model includes all of the variables in the data set except county (which has too many values to be a useful indicator variable) and year, which is a constant (1987). Below are the results.

```
all_in_model_log_level <- lm(log_crmrte ~ prbarr + prbconv + prbpris
                             + avgsen + polpc + density
                             + taxpc + west + central + urban
                             + pctmin80 + wcon
                             + wtuc + wtrd + wfir + wser + wmfg
                             + wfed + wsta + wloc
                             + mix + pctymle,
                             data = data_crmrte)
se.all_in_model_log_level = sqrt(diag(vcovHC(all_in_model_log_level))) #HC White SE
coeftest(all_in_model_log_level, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4.0001e+00  8.0852e-01 -4.9474 5.337e-06 ***
## prbarr      -1.9106e+00  3.8228e-01 -4.9980 4.412e-06 ***
## prbconv     -6.9639e-01  1.3595e-01 -5.1223 2.754e-06 ***
## prbpris     -8.0121e-02  3.7251e-01 -0.2151  0.830353
## avgsen      -9.6865e-03  1.4217e-02 -0.6813  0.498010
## polpc        1.5590e+02  8.4796e+01  1.8386  0.070411 .
## density      1.0615e-01  5.7028e-02  1.8613  0.067095 .
## taxpc        3.4642e-03  6.8480e-03  0.5059  0.614611
## west        -1.3778e-01  1.3894e-01 -0.9917  0.324920
## central     -1.2697e-01  8.9844e-02 -1.4132  0.162213
## urban       -1.1830e-01  2.2651e-01 -0.5223  0.603204
## pctmin80     8.7072e-03  3.2098e-03  2.7127  0.008474 **
## wcon         6.9495e-04  8.3219e-04  0.8351  0.406642
## wtuc         1.1665e-04  6.2233e-04  0.1874  0.851880
## wtrd         2.4260e-04  1.7677e-03  0.1372  0.891255
## wfir        -7.3904e-04  1.0822e-03 -0.6829  0.497009
## wser        -1.6589e-03  1.2455e-03 -1.3319  0.187414
## wmfg        -1.0096e-04  4.4063e-04 -0.2291  0.819466
## wfed         2.6813e-03  1.0321e-03  2.5978  0.011524 *
## wsta        -1.3460e-03  8.7400e-04 -1.5400  0.128265
## wloc         1.0065e-03  2.1753e-03  0.4627  0.645083
## mix         -2.7416e-01  6.0320e-01 -0.4545  0.650934
## pctymle      3.0837e+00  1.1764e+00  2.6212  0.010833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Low value of AIC for this model specification which shows good parsimony adjusted fit
AIC(all_in_model_log_level)
```

```
## [1] 18.41539
```

```
stargazer(simple_regression_model, all_in_model_log_level,
          type = "text", omit.stat = "f",
          se = list(se.simple_regression_model, se.all_in_model_log_level),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ================================================
##                           Dependent variable:
##                     -------------------------------
##                                log_crmrte
##                           (1)             (2)
## ------------------------------------------------
## prbarr                                  -1.911***
##                                          (0.382)
##
## prbconv                                 -0.696***
##                                          (0.136)
##
## prbpris                                  -0.080
##                                          (0.373)
##
## avgsen                                   -0.010
##                                          (0.014)
##
## polpc                                    155.903
##                                         (84.796)
##
## density                 0.228***          0.106
##                          (0.030)         (0.057)
##
## taxpc                                     0.003
##                                          (0.007)
##
## west                                     -0.138
##                                          (0.139)
##
## central                                  -0.127
##                                          (0.090)
##
## urban                                    -0.118
##                                          (0.227)
##
## pctmin80                                 0.009**
##                                          (0.003)
##
## wcon                                      0.001
##                                          (0.001)
##
## wtuc                                     0.0001
##                                          (0.001)
##
## wtrd                                     0.0002
##                                          (0.002)
##
## wfir                                     -0.001
##                                          (0.001)
##
## wser                                     -0.002
##                                          (0.001)
##
## wmfg                                     -0.0001
##                                         (0.0004)
##
## wfed                                     0.003**
##                                          (0.001)
##
## wsta                                     -0.001
##                                          (0.001)
##
## wloc                                      0.001
##                                          (0.002)
##
## mix                                      -0.274
##                                          (0.603)
##
## pctymle                                  3.084**
##                                          (1.176)
##
## Constant                -3.870***       -4.000***
##                          (0.069)         (0.809)
##
## ------------------------------------------------
## Observations               90              90
## R2                        0.399           0.858
```

```
## Adjusted R2              0.392           0.812
## Residual Std. Error 0.428 (df = 88) 0.238 (df = 67)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

Unsuprisingly, the r-squared of the “kitchen sink” model is substantially higher (85.8% vs. 39.9%). More importantly, the adjusted r-squared which accounts for the number of variables in the models, is also higher (81.2% vs 39.2%). Interestingly, density is no longer the variable with the highest statistical significance. The coefficients show the effect after all the other variables have been controlled for (partialled out). In the “kitchen sink” model prbarr and prbconv both have the lowest p-values and highest statistical significance.

Model 3: Balanced Model

There is a middle ground between running a simple regression and using the “kitchen sink” model. We took two approaches to building this balanced model. We used a bottom up approach that relied on both the correlation matrix and stepwise regression. We also used a top down approach that started with the “kitchen sink” model and excluded variables. Both methods are discussed below.

For the bottom up approach, we started with the simple regression which used density as its only factor. Recall that decision was largely based on the correlation matrix. Next we used stepwise regression in examining each additional variable. The results are below.

```
base_forward = lm(log_crmrte ~ density,
                            data = data_crmrte)
forward_step = step(base_forward, scope = formula(all_in_model_log_level), direction = "forward")
```

With the top down approach, we started with model 3 and looked to exclude variables that weren’t as predictive. First, we ran hypothesis testing on all five groups, one group at a time.

```
#deterrent
linearHypothesis(all_in_model_log_level,
                 c("prbarr = 0", "prbconv = 0", "prbpris = 0",
                   "avgsen = 0", "polpc = 0"),
                 vcov = vcovHC)
```

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	72	*NA*	*NA*	*NA*
2	67	5	7.102519	2.211751e-05

2 rows

```
#wage
linearHypothesis(all_in_model_log_level,
                 c("wcon = 0", "wtuc = 0", "wtrd = 0",
                   "wfir = 0", "wser = 0", "wmfg = 0",
                   "wfed = 0", "wsta = 0", "wloc = 0"),
                 vcov = vcovHC)
```

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	76	*NA*	*NA*	*NA*
2	67	9	1.772341	0.09003793

2 rows

```
#region
linearHypothesis(all_in_model_log_level,
                 c("west = 0", "central = 0"),
                 vcov = vcovHC)
```

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	69	*NA*	*NA*	*NA*
2	67	2	1.002623	0.3723532

2 rows

```
#urban
linearHypothesis(all_in_model_log_level,
                 c("urban = 0"),
                 vcov = vcovHC)
```

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	68	*NA*	*NA*	*NA*
2	67	1	0.2727676	0.6032039

2 rows

```
#demographic
linearHypothesis(all_in_model_log_level,
                 c("density = 0", "taxpc = 0", "pctmin80 = 0",
                   "mix = 0", "pctymle = 0"),
                 vcov = vcovHC)
```

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	72	*NA*	*NA*	*NA*
2	67	5	3.706389	0.005069108

2 rows

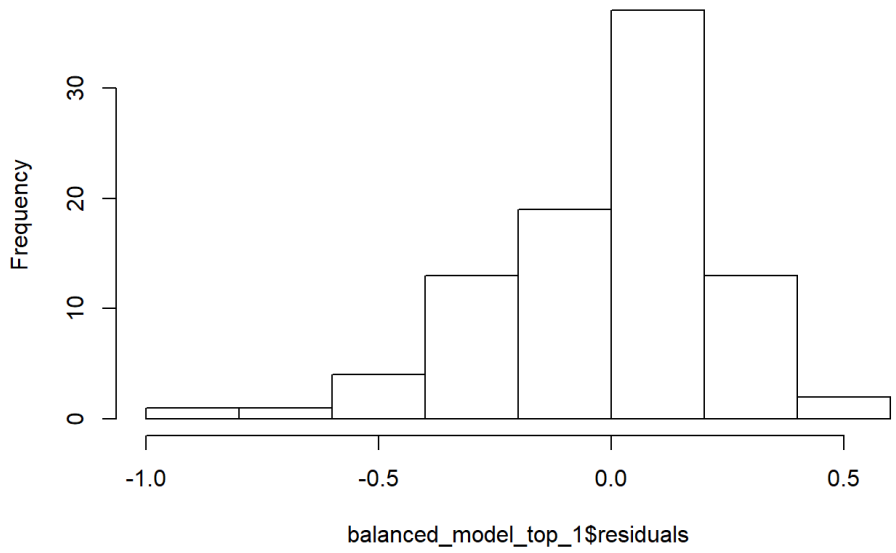
The hypothesis tests show that of the five groups the only groups that are jointly significant are the deterrent data and the demographic data. These tests measure whether removing all the variables within a group reduces the r-squared by a statistically signficant amount. If it does, then at least one variable, and perhaps all variables, within the group should be retained in the model.

We will re-run the models with just the deterrent and demographic data added and compare.

```
balanced_model_top_1 <- lm(log_crmrte ~ prbarr + prbconv + prbpris
                           + avgsen + polpc + density
                           + taxpc + pctmin80 + mix + pctymle,
                           data = data_crmrte)
se.balanced_model_top_1 = sqrt(diag(vcovHC(balanced_model_top_1)))
coeftest(balanced_model_top_1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -3.3584719   0.3588734 -9.3584 1.945e-14 ***
## prbarr       -1.9626880   0.3991189 -4.9176 4.673e-06 ***
## prbconv      -0.7678814   0.1366773 -5.6182 2.786e-07 ***
## prbpris      -0.0622711   0.4833049 -0.1288  0.897808
## avgsen       -0.0042601   0.0141588 -0.3009  0.764297
## polpc       175.4547516  82.2926450  2.1321  0.036107 *
## density       0.1130977   0.0350585  3.2260  0.001827 **
## taxpc         0.0021292   0.0055612  0.3829  0.702849
## pctmin80      0.0125104   0.0016232  7.7075 3.231e-11 ***
## mix          -0.7438992   0.5451903 -1.3645  0.176292
## pctymle       1.3902094   1.6264772  0.8547  0.395282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Residuals seem like they might not be normally distributed
hist(balanced_model_top_1$residuals)
```




```
# Shapiro test confirm it by rejecting normality
shapiro.test(balanced_model_top_1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  balanced_model_top_1$residuals
## W = 0.94861, p-value = 0.001374
```

```
#AIC for the model is pretty low
AIC(balanced_model_top_1)
```

```
## [1] 27.70917
```

```
#The residuals vs Fitted plot for this model shows curvature
# hence likely violation of ZCM and Nr error
# Although the model has high adj R-squares,
#it's likely the model specification is incorrect.

stargazer(all_in_model_log_level, balanced_model_top_1,
          type = "text", omit.stat = "f",
          se = list(se.all_in_model_log_level, se.balanced_model_top_1),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

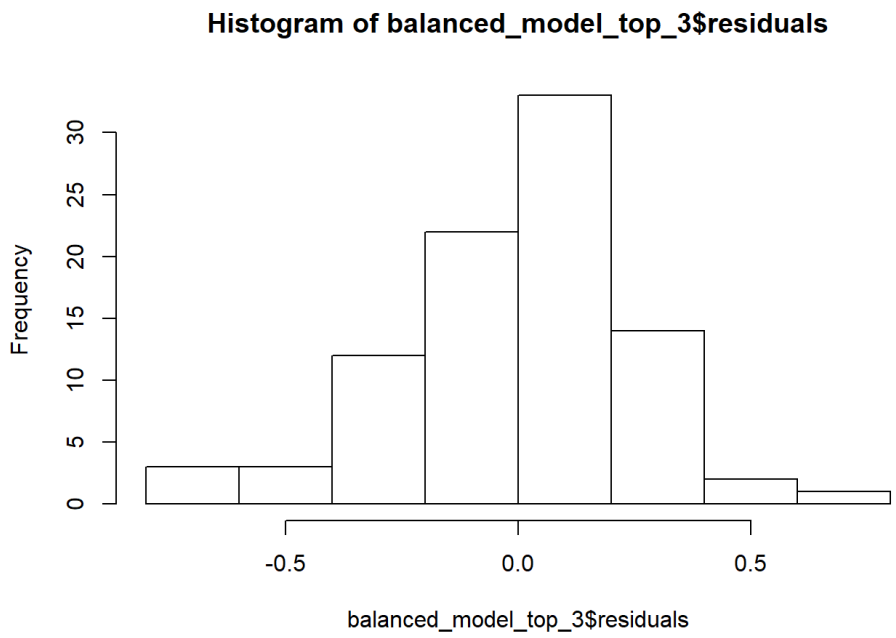
```
##
## ================================================
##                           Dependent variable:
##                      -------------------------------
##                                log_crmrte
##                           (1)             (2)
## ------------------------------------------------
## prbarr                  -1.911***       -1.963***
##                          (0.382)         (0.399)
##
## prbconv                 -0.696***       -0.768***
##                          (0.136)         (0.137)
##
## prbpris                  -0.080          -0.062
##                          (0.373)         (0.483)
##
## avgsen                   -0.010          -0.004
##                          (0.014)         (0.014)
##
## polpc                    155.903        175.455*
##                         (84.796)        (82.293)
##
## density                   0.106          0.113**
##                          (0.057)         (0.035)
##
## taxpc                     0.003           0.002
##                          (0.007)         (0.006)
##
## west                     -0.138
##                          (0.139)
##
## central                  -0.127
##                          (0.090)
##
## urban                    -0.118
##                          (0.227)
##
## pctmin80                 0.009**        0.013***
##                          (0.003)         (0.002)
##
## wcon                      0.001
##                          (0.001)
##
## wtuc                     0.0001
##                          (0.001)
##
## wtrd                     0.0002
##                          (0.002)
##
## wfir                     -0.001
##                          (0.001)
##
## wser                     -0.002
##                          (0.001)
##
## wmfg                     -0.0001
##                         (0.0004)
##
## wfed                     0.003**
##                          (0.001)
##
## wsta                     -0.001
##                          (0.001)
##
## wloc                      0.001
##                          (0.002)
##
## mix                      -0.274          -0.744
##                          (0.603)         (0.545)
##
## pctymle                  3.084**          1.390
##                          (1.176)         (1.626)
##
## Constant                -4.000***       -3.358***
##                          (0.809)         (0.359)
##
## ------------------------------------------------
## Observations               90              90
## R2                        0.858           0.795
```

```
## Adjusted R2              0.812           0.769
## Residual Std. Error 0.238 (df = 67) 0.264 (df = 79)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

Our adjusted r-squared has only fallen from 81.2% to 76.9% but we have dropped 12 variables. This is a much more parisimous model.

Three of the five groups have been eliminated, with only the deterrent and demographic groups remaining. We will use step wise regression to evaluate.

```
base_backward = lm(log_crmrte ~ prbarr + prbconv + prbpris
                           + avgsen + polpc + density
                           + taxpc + pctmin80 + mix + pctymle,
                           data = data_crmrte)


balanced_model_top_3 <- lm(log_crmrte ~ density
                         + polpc + pctmin80
                         + prbarr + prbconv,
                           data = data_crmrte)
se.balanced_model_top_3 = sqrt(diag(vcovHC(balanced_model_top_3)))
# Residuals look normally distributed
hist(balanced_model_top_3$residuals)
```




```
# Shapiro test on residuals confirms normality
shapiro.test(balanced_model_top_3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  balanced_model_top_3$residuals
## W = 0.97312, p-value = 0.05875
```

```
coeftest(balanced_model_top_3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  -3.2602986   0.1756202 -18.5645 < 2.2e-16 ***
## density       0.1157017   0.0283962   4.0746 0.0001041 ***
## polpc       193.3903420  53.9432366   3.5851 0.0005646 ***
## pctmin80      0.0120992   0.0015706   7.7036 2.369e-11 ***
## prbarr       -2.2755220   0.3325058  -6.8436 1.176e-09 ***
## prbconv      -0.7577049   0.1157797  -6.5444 4.461e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# This model also has low AIC proving good parsimony adjusted fit
AIC(balanced_model_top_3, k=2)
```

```
## [1] 23.3983
```

```
stargazer(all_in_model_log_level, balanced_model_top_3,
          type = "text", omit.stat = "f",
          se = list(se.all_in_model_log_level, se.balanced_model_top_3),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ================================================
##                           Dependent variable:
##                     -------------------------------
##                                log_crmrte
##                           (1)             (2)
## ------------------------------------------------
## prbarr                 -1.911***       -2.276***
##                         (0.382)         (0.333)
##
## prbconv                -0.696***       -0.758***
##                         (0.136)         (0.116)
##
## prbpris                 -0.080
##                         (0.373)
##
## avgsen                  -0.010
##                         (0.014)
##
## polpc                   155.903       193.390***
##                        (84.796)        (53.943)
##
## density                  0.106         0.116***
##                         (0.057)         (0.028)
##
## taxpc                    0.003
##                         (0.007)
##
## west                    -0.138
##                         (0.139)
##
## central                 -0.127
##                         (0.090)
##
## urban                   -0.118
##                         (0.227)
##
## pctmin80                0.009**         0.012***
##                         (0.003)         (0.002)
##
## wcon                     0.001
##                         (0.001)
##
## wtuc                    0.0001
##                         (0.001)
##
## wtrd                    0.0002
##                         (0.002)
##
## wfir                    -0.001
##                         (0.001)
##
## wser                    -0.002
##                         (0.001)
##
## wmfg                    -0.0001
##                        (0.0004)
##
## wfed                    0.003**
##                         (0.001)
##
## wsta                    -0.001
##                         (0.001)
##
## wloc                     0.001
##                         (0.002)
##
## mix                     -0.274
##                         (0.603)
##
## pctymle                 3.084**
##                         (1.176)
##
## Constant               -4.000***       -3.260***
##                         (0.809)         (0.176)
##
## ------------------------------------------------
## Observations              90              90
## R2                       0.858           0.782
```

```
## Adjusted R2              0.812           0.769
## Residual Std. Error 0.238 (df = 67) 0.264 (df = 84)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

The difference between the backward and forward model is that the backward model chooses variables for exclusion based on comparing significance while the forward model looks for significance in inclusion. We also used the f-tests (hypothesis tests) to give the backward stepwise regression a head start.

The backward stepwise regression yielded a more reasonable model so that is the model we are choosing for our balanced model. This model strikes a nice balance between parsimony and explanatory power. The variables included are prbarr, prbconv, polpc, density and pctmin80. Five out of the original twenty four independent variables are included. The adjusted r-squared is only 3% lower (76.9% vs. 81.2%). It includes a blend of actionable items for the campaign in the deterrent data as well as demographic variables that perhaps can focus the campaign's efforts.

Final Model

$$log(crmrte) = \beta_0 + \beta_1 \cdot prbarr + \beta_2 \cdot prbconv + \beta_3 \cdot polpc + \beta_4 \cdot density + \beta_5 \cdot pctmin80$$

$$\begin{aligned}\beta_0 &= -3.26\\ \beta_1 &= -2.276\\ \beta_2 &= -0.758\\ \beta_3 &= 193.390\\ \beta_4 &= 0.116\\ \beta_5 &= 0.012\end{aligned}$$

The above equation shows the final model and its coefficients.

- The deterrent variables (prbarr and prbconv) have negative impact on crime rate indicated by the negative coefficients. As the prbarr increases by one unit (.01), predicted crime rate decreases by 2.28%. As prbconv increases by one unit (.01), predicted crime decreases by 0.76%.
- The demographic variable density and pctmin80 have a positive impact on crime rate. As density increases by one unit (.01), predicted crime increases by 0.12%. As pctmin80 increases by one unit (.01), predicted crime rate increases by 0.01%.
- The demographic variable polpc has an unexpected positive sign and large magnitude. This does not mean that crime rate increase with increase in police per capita. This effect is caused because of the simultaneity bias between "polpc" and "crm_rate" which causes violation of the OLS assumptions (which is discussed further in the next few sections) and hence leads to the estimated causal effect of crime rate to suffer from omitted variable bias .Polpc is a control variable in our model and it's coefficient is in violation of the condition mean independence requirement and hence does not have a causal interpretation

Here is a view of the coefficients along with predictions at the 25th percentile, median, mean, and 75th percentile. This gives us both predictions and a sense of the distribution.

Independent Variables	Percentiles			
	25th	median	mean	75th
crmrte	0.020604	0.03	0.03351	0.0402

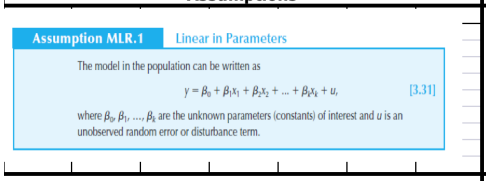
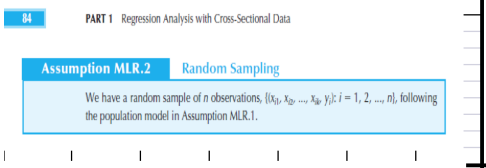
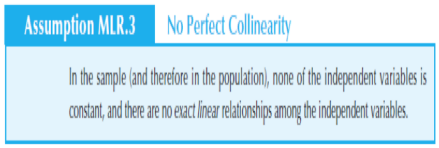
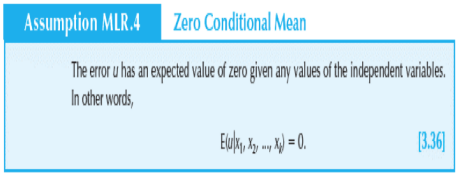
Coefficient	Dependent Variables	Percentiles				Predictions			
		25th	median	mean	75th	25th	median	mean	75th
(3.260)	Intercept					(3.2600)	(3.2600)	(3.2600)	(3.2600)
(2.276)	prbarr	0.2050	0.2710	0.2949	0.3449	(0.4665)	(0.6167)	(0.6712)	(0.7849)
(0.758)	prbconv	0.3442	0.4528	0.5513	0.5851	(0.2609)	(0.3432)	(0.4179)	(0.4435)
193.390	polpc	0.0012	0.0015	0.0017	0.0019	0.2394	0.2872	0.3291	0.3647
0.116	density	0.5472	0.9623	1.4288	1.5693	0.0635	0.1116	0.1657	0.1820
0.012	pctmin80	10.0240	24.3117	25.4955	38.1830	0.1203	0.2917	0.3059	0.4582
					Sum	(3.5642)	(3.5294)	(3.5483)	(3.4836)
					Prediction (exp)	0.0283	0.0293	0.0288	0.0307

Predictions

There are a couple of things to note in looking at this table. First, let's look at the interquartile ranges to assess practical significance. At first glance, pctmin80 doesn't appear to be practically significant as a one unit increase only leads to a .01% increase in predicted crime rate. However, a one unit increase is quite small for this variable. polpc has the tightest interquartile range. We view polpc, density, and pctmin80 more as control variables. They are important for making prediction but we don't view them as part of a political campaign. If we look at prbarr and prbconv, we can realistically hope to improve both variables by 10 units (.10). A ten unit increase in prbarr is similar to going from the 25th to 75th percentile of the existing data set. A ten unit increase in the prbconv is similar to going from the median to the mean. In the event we are able to improve both measures by 10%, we predict crime to fall by 26.17%.

3. An Assessment of the CLM Assumptions

We choose our balanced model for the complete assessment of all 6 classical linear model assumptions.

Assumptions	Ways Assumption Fails	Diagnostic	Conseq of Failed Assump	Solution
	Fit linear model to nonlinear data	plot of observed vs. predicted values or plot of residuals vs. predicted values	Bad predictions, particularly out of range of the sample data	Apply a nonlinear transformation to either independent or dependent variables; add another regressor that is a nonlinear function of another variable; add a possibly omitted variable
	1.) Clustering (researchers can only access a limited number 2.) Autocorrelation common for time series data)	1.) Use knowledge of where data comes from 2.) Durbin Watson Statistic	1.) Observing less variation than actually exists in the population; betas still unbiased but estimates are much less precise	1.) Use clustered standard errors 2.) No simple fix for serial correlation - use time series model
	Extremely high or perfect multicollinearity (assumption only rules out perfect multicollinearity); often from lagged variables of another variable, a shared common time trend, or variables that capture similar phenomena	Correlation matrix (though difficult for several variables); VIF's (multicollinearity likely between 5 and 10, problem > 10)	When variables are highly correlated but not perfectly collinear, OLS works but estimates will be much less precise. R-squared may be high but t-stats are low; regression becomes sensitive to small changes in specification and adding or removing a variable changes betas a lot; you might get nonsensical coefficient signs and magnitudes; confidence intervals might be very wide	Drop redundant variables
	The error exhibits a pattern that is not in a fairly constant band around zero or it shows a pattern that results in nonzero errors for different x's	For one variable, plot residuals vs predictor (should see flat average line around zero). For multiple regression, plot residuals vs. fitted values (predicted values). Should again see a flat band or line. Also use domain knowledge on any omitted variables.	Endogeneity is a violation of zero-conditional mean and results in OLS coefficients that are biased and inconsistent. If the explanatory variables are uncorrelated with the error term they are exogenous	1.) Change the functional form (log of independent or dependent variable, x and x^2, etc.); might lose interpretability though. 2.) Adding new variables. 3.) Decide we can't meet zero conditional mean but we can meet exogeneity. If we satisfy the first three assumptions and exogeneity (Cov(x,u) = 0 for all x, dependent variables) then OLS estimators are consistent (unbiased as n->infinity)

MLR 1-4'

MLR.1: The model is linear in parameters (and the error term)

we haven't constrained the error term, so the model can be any joint distribution. Therefore the linear model assumption is not violated

```
balanced_model_top_3 <- lm(log_crmrte ~  density
                           + polpc + pctmin80
                           + prbarr + prbconv,
                             data = data_crmrte)
```

MLR.2: Random sampling

First thing to note is that we are dealing with a single cross-section (1987) of a multi-year panel data.

Secondly this is observational data and not experimental so perfect random sampling is hard to achieve.

CORNWELL – TRUMBULL (1994) specifically state they choose panel data because cross – section data were not able to capture the real effect of the crime rate on several independent regressors.

The authors identify that the time-series component of the panel data is able to identify specific characteristics of county heterogeneity, which is correlated with the criminal justice variables.

In exploring the effects of county specific heterogeneity, counties next to each other may exhibit similar behaviour. While that may be valid for prediction model the standard errors may be understated causing violation of random sampling

While the balanced model achieves high level of statistical significance for the co-efficients, it's important to be mindful of the limitations of the dataset.

```
se.balanced_model_top_3 = sqrt(diag(vcovHC(balanced_model_top_3)))
coeftest(balanced_model_top_3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  -3.2602986   0.1756202 -18.5645 < 2.2e-16 ***
## density       0.1157017   0.0283962   4.0746 0.0001041 ***
## polpc       193.3903420  53.9432366   3.5851 0.0005646 ***
## pctmin80      0.0120992   0.0015706   7.7036 2.369e-11 ***
## prbarr       -2.2755220   0.3325058  -6.8436 1.176e-09 ***
## prbconv      -0.7577049   0.1157797  -6.5444 4.461e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MLR.3: No perfect multicollinearity

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. We have perfect multicollinearity if, for example as in the equation above, the correlation between two independent variables is equal to 100% or negative 100%.

We can check for multicollinearity by making a correlation matrix (though there are other complex ways of checking them like Variance Inflation Factor, which are outside the scope of this study). As seen from the correlation matrix below, there is no perfect multicollinearity in the model but we observe some meaningful correlations between (Prbarr, polpc) and (Prbarr,density). Indicating partial multicollinearity in the data. This correlation makes polpc useful as a control variate in the final regression model but the co-efficient does not have causal interpretaion due to omitted variable bias.

These linear realtionships among the X's don't invalidate the MLR parameters but they lower precision and increase the std-errors in the mdoel

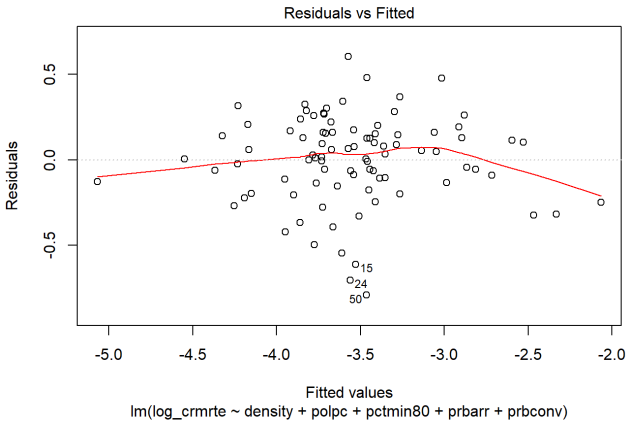
```
balanced_model <- c( "density","polpc", "pctmin80","prbarr","prbconv")
balanced_model_data <- data_crmrte[balanced_model]
round(cor(balanced_model_data)*100,0) # correlations displayed as % for convenience
```

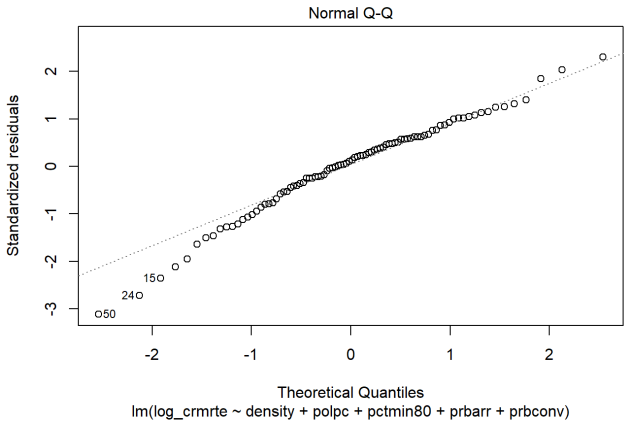
```
##          density polpc pctmin80 prbarr prbconv
## density      100    16       -7    -30     -23
## polpc         16   100      -17     43      17
## pctmin80      -7   -17      100      5       6
## prbarr       -30    43        5    100      -6
## prbconv      -23    17        6     -6     100
```

MLR.4: Zero Conditional Mean / exogeneity

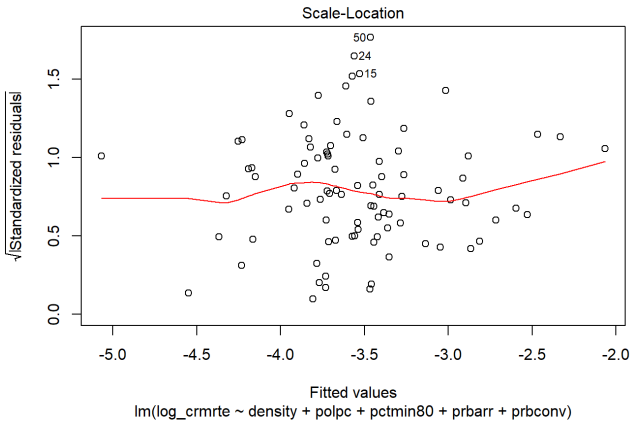
ZCM is best analysed by studying the regression plots of the residuals. Let's start by looking at the regression plots of the balanced model

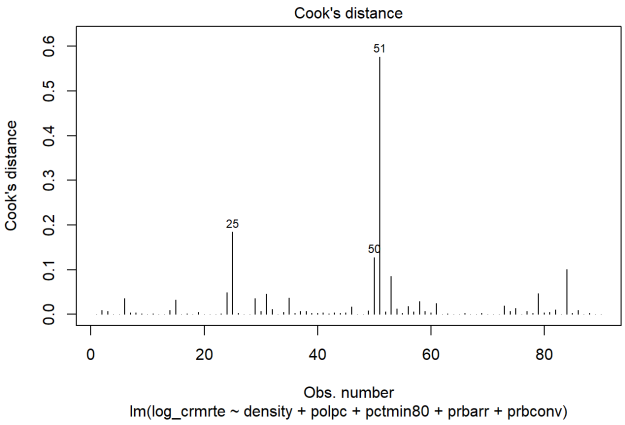
```
plot(balanced_model_top_3, 1)
plot(balanced_model_top_3, 2)
```






```
plot(balanced_model_top_3, 3)
plot(balanced_model_top_3, 4)
```





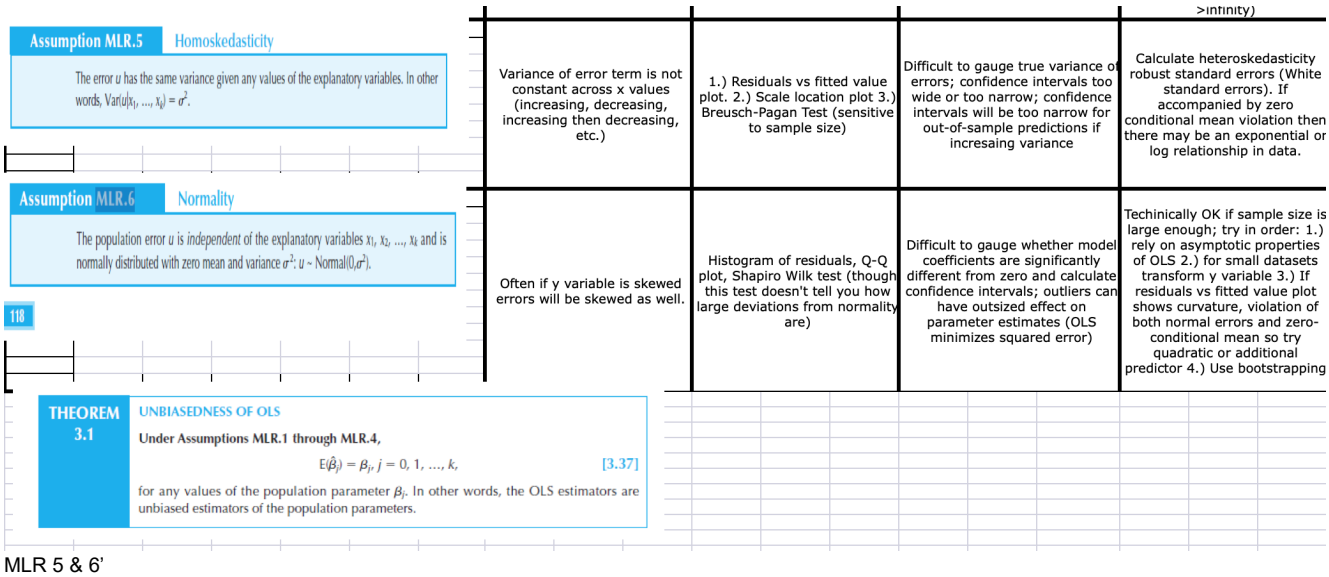
CLM assumptions analysis from plots

- Plot 1. The residuals vs. fitted plot indicates that the zero conditional mean assumption is NOT perfectly satisfied but the red line is close enough to zero, a big improvement compared to some of the other models we tested. The non-uniform thickness of the residuals-especially in the middle- indicates possible heteroskedasticity.
- Plot 2. The Q-Q plot shows that the residuals are not perfectly normally distributed, but the log transform of the crime rate improved the positive skew in the data but has introduced some negative skew
- Plot 3. The scale location plot indicated the presence of heteroskedasticity especially in the middle where the thickness of the band varies and outliers such as '50' and '24' are generating large standardized residuals
- Plot 4. The residuals vs leverage plot shows some of the outliers we had discussed earlier (observations 51, 25, 84) but the most significantly outlier is observation 51 or county 115 (having high leverage and Cook's distance >1). This outlier significantly affects our model estimate and likely increases model error.

```
round(cor(balanced_model_top_3$residuals, balanced_model_data)*100,5)
```

```
##      density polpc pctmin80 prbarr prbconv
## [1,]       0     0        0      0       0
```

Finally we check the correlation between the X's and the errors in the model to ensure there is no endogeneity in the model. The zero correlation is a necessary but not sufficient condition for the presence of omitted variable in the regression. There is a more extensive discussion of omitted variable and their implication on model endogeniety in section 5 of the report.



MLR 5 & 6'

MLR.5: Homoskedasticity

Homoskedasticity describes a situation in which the error term has the same variance across all values of the independent variables.

The regression plots indicate the presence of some heteroskedasticity in the errors let's test if they are statistically significant using the Breusch-Pagan test.

- The Breusch-Pagan test below allows us to test for heteroskedasticity under the

$H_0 : Homeskedasticity$

```
bptest(balanced_model_top_3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  balanced_model_top_3
## BP = 7.4305, df = 5, p-value = 0.1905
```

From the BP test, surprisingly we find p-value is not statistically significant, therefore we fail to reject $H_0 : Homeskedasticity.$

However, we will still choose to be more conservative and use HC consistent std-errors (Huber-white Std-errors) using coeftest function from the sandwich package in R. This conservative approach we have taken throughout this report in our model selection process in choosing regressors for different models

MLR.6: Normality of the error term

Often, if the Y variable is skewed, the error terms will be skewed as well.

We can check the normality using the Q-Q plot to vizualize the distribution of residuals.

We saw in the earlier section that the crime rate has some positive skew, but we were able to reduce the skew by applying log transform to the crime rate.

We can also run a Shapiro - Wilk test for normality of the residuals

$H_0 : Normality$

```
shapiro.test(balanced_model_top_3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  balanced_model_top_3$residuals
## W = 0.97312, p-value = 0.05875
```

The p-value is significant, therefore we reject $H_0 : Normality$

The non-normality of the residuals is statisitically significant for this model.

There is some negative skew from outlier 51 in the transformed variable, however, since we have n>30 under CLT we have OLS estimators are normally distributed.

4. A Regression Table

The results were displayed in stargazer using HC standard errors as part of model selction

- This section has been fully covered under section 2 of the report
- We have include statistical F-tests besides the standard t-tests for regression coefficients to check model validity.
- Additionally the practical significance of the model variable chosen have also been discussed in detail
- Below is the summary of the regression models and the AIC & BIC scores which provides a parsimony adjusted measure of fit

```
stargazer(simple_regression_model, all_in_model_log_level, balanced_model_top_1, balanced_model_top_3,
          type = "text", omit.stat = "f",
          se = list(se.simple_regression_model, se.all_in_model_log_level,
                    se.balanced_model_top_1, se.balanced_model_top_3),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ================================================================================
##                                        Dependent variable:
##                    ------------------------------------------------------------
##                                           log_crmrte
##                           (1)             (2)             (3)             (4)
## --------------------------------------------------------------------------------
## prbarr                                  -1.911***       -1.963***       -2.276***
##                                          (0.382)         (0.399)         (0.333)
##
## prbconv                                 -0.696***       -0.768***       -0.758***
##                                          (0.136)         (0.137)         (0.116)
##
## prbpris                                  -0.080          -0.062
##                                          (0.373)         (0.483)
##
## avgsen                                   -0.010          -0.004
##                                          (0.014)         (0.014)
##
## polpc                                    155.903         175.455*        193.390***
##                                         (84.796)        (82.293)        (53.943)
##
## density                 0.228***         0.106           0.113**         0.116***
##                          (0.030)         (0.057)         (0.035)         (0.028)
##
## taxpc                                    0.003           0.002
##                                          (0.007)         (0.006)
##
## west                                     -0.138
##                                          (0.139)
##
## central                                  -0.127
##                                          (0.090)
##
## urban                                    -0.118
##                                          (0.227)
##
## pctmin80                                 0.009**         0.013***        0.012***
##                                          (0.003)         (0.002)         (0.002)
##
## wcon                                     0.001
##                                          (0.001)
##
## wtuc                                     0.0001
##                                          (0.001)
##
## wtrd                                     0.0002
##                                          (0.002)
##
## wfir                                     -0.001
##                                          (0.001)
##
## wser                                     -0.002
##                                          (0.001)
##
## wmfg                                     -0.0001
##                                         (0.0004)
##
## wfed                                     0.003**
##                                          (0.001)
##
## wsta                                     -0.001
##                                          (0.001)
##
## wloc                                     0.001
##                                          (0.002)
##
## mix                                      -0.274          -0.744
##                                          (0.603)         (0.545)
##
## pctymle                                  3.084**          1.390
##                                          (1.176)         (1.626)
##
## Constant                -3.870***       -4.000***       -3.358***       -3.260***
##                          (0.069)         (0.809)         (0.359)         (0.176)
##
## --------------------------------------------------------------------------------
## Observations               90              90              90              90
## R2                        0.399           0.858           0.795           0.782
```

```
## Adjusted R2              0.392           0.812           0.769           0.769
## Residual Std. Error 0.428 (df = 88) 0.238 (df = 67) 0.264 (df = 79) 0.264 (df = 84)
## ================================================================================
## Note:                                            *p<0.05; **p<0.01; ***p<0.001
```

Parismony adjusted model performance

Though AIC and BIC are both Maximum Likelihood estimate driven and penalize free parameters in an effort to combat overfitting, they do so in ways that result in significantly different behavior. Lets look at one commonly presented version of the methods (which results from stipulating normally distributed errors and other well behaving assumptions):

AIC = -2*ln(likelihood)* + 2k, and BIC = -2*ln(likelihood)* + *ln(N)*k,

where: k = model degrees of freedom (K=2 is default for OLS) N = number of observations

The quick explanation is:

- AIC is best for prediction as it is asymptotically equivalent to cross-validation.
- BIC is best for explanation as it is allows consistent estimation of the underlying data generating process.

When N is large the two models will produce quite different results. Then the BIC applies a much larger penalty for complex models, and hence will lead to simpler models than AIC for very large N.

So we check both IC for our model and in both cases a lower value implies a better parsimony adjusted outcome.

```
AIC(simple_regression_model)
```

```
## [1] 106.5187
```

```
AIC(all_in_model_log_level)
```

```
## [1] 18.41539
```

```
AIC(balanced_model_top_1)
```

```
## [1] 27.70917
```

```
AIC(balanced_model_top_3)
```

```
## [1] 23.3983
```

```
BIC(simple_regression_model)
```

```
## [1] 114.0182
```

```
BIC(all_in_model_log_level)
```

```
## [1] 78.41082
```

```
BIC(balanced_model_top_1)
```

```
## [1] 57.70689
```

```
BIC(balanced_model_top_3)
```

```
## [1] 40.89697
```

5. Omitted Variables

We know that two conditions must hold true for omitted-variable bias to exist in linear regression: - The omitted variable must be a determinant of the dependent variable (i.e., its true regression coefficient must not be zero); and - The omitted variable must be correlated with an independent variable specified in the regression (i.e., cov(z,x) must not equal zero).

In our model the presence of omitted variables leads to the error term being correlated with the regressors. - The presence of omitted-variable bias causes the OLS estimator to be biased and inconsistent. - The direction of the bias depends on the estimators as well as the covariance between the regressors and the omitted variables. - A positive covariance of the omitted variable with both a regressor and the dependent variable will lead the OLS estimate of the included regressor's coefficient to be greater than the true value of that coefficient.

We've identified several key omitted variables that we feel most influence the crime rate but are not represented in the data here.

1. Unemployment Rate - Unemployment is a key indicator for crime rate. We may be able to infer some indication of the frequency of seasonal or part-time work in the construction or service industries from the `wcon` or `wser` variables as they shows an average weekly wage which mght indicate how often workers are employed. However, this estimate is likely not accurate enough to be considered meaningful. Unemployment rates among Americans from minority groups in many locations in the United States are significantly higher than for Americans who identify as white. We can posit that this may have positive bias on `pctmin80`, the percentage of minority residents in the county, and that it would give it more significance than it may deserve in predicting crime rate.
2. Inflation Rate (Consumer Price Index) - Inflation and crime rates are correlated with a positive relationship and the causal link is from inflation and unemployment to crime. Link (https://www.researchgate.net/publication/236736987_Will_Inflation_Increase_Crime_Rate_New_Evidence_from_Bounds_and_Modified_Wald_Tes Inflation causes the purchasing power to reduce and cost of living to increase, consequently crime rates rise as the inflation rate rises. Because of the lag between price and wage adjustments, inflation lowers the real income of low-skilled labor, but rewards property criminals due to the rising demand and subsequent high profits in the illegal market. Inflation in the year represented, 1987, would not be sufficient though as the reduction in purchasing power does not happen immediately, it takes time for inflation to gradually reduce purchasing power. None of the data provided in the study gives us an indication of the inflation rate in a time period before the study however inflation rate and unemployment may be correlated and may have positive bias on one another. These then also would have a positive bias on the percentage of minority residents in the county in predicting crime rate, biasing `pctmin80` away from zero.
3. Childhood Blood Lead Levels (with 18 year lag offset) - The lead–crime hypothesis is the proposed link between elevated blood lead levels in children and increased rates of crime, delinquency, and recidivism later in life. Studies linking blood lead levels (BLL) in children to crime rate typically seek to quantify the BLL 17-18 years before the examined crime rate. One such study used a unique dataset linking preschool blood lead levels (BLLs), birth, school, and detention data for 120,000 children born 1990-2004 in Rhode Island, to estimate the impact of lead on behavior Link (https://www.nber.org/papers/w23392.pdf). The `density` varaible would most likely have a positive correlation with BLL as urban and more industrialized areas typically had greater levels of lead poisoning in groundwater and street surfaces due to heavier vehicle traffic and industrial emissions in denser areas. As density had a positive coefficient, we believe that the omitted variable bias is positive away from zero. This may lead ascribing greater significance to `density` in predicting `crmrte`, particularly in the period up until 18 years after the phase out of leaded gas, which 1987 was within.
4. Income Inequality metrics: There are several measures of income inequaity that could be included in the data: Mean Log Deviation (https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/mld.html) or Theil Index (https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/theil-index.html) or Gini Index (https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html) for each of the counties. Income inequality has been shown to have a significant effect on violent crime in particular. One World Bank report states that inequality predicts about half of the variance in murder rates between American states and between countries around the world. Link (https://siteresources.worldbank.org/DEC/Resources/Crime%26Inequality.pdf) Income inequality measures are often measured as 0 (perfectly equal income distribution) to 1 (perfectly unequal income distribution, or 1 household has all the income). We would thus expect these to have a positive bias, in that an increase in income inequality would lead to an increase in violent crime. A correlation in American cities between density and income inequality has been shown by Edward Glaseer (https://scholar.harvard.edu/files/glaeser/files/urban_inequality.pdf) leading us to believe that income inequality shows a positive bias on `density` We believe that the omitted variable bias is positive away from zero and that this plays some role in over-estimating the significance of density in predicting crime rate.

Summary:

Unemployment rate is positively correlated with crime rate and positively correlated with percentage of minorities in each county. The coefficient on `pctmin80` is 0.012099, therefore the omitted variable bias is positive away from zero.

Inflation rate is also positively correlated with crime rate and positively correlated with percentage of minorities in each county. The coefficient on `pctmin80` is 0.012099, therefore the omitted variable bias is positive away from zero.

Childhood BLL is positively correlated with crime rate and positively correlated with density. The coefficient on `density` is 0.115702, so we surmise that the bias is positive away from zero.

Income inequality is positively correlated with crime rate and positively correlated with density. The coefficient on `density` is 0.012099 and we surmise that as density rises that income inequality rises as well, so the bias is positive away from zero.

6. A Conclusion

Using the 1987 dataset, we were able to identify the key demographic and deterrent variables that affect crime. Our final model shows that arrests (prbarr), convictions (prbconv), and police presence (polpc) are deterrents that affect crime rates. As indicated by the negative coefficients, an increase in arrests and convictions predict a decrease in crime rate, suggesting that these variables are key deterrents in reducing crime. While the probability of arrest and convictions are statistically significant, they also have practical significance. Our study shows improving the measures of arrest and convincitions by 10% could potentially lead a reduction in crime rate by 25%. This is a number that can resonate with voters and we recommend that this headlining statistic lead the campaign.

While the coefficeint for police presence predicts an increase in crime rates, it should be noted that this does not indicate that additional police cause an increase in crime. The results suggest that with additional police there will be an initial increase in apprehension of criminals, resulting in an increase in police reports. The actual reduction in committed crimes due to a larger police presence may see a lag effect which could potentially be corrected by the use of Fixed Effects regression model. In addition, police presence and crime rate affect each other and causes simultaneity bias in the model, which states that the dependent and indepedent variable influence each other at the same time. Cornwell and Trunbull(1994) also raise this issue in their paper and they employ 2 SLS model to adjust for the effects of simultaneity. While this issue is worth reviewing further, it is outside the scope of this report.

The demographic categories of density, minorities (pctmin80), and offense (mix) are statistically significant variables of crime rate. While the results show a strong statistical relationship with crime rates and percentage of minorities, this result does not definitively denote a causal affect that an increase in minorities predicts an increase in crime. The complex nature of race relations should be considered. The relationship between minorties and crime could be a result of racial bias within the police force and criminal justice system, leading to a disproportionate number of minorities being arrested and convicted for crimes.

While the results of this report have shown that the economic data does not hold statistical significance in our analysis, that is not an indication that economic factors are not key determinants on crime. The wage statistics provided in the dataset, while important, do not address issues of poverty and inequality which are traditional drivers of crime. As described in the report, omitted variables such as unemployment rate, inflation rate, and income inequality are key economic factors that have a significant affect on crime and should be examined further.

Based on the results of our study, we propose a political strategy that will focus on deterrents and demographic factors to address crime:

- The first recommendation is to increase the number of arrests by focusing on providing the police force with the proper funding, training and resources. As shown in the results, the prbarr is a practically significant deterrent of crime showing that a .01 increase in the variable would predict a 2.25% decrease in crime rate.
- The second recommendation is to increase the number of convictions by providing additional resources to the criminal justice system. However, it is recommended that punishment for convictions should not be prison sentences unless necessary and justified. Our study suggests that prison sentences are not a significant deterrent of crime, while also having adverse effects of overcrowding of the prison system and burdening taxpayers.
- The third recommendation is to focus on developing local economies and increasing job opportunities in areas of high unemployment. The results of this study show a significant associative relationship between minorities and crime rate that should be addressed. While the dataset is limited in determining causal reasons for this issue, our team has identified employment rate as an omitted variable that disproportionately impacts minorities. We recommend providing resources that can stimulate local economies and boost the employment rate.