

# Lab 3: Reducing Crime (DRAFT: Stage 1)

*C. Akkineni, A. Thorp, K. Hanna*

*November 27, 2018*

## Contents

<b>Introduction (Stage 1: Draft Report)</b>	<b>2</b>
<b>Exploratory Data Analysis</b>	<b>3</b>
Data Summary . . . . .	3
Data Clean Up . . . . .	3
Concerns about data . . . . .	3
Univariate Analysis . . . . .	3
Key Interests . . . . .	4
Explanatory . . . . .	4
<b>Model Analysis</b>	<b>9</b>
Model1 - Minimum Specification . . . . .	9
Checking if our Assumptions are valid in the model- . . . . .	9
Model2 - Optimal Specification . . . . .	14
Checking if our Assumptions are valid in the model- . . . . .	14
Model3 - Optimal-2 Specification . . . . .	18
Checking if our Assumptions are valid in the model- . . . . .	19
Model4 - Using all variables Specification . . . . .	23
Checking if our Assumptions are valid in the model- . . . . .	24
<b>Omitted Variables</b>	<b>32</b>
<b>Conclusion</b>	<b>34</b>
<b>Appendix A: Codebook</b>	<b>35</b>

## Introduction (Stage 1: Draft Report)

The team has been hired to provide research for a political campaign and help the campaign understand the determinants of crime and to help with policy suggestions that are applicable to local government.

```
library(knitr)
library(kableExtra)
library(car)

## Loading required package: carData

library(stargazer)

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

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

codebook <- read.csv('codebook.csv')
crime <- read.csv('crime_v2.csv')

# Convert columns to factors and logical.
crime$county <- as.factor(crime$county)
crime$year <- as.factor(crime$year)
crime$west <- as.logical(crime$west)
crime$central <- as.logical(crime$central)
crime$urban <- as.logical(crime$urban)

# Create a log of the dependent variable
crime$logcrmrte <- log(crime$crmrte)
crime$east <- !(crime$west | crime$central)

# Create an average of all weekly wages values.
crime$avgwage = (crime$wcon + crime$wtuc + crime$wtrd + crime$wfir +
                 crime$wser + crime$wmfg + crime$wfed + crime$wsta + crime$wloc)/9

# Possible transformation for prbconv
#crime$adjprbconv = crime$prbconv/max(crime$prbconv)

# Reorder to place logcrmrte next to crmrte and east next to central
# It's unsafe to keep this in, but handy for viewing data
#crime <- crime[,c(1,2,3,26,4:13,27,14:25,28)]
```

# Exploratory Data Analysis

## Data Summary

We were provided with a dataset of crime statistics for a selection of counties in North Carolina. After performing data clean up (outlined below) the data set contained 90 county observations each having 25 variables (outlined in the codebook found in Appendix A).

## Data Clean Up

### Null Rows

The dataset contained a an apostrophe 6 rows after the data which caused the csv reader to create 6 invalid rows. We feel it is safe to remove these rows as they contain no data.

```
# Delete the 6 empty observations at the end, including the row with the apostrophe.  
# We can use complete.cases to do this as these 6 observations are the only incomplete observations.  
crime = crime[complete.cases(crime), ]  
  
# Fix prbconv which is a factor rather than numeric due to the apostrophe  
# Convert from factor to numeric  
crime$prbconv = as.numeric(as.character(crime$prbconv))
```

We found two identical observations for county 193. There is no logical reason to have two identical observations in this cross-sectional data, so we feel strongly that removing one of these two observations can only benefit our analysis.

```
# county 193 is duplidated, remove one  
crime = crime[!duplicated(crime), ]
```

## Concerns about data

There are three probability columns in the given dataset. Check if any of the columns has invalid values - i.e., any of the columns have less than zero or greater than 1 values.

### prbarr (Probability of Arrest)

We found that county 115 contained a value of 1.09 in prbarr (probability of arrest) which is not possible. We beleive this to be a coding error, it is a ratio and not a probability.

### prbconv (Probability of Conviction)

We found 10 observations with values greater than 1, which, again, is not a possible value for probability. The documentation in the codebook specifies that “(t)he probability of conviction is proxied by the ratio of convictions to arrests”, which leaves some ambiguity, however it is plausible to have values greater than 1 as a single arrest can result in multiple convictions.

## Univariate Analysis

```
quick_uni_analysis = function(variable, description, roundto = 8) {  
  hist(variable, xlab = paste(tools::toTitleCase(description),  
    paste('\n Shapiro:',
```

```

round(as.numeric(shapiro.test(variable)[2]), roundto)
)), main = "")

hist(log(variable),
     xlab = tools::toTitleCase(paste('Log of', description,
                                     paste('\n Shapiro:', round(as.numeric(shapiro.test(log(variable))[2]), roundto)
                                     ))), main = "", ylab = '')
}

```

## Key Interests

### Crimes Committed Per Person

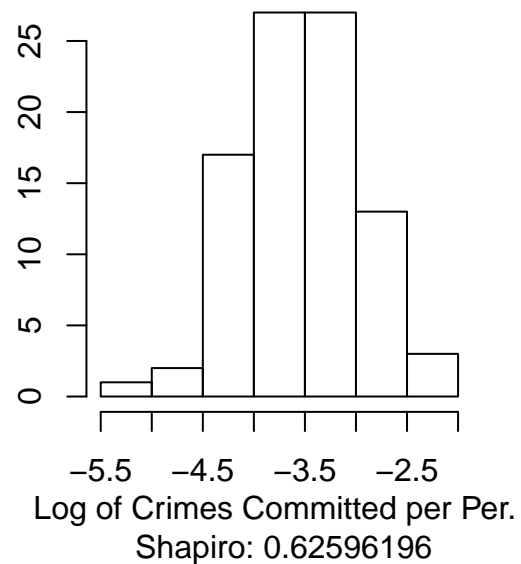
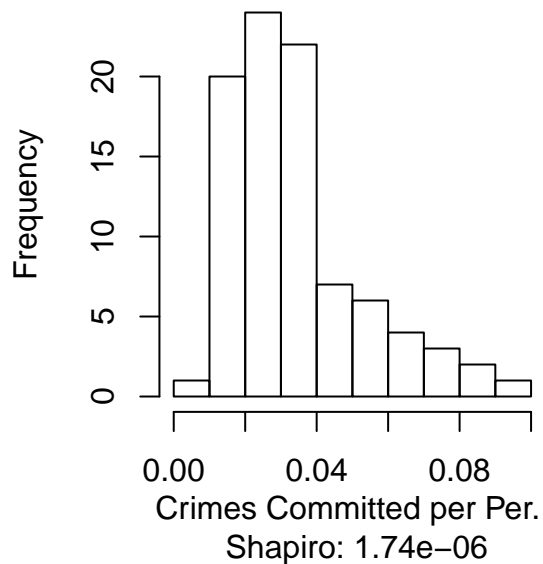
**Campaign Significance:** Crime rate is a politicized and effects economy

This is the key variable we will be regressing to in our modeling.

```

par(mfrow=c(1,2))
quick_uni_analysis(crime$crmrte, 'crimes committed per per.')

```



Crimes committed per capita has a positive skew, applying a natural log transformation creates a more symmetrical distribution and results in a Shapiro-Wilk test p-value that we cannot reject.

The transformed variable is preferable for modelling.

## Explanatory

### Diagrams of Key Variables With and Without Log Transformations

```

par(mfrow=c(5,4))
quick_uni_analysis(crime$prbarr, 'Probability of Arrest', roundto = 10)
quick_uni_analysis(crime$prbconv, 'Probability of Conviction', roundto = 10)

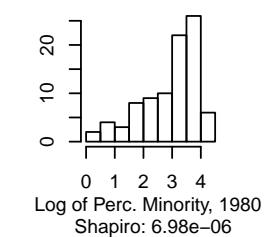
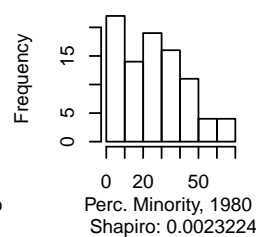
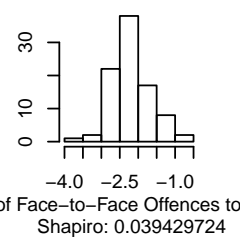
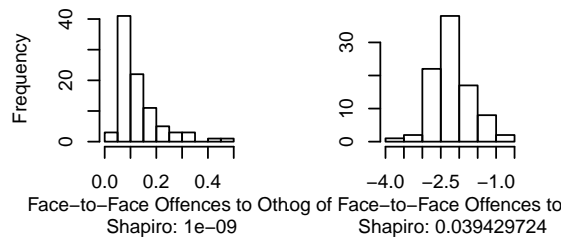
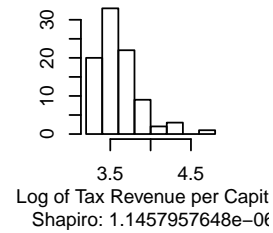
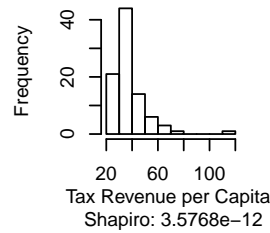
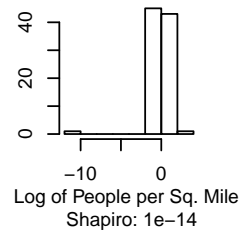
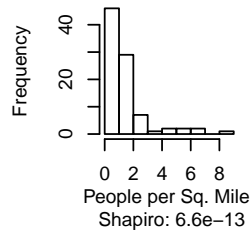
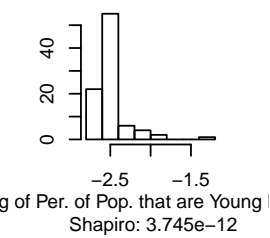
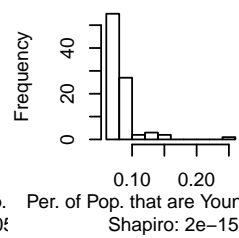
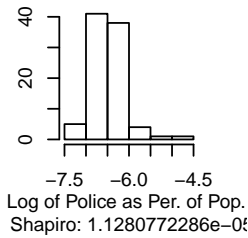
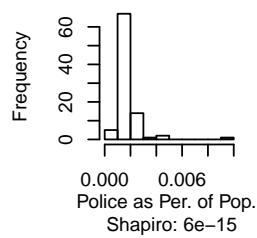
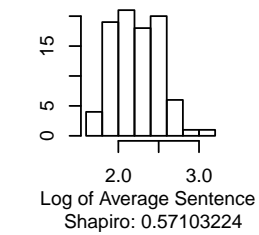
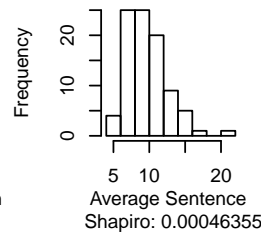
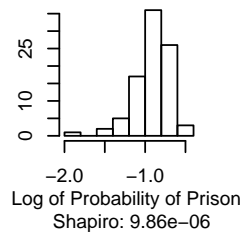
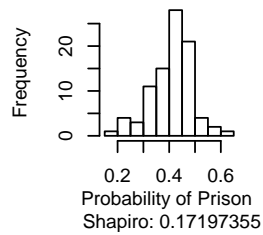
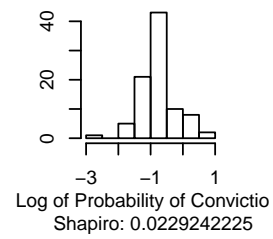
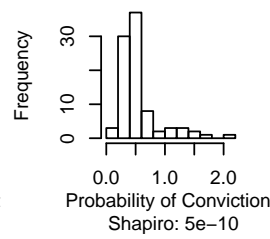
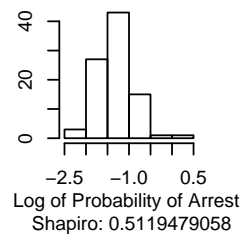
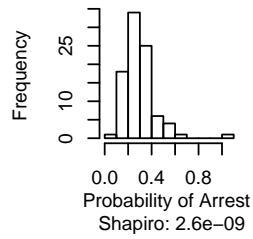
quick_uni_analysis(crime$prbpris, 'Probability of Prison')
quick_uni_analysis(crime$avgsen, 'Average Sentence')

quick_uni_analysis(crime$polpc, 'Police as Per. of Pop.', roundto = 15)
quick_uni_analysis(crime$pctymle, 'Per. of Pop. That Are Young Males', roundto = 15)

quick_uni_analysis(crime$density, 'people per sq. mile', roundto = 14)
quick_uni_analysis(crime$taxpc, 'tax revenue per capita', roundto = 16)

quick_uni_analysis(crime$mix, 'Face-to-face offences to other', roundto = 9)
quick_uni_analysis(crime$pctmin80, 'perc. minority, 1980')

```



```
par(mfrow=c(1,1))
```

### **Probability of Arrest**

Probability of Arrest has a positive skew, applying a natural log transformation creates a more symmetrical distribution and results in a Shapiro-Wilk test p-value that we cannot reject.

The transformed variable is preferable for modelling.

### **Probability of Conviction**

Log is preferable - both for interpretation and for better adhering to modeling assumptions. However, even the logged version fails a Shapiro-Wilk normality test. Something to keep in mind.

### **Probability of Prison**

From an interpretation standpoint, the logged version is preferable, although from an modeling assumption standpoint, the unlogged version is preferable.

### **Average Sentence**

The logged version is preferable from both an interpretation and modeling assumption standpoint.

### **Police as a Percentage of Population**

Both logged and un-logged versions of police as a percentage of the population are non-normal. Neither is inherently preferable from a modeling assumptions standpoint.

The number of police can conceivably be either the cause of or result of crime rates. However, there is a fair amount of research showing increasing the police workforce causes a reduction in crime according to Vollaard & Hamad <sup>(1)</sup> . Without timeseries data including changes to Police as a Percentage of Population we cannot determine the true effects. However, many studies show that increasing the size of a police force does reduce crime.

Both logged and un-logged versions of the percent of population that is young and male are non-normal. Neither is inherently preferable from a modeling assumptions standpoint.

<sup>(1)</sup> The Journal of Law & Economic <https://www.jstor.org/stable/10.1086/666614>

### **Per. of Pop. That Are Young Males**

### **People per Square Mile**

### **tax revenue per capita**

### **Face-to-face offences to Other (Offence Mix)**

This is a ratio of face-to-face crimes to all other crimes. Face-to-face crimes include violent crimes and those with a higher probability of violence, hence the more severe crimes. Focusing resources which reduces this ratio along with the overall crime rate would be more beneficial.

**Campaign Significance:** Violent crimes create fear and fear is a strong motivator for voters.

Mix of face-to-face crimes to other crimes has a positive skew, applying a natural log transformation creates a more symmetrical distribution however, the resulting Shapiro-Wilk test would be rejected at 0.039. That said, the log transformation is

The transformed variable is preferable for modelling.

### **Percentage Minority, 1980**



## Model Analysis

```
crime$log_crmrte <- log(crime$crmrate)
crime$log_prbarr <- log(crime$prbarr)
crime$log_prbconv <- log(crime$prbconv)
crime$log_prbpris <- log(crime$prbpris)
crime$log_avgsen <- log(crime$avgsen)
crime$log_polpc <- log(crime$polpc)
crime$log_density <- log(crime$density)
crime$log_taxpc <- log(crime$taxpc)
crime$log_pctmin80 <- log(crime$pctmin80)
crime$log_mix <- log(crime$mix)
crime$log_pctymle <- log(crime$pctymle)
```

### Model1 - Minimum Specification

Crime-Determinants: we anticipate crime rate depends on average sentencing as higher crime rate tends to have higher sentencing. And Increased avg sen suggests there are severe crimes happening in a given county. The two probability variables prbarr and prbconv will have strong correlations with prbpris and so we are including them as well in our model so that we can measure how much these variables influence crime rate.

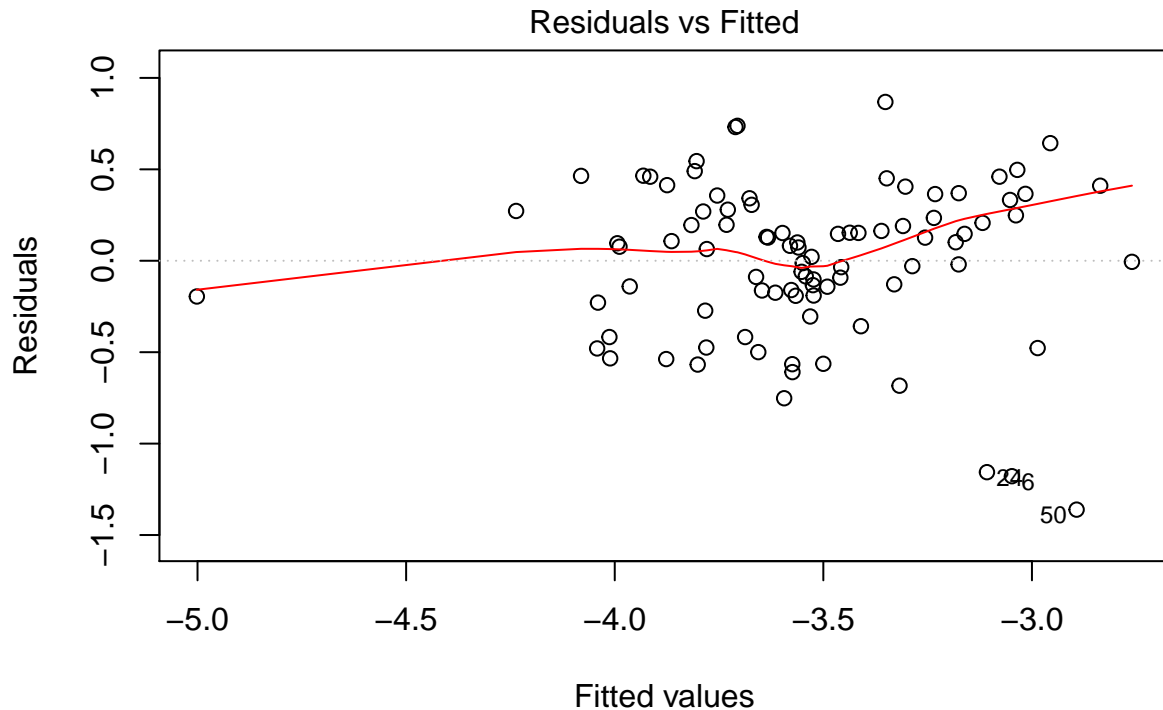
$$\log(\text{crmrate}) = \beta_0 + \beta_1 \log(\text{prbarr}) + \beta_2 \log(\text{prbconv}) + \beta_3 \log(\text{prbpris}) + \beta_4 \log(\text{avg sen}) + u$$

```
model1 = lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avg sen), data = crime)
```

### Checking if our Assumptions are valid in the model-

Assumption 1 - Linear model -> the specified model has the dependent variable linear with explanatory variables

```
plot(model1, which=1)
```



$\text{lm}(\log(\text{crm rte}) \sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{prbpris}) + \log(\text{avgsen}))$

from the Residuals vs Fitted plot, we don't see any non-linear relationship. So, it is a VALID Assumption.

Assumption - 2 Random Sampling Since the dataset has counties information from only few regions it is not truly randomly sampled. But, we were provided with information that how the data is collected. Based on this we think it is a VALID Assumption.

Assumption -3 Multicollinearity

```
cor_model1 <- data.matrix(subset(crime, select = c("log_crm rte", "log_prbarr", "log_prbconv", "log_prbpris", "log_avgsen")))
cor(cor_model1)
```

```
##          log_crm rte    log_prbarr log_prbconv log_prbpris    log_avgsen
## log_crm rte    1.00000000 -0.435753889 -0.37249610  0.06960729  0.023417173
## log_prbarr   -0.43575389  1.000000000 -0.20235541 -0.02801506  0.003832922
## log_prbconv  -0.37249610 -0.202355412  1.00000000  0.01053523  0.011629954
## log_prbpris   0.06960729 -0.028015058  0.01053523  1.00000000 -0.123723380
## log_avgsen    0.02341717  0.003832922  0.01162995 -0.12372338  1.000000000
```

We are not seeing any obvious signs of multicollinearity.

```
vif(model1)
```

```
##  log(prbarr) log(prbconv) log(prbpris)  log(avgsen)
##    1.043435    1.042909    1.016372    1.015728
```

Computing VIF also indicates that there is no multicollinearity. So, it is a VALID Assumption.

Assumption - 4 Exogeneity (Zero Conditional Mean) From the Residuals Vs Fitted Plot, the red line is very influenced by the outliers on the ends. It is a Most-Likely valid Assumption.

Additional Assumptions - Homoscedasticity From the same Residuals Vs Fitted plot, we see it is very scattered with extreme outliers. So, it is not easy to determine Homoscedasticity from this plot only.

```
bptest(model1)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: model1  
## BP = 10.702, df = 4, p-value = 0.03012
```

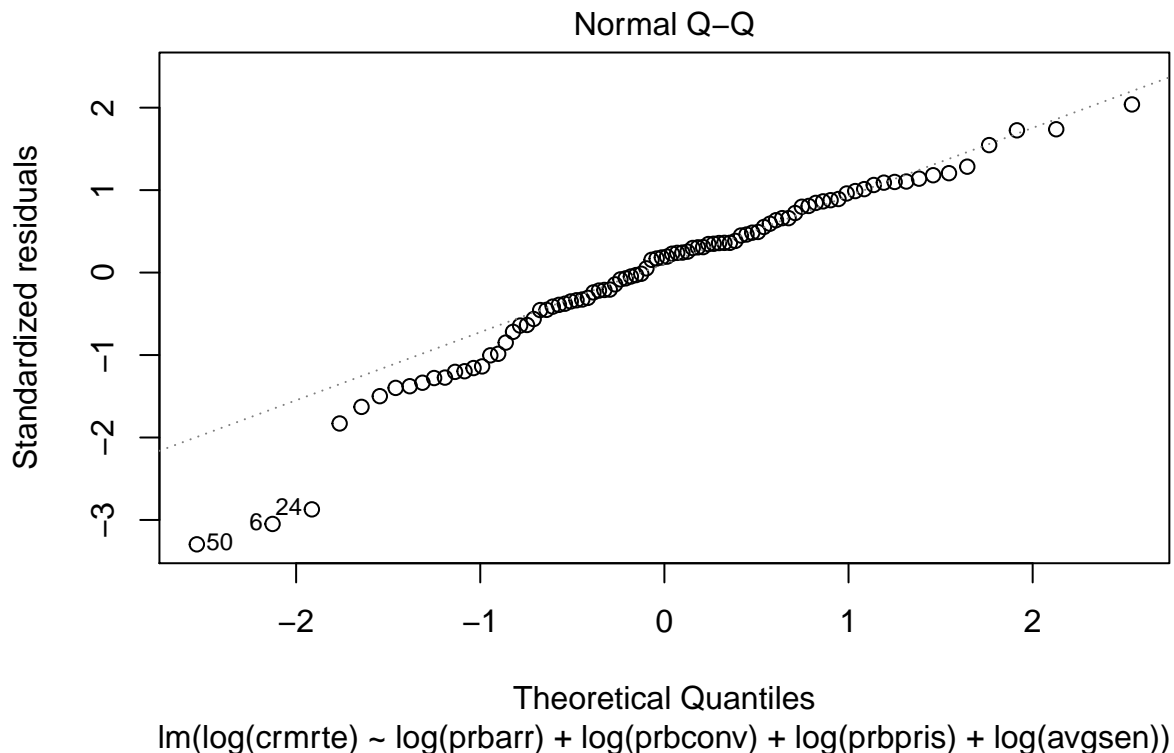
```
ncvTest(model1)
```

```
## Non-constant Variance Score Test  
## Variance formula: ~ fitted.values  
## Chisquare = 7.467969, Df = 1, p = 0.0062806
```

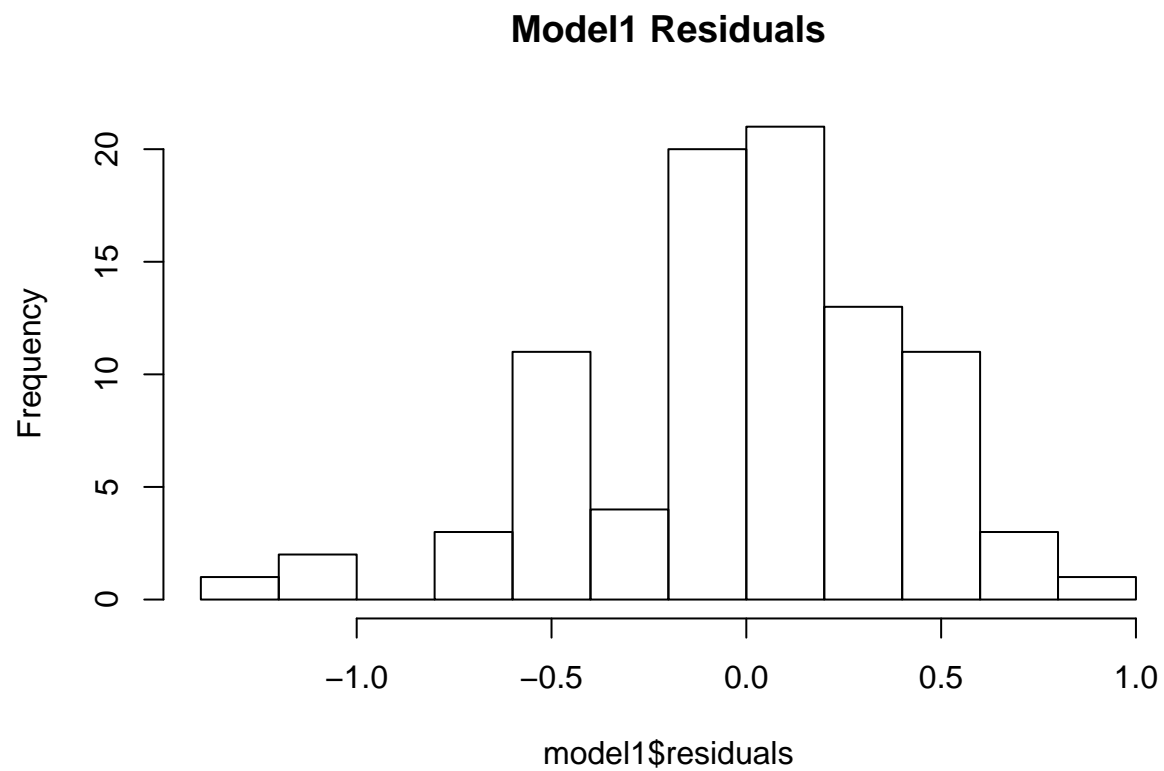
Both tests are showing small p-values showing that we have to reject the hypothesis. So Homoscedasticity is not a valid assumption here indicating that our explanatory variables may not be able to explain crime rate highly significant.

Normality of Residuals

```
plot(model1, which=2)
```



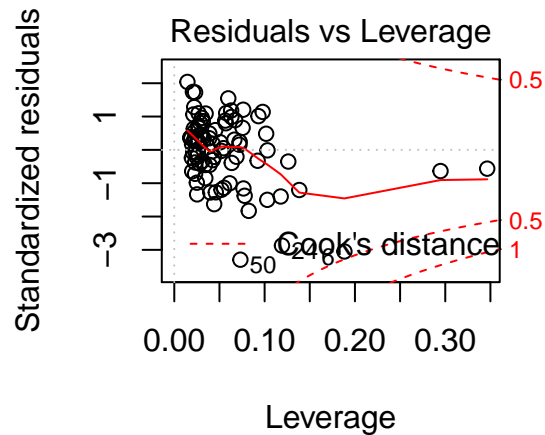
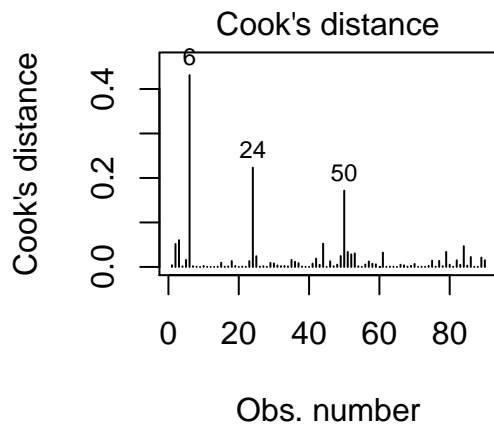
```
hist(model1$residuals,main="Model1 Residuals")
```



Other than a few outliers, the distribution is relatively normal for our given sample size. So, it is a VALID assumption.

Cook's Distance:

```
plot(model1, which=4)  
plot(model1, which=5)
```



$\epsilon \sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{prbpri})$ 
 $\epsilon \sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{prbpri})$

There are some influential values however cook's distance is within the bounds

```
AIC(model1)
```

```
## [1] 109.9382
```

The AIC for this model is 110.0643

```
summary(model1)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##     log(avgsen), data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36104 -0.19129  0.07939  0.27754  0.86843
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.86792    0.43153  -11.281  < 2e-16 ***
## log(prbarr)   -0.72397    0.11532   -6.278 1.39e-08 ***
## log(prbconv) -0.47251    0.08311   -5.686 1.80e-07 ***
## log(prbpris)  0.15967    0.20644    0.773  0.441
## log(avgsen)   0.07642    0.16347    0.467  0.641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 85 degrees of freedom
## Multiple R-squared:  0.4162, Adjusted R-squared:  0.3888
## F-statistic: 15.15 on 4 and 85 DF,  p-value: 2.171e-09
```

Looking at the summary of model2, 38.98% of crime rate is being explained by our model. Looking at the co-efficient values -> prbarr, prbconv has negative effect on crime. While other independent variables prbpris and avgsen has postive effect on crime. Looking at the  $\text{Pr}(>|t|)$  values -> crime rate fluctuates more with prbarr, prbconv

This says that crime rate will decrease if people are arrested and convicted.

## Model2 - Optimal Specification

In addition to the explanatory variables introduced in our #Model1, we have decided to include the following variables in the model.

Demographics - The team anticipates that crime behavior alters with demographics information such as race, gender, age. So, we are including pctymle and pctmin80 variables in our model.

Density - The team is interested to see how density is altering crime rates. The team expects that this should have a negative effect on crime.

Income Variables - The team expects that higher tax money means less crimes. Similar crimes would be less with better policing. So, we are including polpc and taxpc variables in our model.

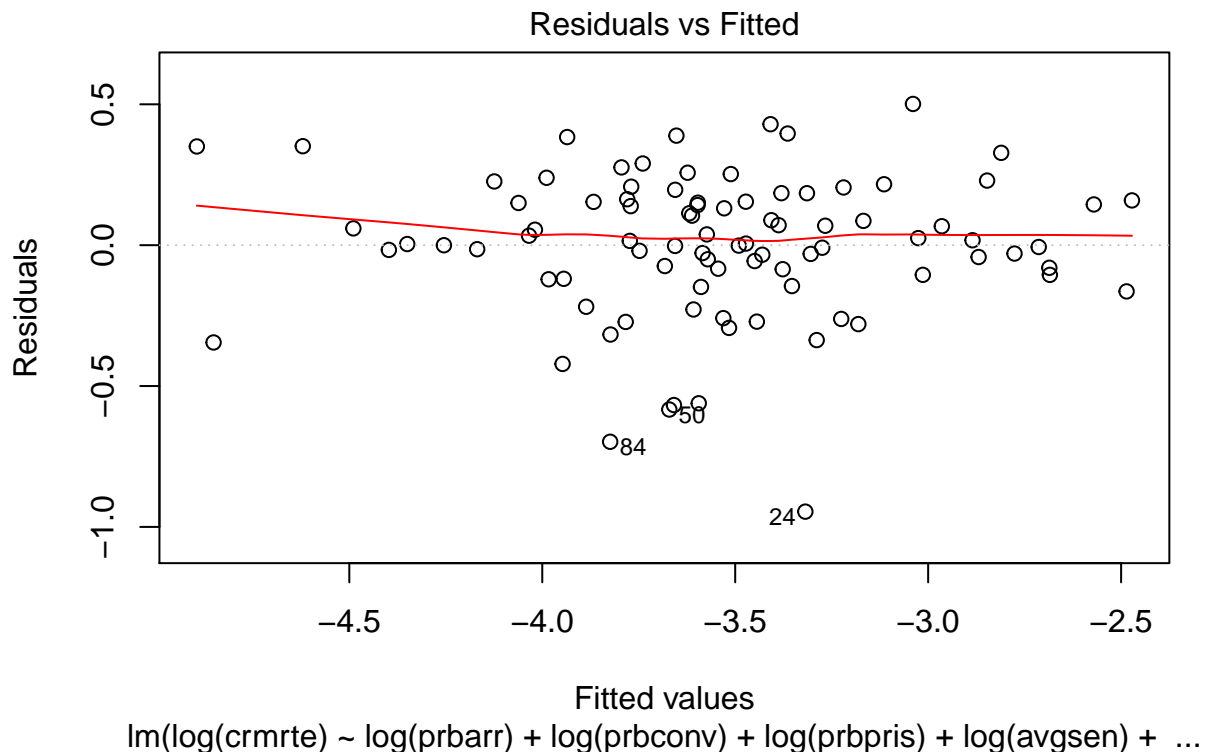
$$\begin{aligned} \log(\text{crmrte}) = & \beta_0 + \beta_1 \log(\text{prbarr}) + \beta_2 \log(\text{prbconv}) + \beta_3 \log(\text{prbpris}) + \\ & \beta_4 \log(\text{avgse}) + \beta_5 \log(\text{polpc}) + \beta_6 \log(\text{taxpc}) + \\ & \beta_7 \log(\text{density}) + \beta_8 \log(\text{pctymle}) + \beta_9 \log(\text{pctmin80}) + u \end{aligned}$$

```
model2 = lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgse)
           + log(polpc) + log(taxpc) + log(density) + log(pctymle) + log(pctmin80), data = crime)
```

### Checking if our Assumptions are valid in the model-

Assumption 1 - Linear model -> the specified model has the dependent variable linear with explanatory variables

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



from the Residuals vs Fitted plot, we don't see any non-linear relationship. So, it is a VALID Assumption.

Assumption - 2 Random Sampling Since the dataset has counties information from only few regions it is not truly randomly sampled. But, we were provided with information that how the data is collected. Based on this we think it is a VALID Assumption.

Assumption -3 Multicollinearity

```
cor_model2 <- data.matrix(subset(crime, select = c("log_crmrte", "log_prbarr", "log_prbconv", "log_prbpris", "log_avgsen", "log_polpc", "log_taxpc", "log_density", "log_pctymle", "log_pctmin80")))
cor(cor_model2)
```

```
##          log_crmrte  log_prbarr log_prbconv log_prbpris  log_avgsen
## log_crmrte    1.00000000 -0.435753889 -0.37249610  0.06960729  0.023417173
## log_prbarr   -0.43575389  1.000000000 -0.20235541 -0.02801506  0.003832922
## log_prbconv  -0.37249610 -0.202355412  1.00000000  0.01053523  0.011629954
## log_prbpris   0.06960729 -0.028015058  0.01053523  1.00000000 -0.123723380
## log_avgsen    0.02341717  0.003832922  0.01162995 -0.12372338  1.000000000
## log_polpc     0.28453961  0.054242849 -0.13863932 -0.01046407  0.395079641
## log_taxpc     0.33984322 -0.111996061 -0.20622995 -0.05671890  0.077900054
## log_density   0.49364251 -0.354779226 -0.07258589  0.41571758  0.159651910
## log_pctymle   0.31175403 -0.282255476 -0.22382225 -0.02486072  0.103327744
## log_pctmin80  0.39729097 -0.051223276  0.07171367  0.02897773 -0.164025122
##          log_polpc  log_taxpc log_density  log_pctymle log_pctmin80
## log_crmrte    0.28453961  0.33984322  0.49364251  0.31175403  0.397290965
## log_prbarr    0.05424285 -0.11199606 -0.35477923 -0.282255476 -0.051223276
## log_prbconv   -0.13863932 -0.20622995 -0.07258589 -0.223822249  0.071713668
## log_prbpris   -0.01046407 -0.05671890  0.41571758 -0.024860719  0.028977726
## log_avgsen    0.39507964  0.07790005  0.15965191  0.103327744 -0.164025122
## log_polpc     1.00000000  0.38298724  0.02304206  0.176404361 -0.161898455
## log_taxpc     0.38298724  1.00000000  0.10798692 -0.073609432  0.121027860
## log_density   0.02304206  0.10798692  1.00000000  0.175156111 -0.018819680
## log_pctymle   0.17640436 -0.07360943  0.17515611  1.000000000  0.008048105
## log_pctmin80 -0.16189846  0.12102786 -0.01881968  0.008048105  1.000000000
```

We are not seeing any obvious signs of multicollinearity.

```
vif(model2)
```

```
##      log(prbarr)  log(prbconv)  log(prbpris)  log(avgsen)  log(polpc)
##      1.469188    1.297357      1.344846      1.332245      1.587111
##      log(taxpc)  log(density)  log(pctymle) log(pctmin80)
##      1.465737    1.595347      1.365754      1.105714
```

Computing VIF also indicates that there is no multicollinearity. So, it is a VALID Assumption.

Assumption - 4 Exogeneity (Zero Conditional Mean) From the Residuals Vs Fitted Plot, the red line is very influenced by the outliers on the ends. But it is close to x-axis. It is a Most-Likely valid Assumption.

Additional Assumptions - Homoscedasticity From the same Residuals Vs Fitted plot, we see it is very scattered with extreme outliers. So, it is not easy to determine Homoscedasticity from this plot only.

```
bptest(model2)
```

```
##
## studentized Breusch-Pagan test
##
## data:  model2
## BP = 17.355, df = 9, p-value = 0.04344
```

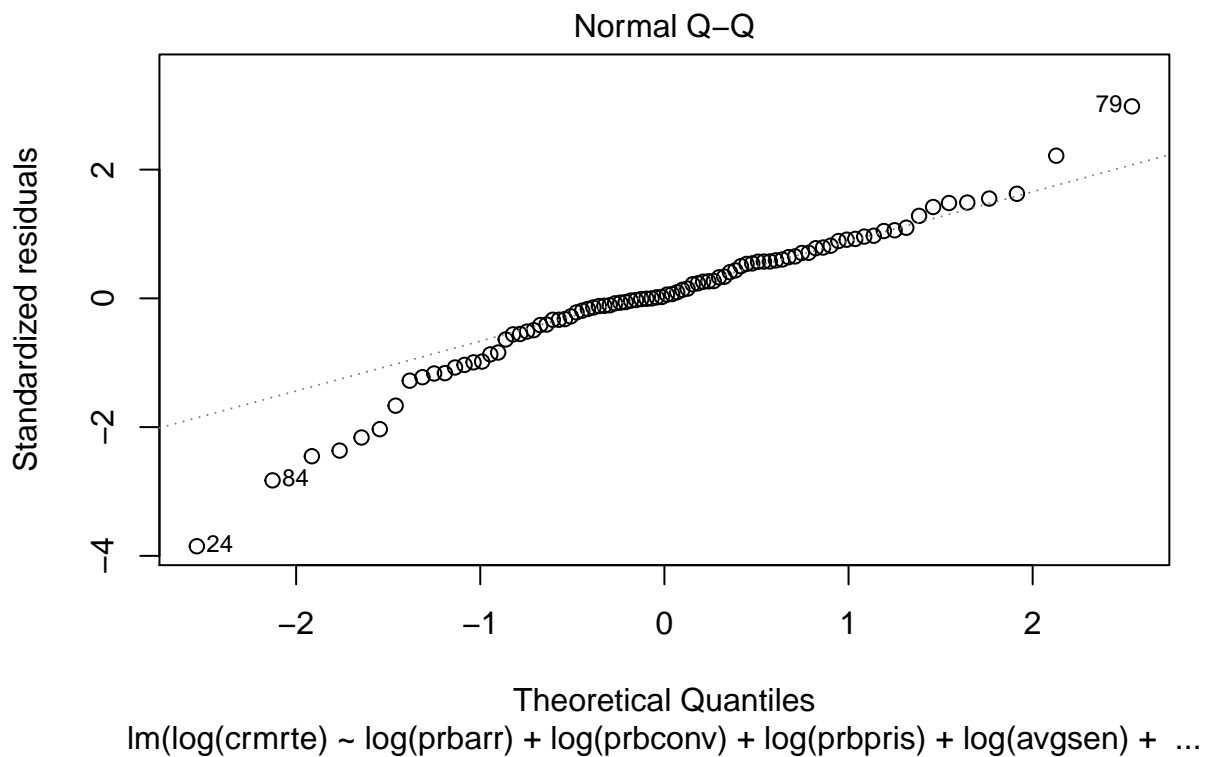
```
ncvTest(model2)
```

```
## Non-constant Variance Score Test  
## Variance formula: ~ fitted.values  
## Chisquare = 0.5980037, Df = 1, p = 0.43934
```

Both tests are showing small p-values showing that we fail to reject the hypothesis. So Homoscedasticity is a valid assumption here.

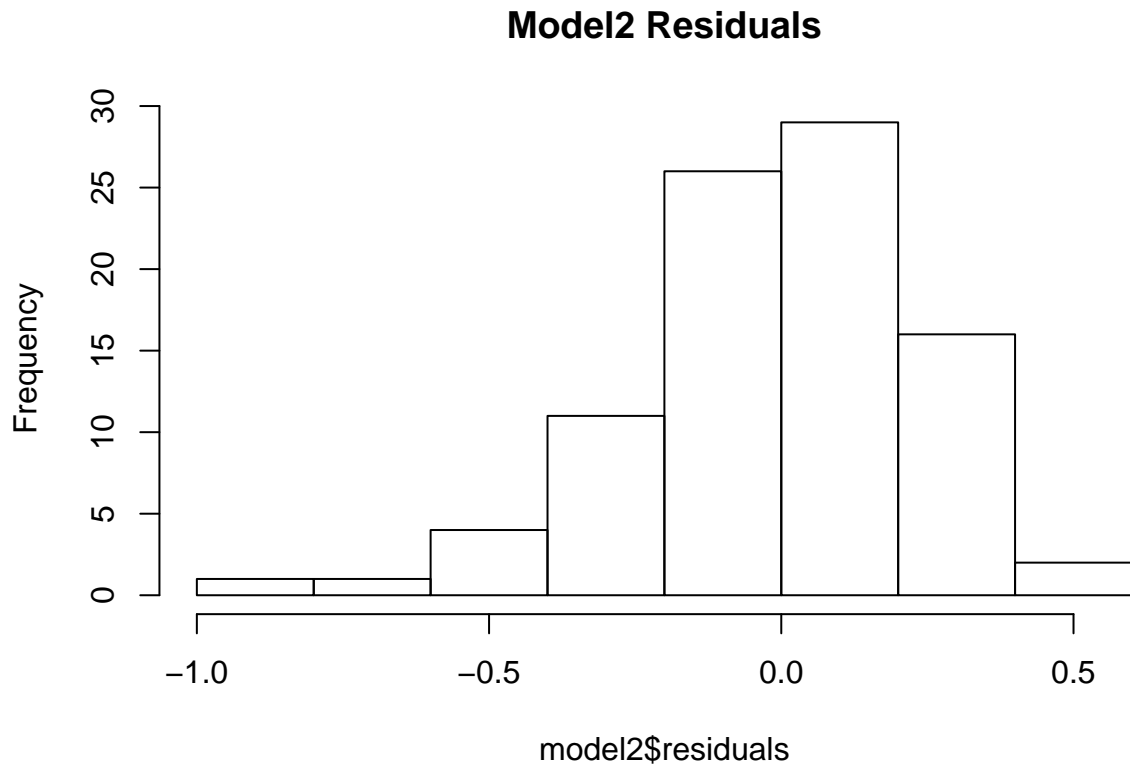
Normality of Residuals

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



```
hist(model2$residuals, main="Model2 Residuals")
```

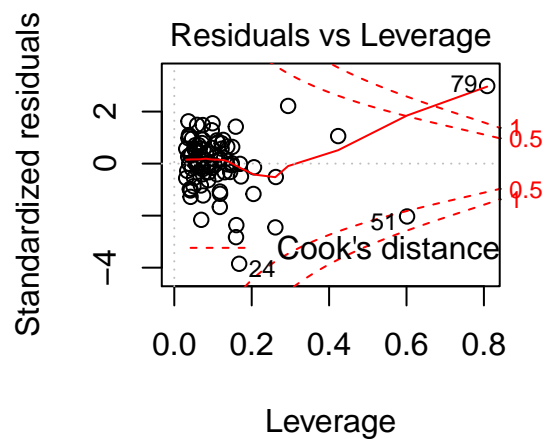
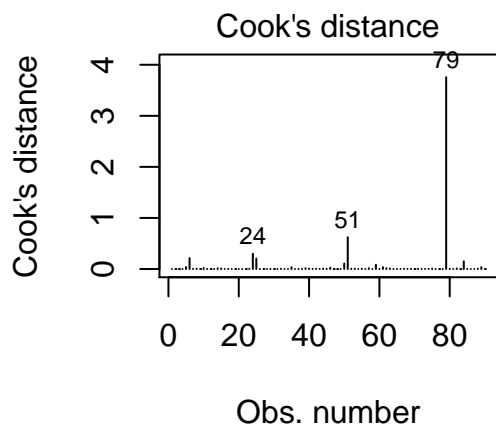




Other than a few outliers, the distribution is relatively normal for our given sample size. So, it is a VALID assumption.

Cook's Distance:

```
plot(model2, which=4)
plot(model2, which=5)
```



$\sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{prbpris}) \sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{prbpris}$  There are some influential values [24,51,79] however cook's distance is within the bounds

```
AIC(model2)
```

```
## [1] 30.5825
```

The AIC for this model is 30.50572

```
summary(model2)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##     log(avgsen) + log(polpc) + log(taxpc) + log(density) + log(pctymle) +
##     log(pctmin80), data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94572 -0.10537  0.01043  0.15780  0.50122
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.41661    0.98533  -1.438   0.1544
## log(prbarr)   -0.52248    0.08586  -6.086 3.80e-08 ***
## log(prbconv)  -0.40321    0.05816  -6.933 9.55e-10 ***
## log(prbpris)  -0.33191    0.14900  -2.228   0.0287 *
## log(avgsen)   -0.23543    0.11747  -2.004   0.0484 *
## log(polpc)     0.55270    0.09605   5.754 1.54e-07 ***
## log(taxpc)    -0.06765    0.13035  -0.519   0.6052
## log(density)   0.16403    0.02624   6.251 1.87e-08 ***
## log(pctymle)  -0.06600    0.16805  -0.393   0.6956
## log(pctmin80)  0.26553    0.03128   8.489 8.87e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2692 on 80 degrees of freedom
## Multiple R-squared:  0.7837, Adjusted R-squared:  0.7594
## F-statistic: 32.21 on 9 and 80 DF,  p-value: < 2.2e-16
```

Looking at the summary of model2, 75.78% of crime rate is being explained by our model. Looking at the co-efficient values -> Polpc, Density and Pctmin80 has a positive effect on crime. While other independent variables prbarr,prbconv,prbpris,avgsen,taxpc,pctymle has negative effect on crime. Looking at the Pr(>|t|) values -> crime rate fluctuates more with prbarr, prbconv, density, pctmin80 and polpc and has some impact with prbpris and avgsen

This says that crimerate decreases if more people are arrested, convicted, sentenced in prison With better policing in place more crimes would be identified. And as density increases crimerates tend to go up. Also, Minority has a positive impact on crime.

## Model3 - Optimal-2 Specification

In additional to the explanatory variables intorudced in our #Model1, we have decided to include the following variables in the model.

Demographics - The team anticipates that crime behavior alters with demographics information such as race, gender, age. So, we are including pctymle and pctmin80 variables in our model.

Density - The team is interested to see how density is altering crime rates. The team expects that this should have a negative effect on crime.

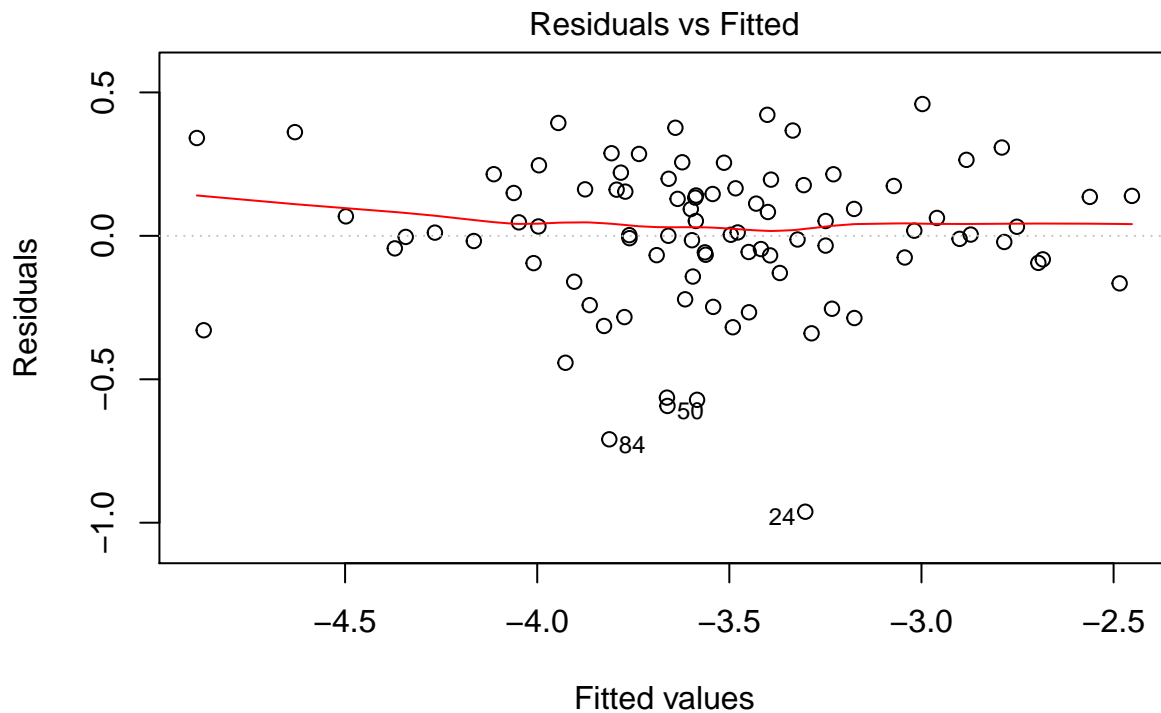
$$\log(\text{crm rte}) = \beta_0 + \beta_1 \log(\text{prbarr}) + \beta_2 \log(\text{prbconv}) + \beta_3 \log(\text{prbpris}) + \beta_4 \log(\text{avg sen}) + \beta_5 \log(\text{polpc}) + \beta_6 \log(\text{density}) + \beta_7 \log(\text{pctmin80}) + u$$

```
model3 = lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc) + log(density) + log(pctmin80) + log(pctmin80))
```

Checking if our Assumptions are valid in the model-

Assumption 1 - Linear model -> the specified model has the dependent variable linear with explanatory variables

```
plot(model3, which=1)
```



$\text{lm}(\log(\text{crm rte}) \sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{polpc}) + \log(\text{density}) + \dots$

from the Residuals vs Fitted plot, we don't see any non-linear relationship. So, it is a VALID Assumption.

Assumption - 2 Random Sampling Since the dataset has counties information from only few regions it is not truly randomly sampled. But, we were provided with information that how the data is collected. Based on this we think it is a VALID Assumption.

Assumption -3 Multicollinearity

```
cor_model3 <- data.matrix(subset(crime, select = c("log_crmrte", "log_prbarr", "log_prbconv", "log_polpc", "log_density")))
cor(cor_model3)
```

```
##           log_crmrte  log_prbarr log_prbconv  log_polpc log_density
```

```
## log_crmrte      1.0000000 -0.43575389 -0.37249610  0.28453961  0.49364251
## log_prbarr     -0.4357539  1.00000000 -0.20235541  0.05424285 -0.35477923
## log_prbconv    -0.3724961 -0.20235541  1.00000000 -0.13863932 -0.07258589
## log_polpc       0.2845396  0.05424285 -0.13863932  1.00000000  0.02304206
## log_density     0.4936425 -0.35477923 -0.07258589  0.02304206  1.00000000
## log_pctmin80    0.3972910 -0.05122328  0.07171367 -0.16189846 -0.01881968
##               log_pctmin80
## log_crmrte      0.39729097
## log_prbarr     -0.05122328
## log_prbconv     0.07171367
## log_polpc      -0.16189846
## log_density    -0.01881968
## log_pctmin80    1.00000000
```

We are not seeing any obvious signs of multicollinearity.

```
vif(model3)
```

```
##      log(prbarr)  log(prbconv)    log(polpc)  log(density) log(pctmin80)
##      1.264585    1.111820      1.237717    1.571169      1.044778
##      log(prbpris)  log(avgsen)
##      1.318298      1.321456
```

Computing VIF also indicates that there is no multicollinearity. So, it is a VALID Assumption.

Assumption - 4 Exogeneity (Zero Conditional Mean) From the Residuals Vs Fitted Plot, the red line is very influenced by the outliers on the ends. But it is close to x-axis. It is a Most-Likely valid Assumption.

Additional Assumptions - Homoscedasticity From the same Residuals Vs Fitted plot, we see it is very scattered with extreme outliers. So, it is not easy to determine Homoscedasticity from this plot only.

```
bptest(model3)
```

```
##
## studentized Breusch-Pagan test
##
## data:  model3
## BP = 8.4026, df = 7, p-value = 0.2984
```

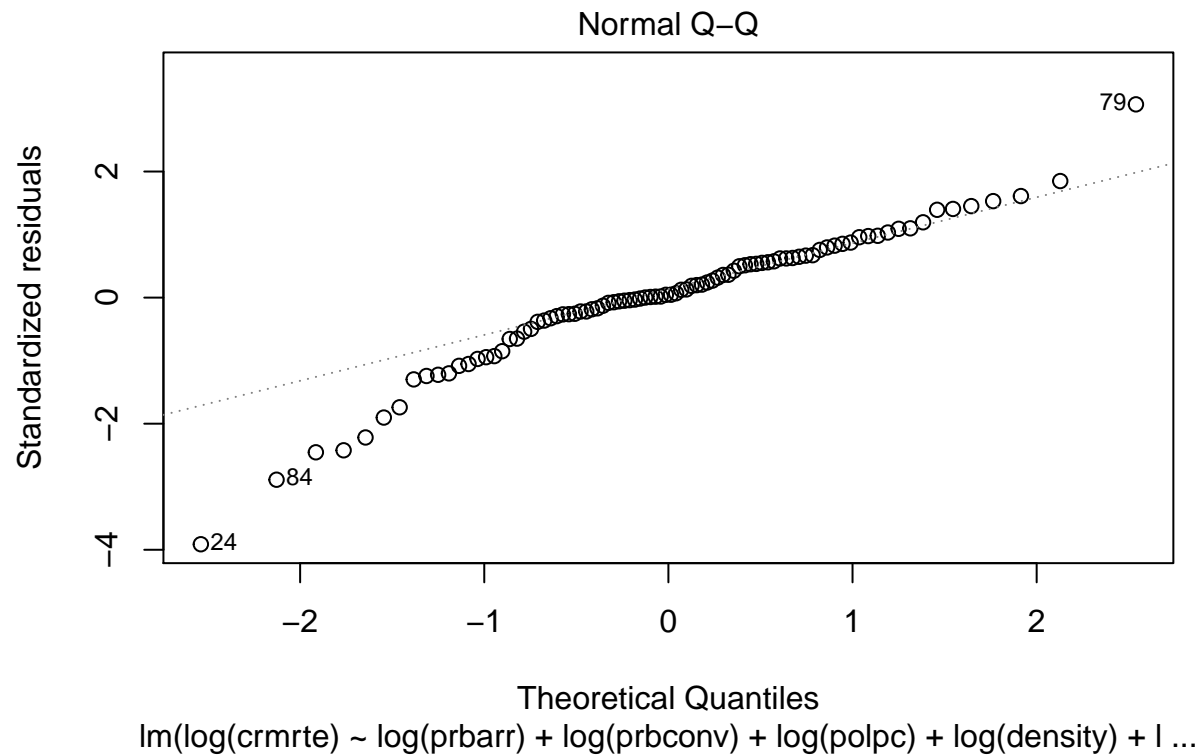
```
ncvTest(model3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.5987105, Df = 1, p = 0.43907
```

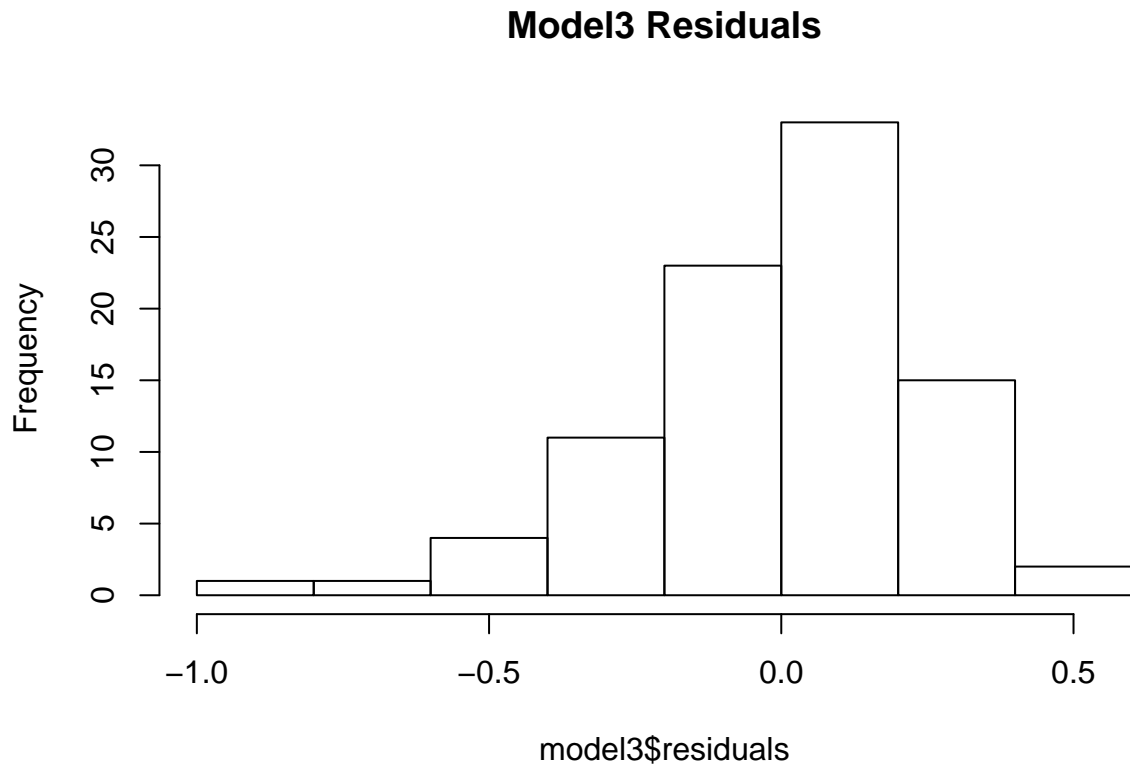
Both tests are showing small p-values showing that we fail to reject the null hypothesis. So Homoscedasticity is a most likely valid assumption here.

Normality of Residuals

```
plot(model3, which=2)
```



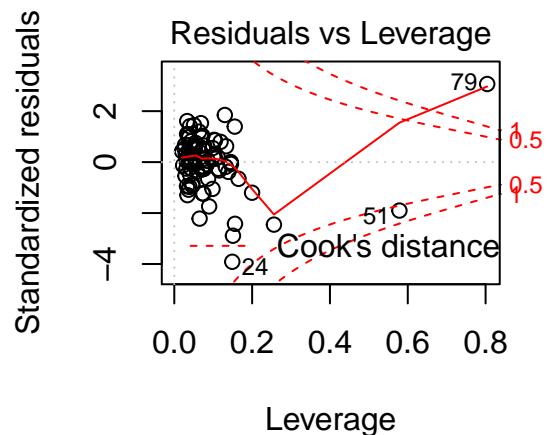
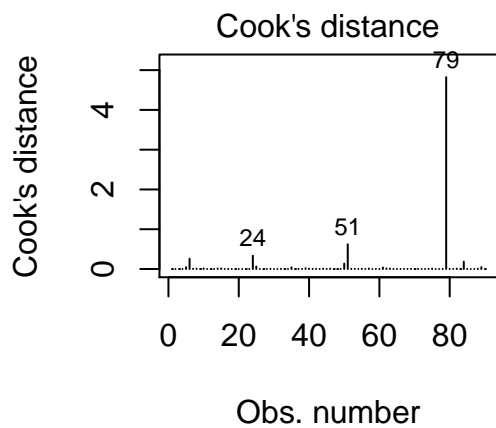
```
hist(model3$residuals, main="Model3 Residuals")
```



Other than a few outliers, the distribution is relatively normal for our given sample size. So, it is a VALID assumption.

Cook's Distance:

```
plot(model3, which=4)
plot(model3, which=5)
```



$\sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{polpc})$  ~  $\log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{polpc})$  There are some influential values [6,51,79] however cook's distance is within the bounds

```
AIC(model3)
```

```
## [1] 26.9529
```

The AIC for this model is 26.92602

```
summary(model3)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc) +
##     log(density) + log(pctmin80) + log(prbpris) + log(avgsen),
##     data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96163 -0.09114  0.01152  0.16184  0.45946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.61525     0.69999  -2.308  0.0235 *
## log(prbarr)   -0.50624     0.07884  -6.421 8.29e-09 ***
## log(prbconv)  -0.39124     0.05329  -7.342 1.38e-10 ***
## log(polpc)     0.52718     0.08395   6.279 1.54e-08 ***
## log(density)   0.16218     0.02577   6.293 1.45e-08 ***
## log(pctmin80)  0.26144     0.03009   8.687 3.01e-13 ***
## log(prbpris)  -0.31998     0.14601  -2.191  0.0313 *
## log(avgsen)   -0.23182     0.11579  -2.002  0.0486 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2664 on 82 degrees of freedom
## Multiple R-squared:  0.7828, Adjusted R-squared:  0.7643
## F-statistic: 42.22 on 7 and 82 DF,  p-value: < 2.2e-16
```

Looking at the summary of model2, 76.26% of crime rate is being explained by our model. Looking at the co-efficient values -> Polpc, Density and Pctmin80 has a positive effect on crime. While other independent variables prbarr,prbconv,prbpris,avgsen has negative effect on crime. Looking at the  $\Pr(>|t|)$  values -> crime rate fluctuates more with prbarr, prbconv, density, pctmin80 and polpc and has some impact with prbpris and avgsen

This says that crimerate decreases if more people are arrested, convicted, sentenced in prison With better policing in place more crimes would be identified. And as density increases crimerates tend to go up. Also, Minority has a positive impact on crime.

## Model4 - Using all variables Specification

The team wanted to check the robustness of all variables in the model except county name, year

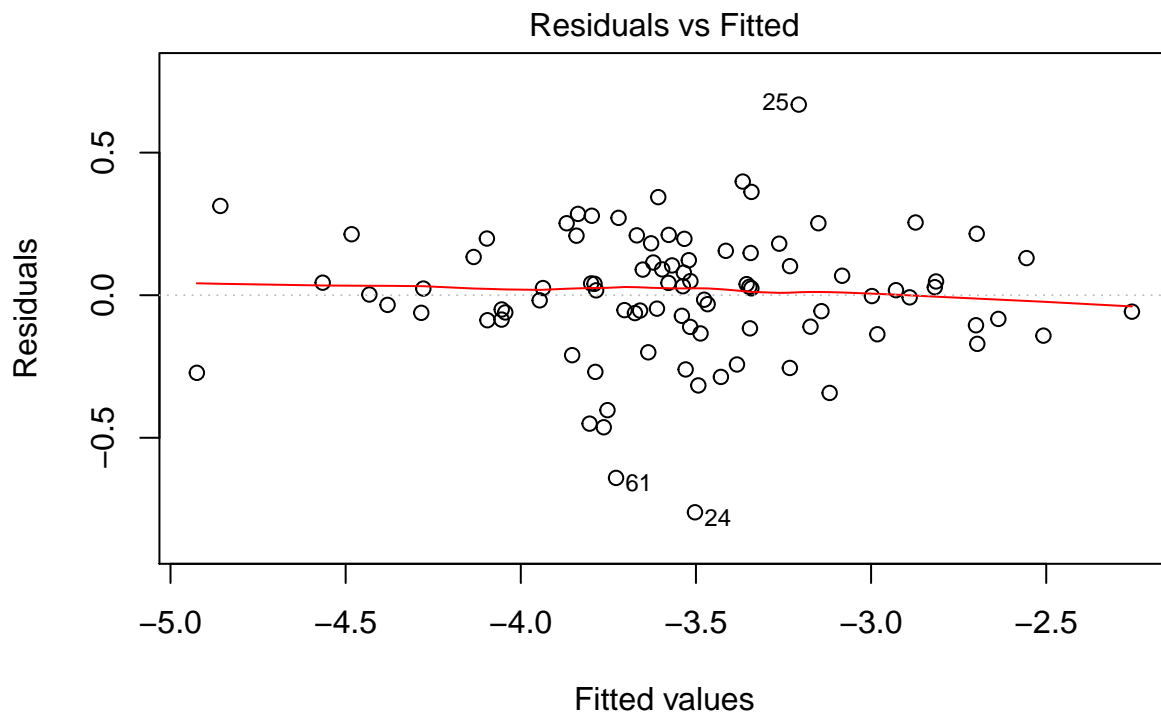
$$\begin{aligned} \log(\text{crmrate}) = & \beta_0 + \beta_1 \log(\text{prbarr}) + \beta_2 \log(\text{prbconv}) + \beta_3 \log(\text{prbpris}) + \\ & \beta_4 \log(\text{avgsen}) + \beta_5 \log(\text{polpc}) + \beta_6 \log(\text{taxpc}) + \beta_7 \log(\text{west}) + \\ & \beta_8 \log(\text{central}) + \beta_9 \log(\text{urban}) + \beta_{10} \log(\text{pctmin80}) + \beta_{11} \log(\text{wcon}) + \beta_{12} \log(\text{wtuc}) + \\ & \beta_{13} \log(\text{wtrd}) + \beta_{14} \log(\text{wfir}) + \beta_{15} \log(\text{wser}) + \beta_{16} \log(\text{wmfg}) + \beta_{17} \log(\text{wfed}) + \\ & \beta_{18} \log(\text{wsta}) + \beta_{19} \log(\text{wloc}) + \beta_{20} \log(\text{mix}) + \beta_{21} \log(\text{density}) + \beta_{22} \log(\text{pctymle}) + u \end{aligned}$$

```
model4 = lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgsen) + log(polpc) + log(de)
+ wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc + log(mix) + log(pctymle), d
```

## Checking if our Assumptions are valid in the model-

Assumption 1 - Linear model -> the specified model has the dependent variable linear with explanatory variables

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



$\text{lm}(\log(\text{crmrate}) \sim \log(\text{prbarr}) + \log(\text{prbconv}) + \log(\text{prbpris}) + \log(\text{avgsen}) + \dots$

from the Residuals vs Fitted plot, we don't see any non-linear relationship. So, it is a VALID Assumption.

Assumption - 2 Random Sampling Since the dataset has counties information from only few regions it is not truly randomly sampled. But, we were provided with information that how the data is collected. Based on this we think it is a VALID Assumption.

Assumption -3 Multicollinearity

```
cor_model4 <- data.matrix(subset(crime, select = c("log_crmrte", "log_prbarr", "log_prbconv", "log_prbpris",
"wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "wsta", "wloc", "log_mix", "log_pctymle")))
cor(cor_model4)
```

```
##          log_crmrte  log_prbarr  log_prbconv  log_prbpris
## log_crmrte      1.0000000000 -0.435753889 -0.372496097  0.069607288
## log_prbarr     -0.4357538891  1.0000000000 -0.202355412 -0.028015058
## log_prbconv    -0.3724960970 -0.202355412  1.000000000  0.010535226
## log_prbpris     0.0696072879 -0.028015058  0.010535226  1.000000000
```



```

## log_avgsen      0.0234171728  0.003832922  0.011629954 -0.123723380
## log_polpc       0.2845396122  0.054242849 -0.138639323 -0.010464075
## log_density     0.4936425080 -0.354779226 -0.072585893  0.415717578
## log_taxpc       0.3398432242 -0.111996061 -0.206229946 -0.056718897
## west            -0.4143995913  0.126566454 -0.003871865 -0.091105655
## central         0.1847192406 -0.155064409 -0.022143708  0.186252205
## urban           0.4914644548 -0.249112191 -0.218656969  0.067685886
## log_pctmin80    0.3972909650 -0.051223276  0.071713668  0.028977726
## wcon            0.3937148597 -0.213455274 -0.085329385 -0.004804471
## wtuc            0.2014649314 -0.123022775  0.009714610  0.169553410
## wtrd            0.3937923986 -0.112966615 -0.113698553  0.165621943
## wfir            0.2932426483 -0.222122040  0.074868735  0.063178022
## wser            -0.1131280124 -0.146031203  0.280195566  0.055783263
## wmfg            0.3075373091 -0.228474626  0.074463187  0.071342452
## wfed            0.5233058481 -0.235181201  0.006173931  0.144473894
## wsta            0.1697020784 -0.229737301 -0.110083307 -0.042511768
## wloc            0.2885667844 -0.125773594  0.065044636  0.111947031
## log_mix         0.0004243981  0.515211592 -0.378912114  0.030162067
## log_pctymle     0.3117540302 -0.282255476 -0.223822249 -0.024860719
##               log_avgsen  log_polpc  log_density  log_taxpc
## log_crmrte      0.023417173  0.284539612  0.493642508  0.339843224
## log_prbarr      0.003832922  0.054242849 -0.354779226 -0.111996061
## log_prbconv     0.011629954 -0.138639323 -0.072585893 -0.206229946
## log_prbpris    -0.123723380 -0.010464075  0.415717578 -0.056718897
## log_avgsen      1.000000000  0.395079641  0.159651910  0.077900054
## log_polpc       0.395079641  1.000000000  0.023042056  0.382987239
## log_density     0.159651910  0.023042056  1.000000000  0.107986921
## log_taxpc       0.077900054  0.382987239  0.107986921  1.000000000
## west            0.075694402  0.108760927 -0.214936953 -0.203869308
## central         -0.139057520  0.013934424  0.288259361  0.054660661
## urban           0.165316448  0.281877660  0.392726476  0.397859371
## log_pctmin80    -0.164025122 -0.161898455 -0.018819680  0.121027860
## wcon            0.010673315  0.103258116  0.370538862  0.270713509
## wtuc            0.247179988  0.218324501  0.394327697  0.175287125
## wtrd            0.117846349  0.193184765  0.421458150  0.191515512
## wfir            0.188502755  0.276361970  0.401743923  0.141022307
## wser            -0.194722056 -0.002394111  0.027641940  0.110201443
## wmfg            0.120525477  0.351209503  0.384974937  0.294237001
## wfed            0.174190820  0.313603184  0.525463707  0.101661299
## wsta            0.129985166  0.099681381  0.004038989 -0.020707347
## wloc            0.129240439  0.392843237  0.302713009  0.219869186
## log_mix         -0.130711469  0.044620962 -0.311452390 -0.005188296
## log_pctymle     0.103327744  0.176404361  0.175156111 -0.073609432
##               west      central      urban  log_pctmin80
## log_crmrte      -0.414399591  0.18471924  0.49146445  0.3972909650
## log_prbarr       0.126566454 -0.15506441 -0.24911219 -0.0512232763
## log_prbconv     -0.003871865 -0.02214371 -0.21865697  0.0717136675
## log_prbpris     -0.091105655  0.18625220  0.06768589  0.0289777255
## log_avgsen       0.075694402 -0.13905752  0.16531645 -0.1640251222
## log_polpc        0.108760927  0.01393442  0.28187766 -0.1618984552
## log_density     -0.214936953  0.28825936  0.39272648 -0.0188196801
## log_taxpc       -0.203869308  0.05466066  0.39785937  0.1210278600
## west            1.000000000 -0.38987611 -0.08681219 -0.7594795676
## central         -0.389876111  1.00000000  0.15927023  0.1081931293

```

## urban	-0.086812195	0.15927023	1.00000000	0.0979915917
## log_pctmin80	-0.759479568	0.10819313	0.09799159	1.0000000000
## wcon	-0.185990737	0.39786886	0.31926691	0.0407306751
## wtuc	0.083384660	0.18855844	0.22632785	-0.1520807592
## wtrd	-0.165704954	0.38668510	0.43172844	0.0517974614
## wfir	-0.039343046	0.29060049	0.40171167	0.0006203079
## wser	-0.059859894	0.19261249	0.05589097	0.1238486289
## wmfg	-0.003744894	0.17368746	0.40362299	-0.0698942458
## wfed	-0.178350352	0.34923553	0.42602595	0.1331392219
## wsta	-0.076443685	0.08527707	0.30194045	0.0848626952
## wloc	-0.115414239	0.33323127	0.33835635	0.0030084019
## log_mix	-0.121013299	-0.06834972	-0.03945711	0.2297713219
## log_pctymle	-0.055597645	-0.08131536	0.12927006	0.0080481047
##	wcon	wtuc	wtrd	wfir
## log_crmrte	0.393714860	0.20146493	0.393792399	0.2932426483
## log_prbarr	-0.213455274	-0.12302277	-0.112966615	-0.2221220397
## log_prbconv	-0.085329385	0.00971461	-0.113698553	0.0748687346
## log_prbpris	-0.004804471	0.16955341	0.165621943	0.0631780215
## log_avgsen	0.010673315	0.24717999	0.117846349	0.1885027548
## log_polpc	0.103258116	0.21832450	0.193184765	0.2763619697
## log_density	0.370538862	0.39432770	0.421458150	0.4017439229
## log_taxpc	0.270713509	0.17528713	0.191515512	0.1410223069
## west	-0.185990737	0.08338466	-0.165704954	-0.0393430459
## central	0.397868862	0.18855844	0.386685102	0.2906004856
## urban	0.319266914	0.22632785	0.431728442	0.4017116651
## log_pctmin80	0.040730675	-0.15208076	0.051797461	0.0006203079
## wcon	1.000000000	0.40937889	0.564058565	0.4889377372
## wtuc	0.409378888	1.00000000	0.351683658	0.3276195550
## wtrd	0.564058565	0.35168366	1.00000000	0.6681545248
## wfir	0.488937737	0.32761955	0.668154525	1.0000000000
## wser	-0.013164375	-0.01924404	-0.020741268	0.0137161401
## wmfg	0.347392816	0.46892658	0.371487416	0.4975834079
## wfed	0.506663944	0.39866915	0.640521866	0.6231788207
## wsta	-0.018856088	-0.15340397	0.007267295	0.2407000594
## wloc	0.517041291	0.33301977	0.581463886	0.5544356350
## log_mix	-0.139217617	-0.22240855	-0.036867273	-0.1566529165
## log_pctymle	0.008266563	-0.08056246	-0.085205330	0.0545098142
##	wser	wmfg	wfed	wsta
## log_crmrte	-0.113128012	0.307537309	0.523305848	0.169702078
## log_prbarr	-0.146031203	-0.228474626	-0.235181201	-0.229737301
## log_prbconv	0.280195566	0.074463187	0.006173931	-0.110083307
## log_prbpris	0.055783263	0.071342452	0.144473894	-0.042511768
## log_avgsen	-0.194722056	0.120525477	0.174190820	0.129985166
## log_polpc	-0.002394111	0.351209503	0.313603184	0.099681381
## log_density	0.027641940	0.384974937	0.525463707	0.004038989
## log_taxpc	0.110201443	0.294237001	0.101661299	-0.020707347
## west	-0.059859894	-0.003744894	-0.178350352	-0.076443685
## central	0.192612486	0.173687456	0.349235529	0.085277071
## urban	0.055890972	0.403622994	0.426025955	0.301940450
## log_pctmin80	0.123848629	-0.069894246	0.133139222	0.084862695
## wcon	-0.013164375	0.347392816	0.506663944	-0.018856088
## wtuc	-0.019244042	0.468926579	0.398669151	-0.153403974
## wtrd	-0.020741268	0.371487416	0.640521866	0.007267295
## wfir	0.013716140	0.497583408	0.623178821	0.240700059

```
## wser      1.000000000  0.008986754  0.020674709  0.037471156
## wmfg      0.008986754  1.000000000  0.518230474  0.052336590
## wfed      0.020674709  0.518230474  1.000000000  0.188250660
## wsta      0.037471156  0.052336590  0.188250660  1.000000000
## wloc      0.076971338  0.450453501  0.519413570  0.164641269
## log_mix   -0.217281012 -0.259530474 -0.216849576 -0.130379017
## log_pctymle -0.052120438  0.074718130  0.003926539  0.285524343
##           wloc      log_mix  log_pctymle
## log_crmrte  0.288566784  0.0004243981  0.311754030
## log_prbarr -0.125773594  0.5152115919 -0.282255476
## log_prbconv 0.065044636 -0.3789121137 -0.223822249
## log_prbpris 0.111947031  0.0301620671 -0.024860719
## log_avgsen  0.129240439 -0.1307114688  0.103327744
## log_polpc   0.392843237  0.0446209622  0.176404361
## log_density 0.302713009 -0.3114523900  0.175156111
## log_taxpc   0.219869186 -0.0051882961 -0.073609432
## west        -0.115414239 -0.1210132985 -0.055597645
## central     0.333231267 -0.0683497164 -0.081315357
## urban       0.338356350 -0.0394571118  0.129270058
## log_pctmin80 0.003008402  0.2297713219  0.008048105
## wcon        0.517041291 -0.1392176166  0.008266563
## wtuc        0.333019765 -0.2224085481 -0.080562458
## wtrd        0.581463886 -0.0368672730 -0.085205330
## wfir        0.554435635 -0.1566529165  0.054509814
## wser        0.076971338 -0.2172810125 -0.052120438
## wmfg        0.450453501 -0.2595304741  0.074718130
## wfed        0.519413570 -0.2168495756  0.003926539
## wsta        0.164641269 -0.1303790170  0.285524343
## wloc        1.000000000 -0.1886121880  0.020230541
## log_mix     -0.188612188  1.0000000000 -0.146667642
## log_pctymle 0.020230541 -0.1466676416  1.000000000
```

We are not seeing any obvious signs of multicollinearity.

```
vif(model4)
```

```
## log(prbarr) log(prbconv) log(prbpris) log(avgsen) log(polpc)
## 1.812595 2.069457 1.536031 1.683955 2.538936
## log(density) log(taxpc) west central urban
## 3.126722 2.087337 4.293576 1.921333 2.073034
## log(pctmin80) wcon wtuc wtrd wfir
## 3.798204 2.214858 1.798198 3.233880 2.841757
## wser wmfg wfed wsta wloc
## 1.367321 2.047091 3.344744 1.797655 2.276723
## log(mix) log(pctymle)
## 2.251470 1.708311
```

Computing VIF also indicates that there is no multicollinearity. So, it is a VALID Assumption.

Assumption - 4 Exogeneity (Zero Conditional Mean) From the Residuals Vs Fitted Plot, the red line is very influenced by the outliers on the ends. But it is close to x-axis. It is a Most-Likely valid Assumption.

Additional Assumptions - Homoscedasticity From the same Residuals Vs Fitted plot, we see it is very scattered with extreme outliers. So, it is not easy to determine Homoscedasticity from this plot only.

```
bptest(model4)
```

```
##
## studentized Breusch-Pagan test
##
## data: model4
## BP = 50.118, df = 22, p-value = 0.0005653
```

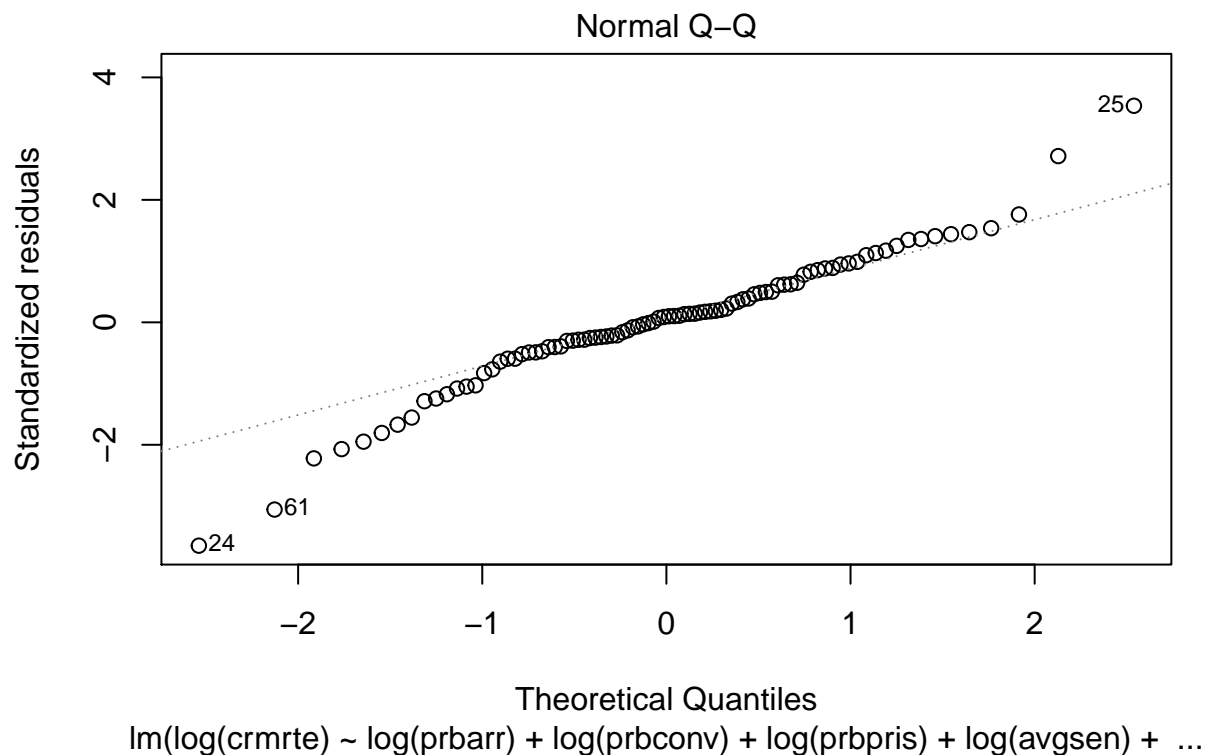
```
ncvTest(model4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1546225, Df = 1, p = 0.69416
```

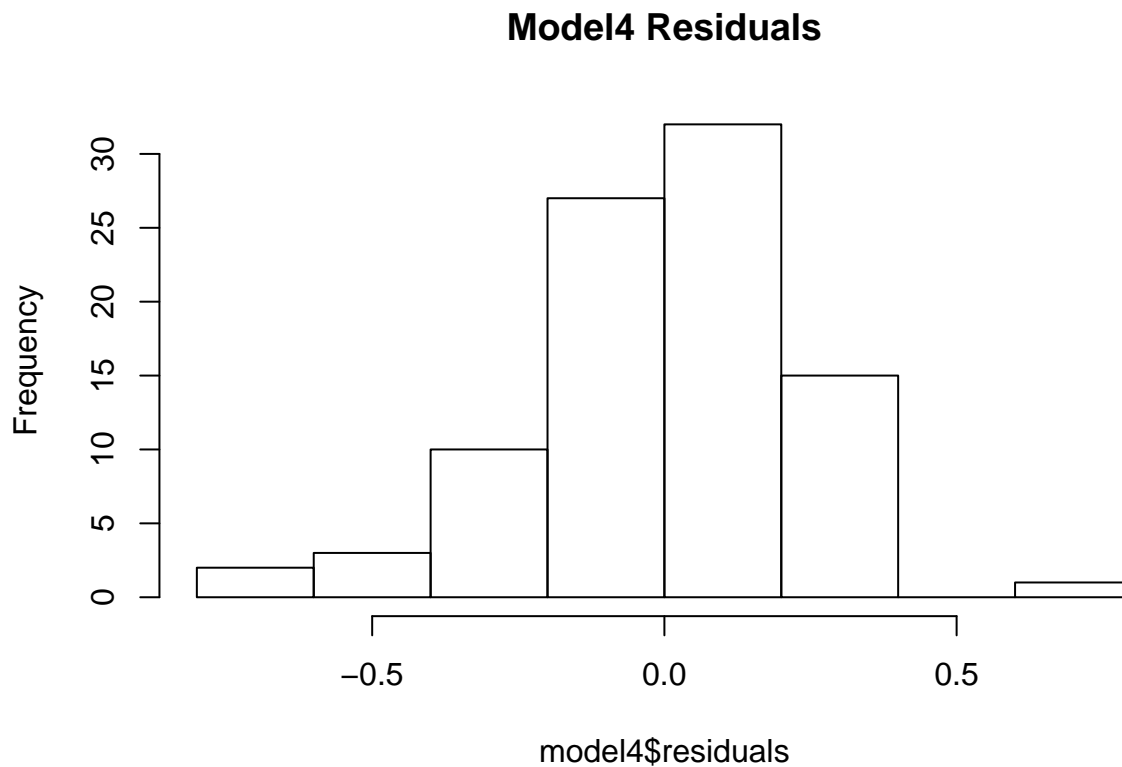
Both tests are showing small p-values showing that we fail to accept the hypothesis. So Homoscedasticity is a most likely valid assumption here.

Normality of Residuals

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



```
hist(model4$residuals, main="Model4 Residuals")
```



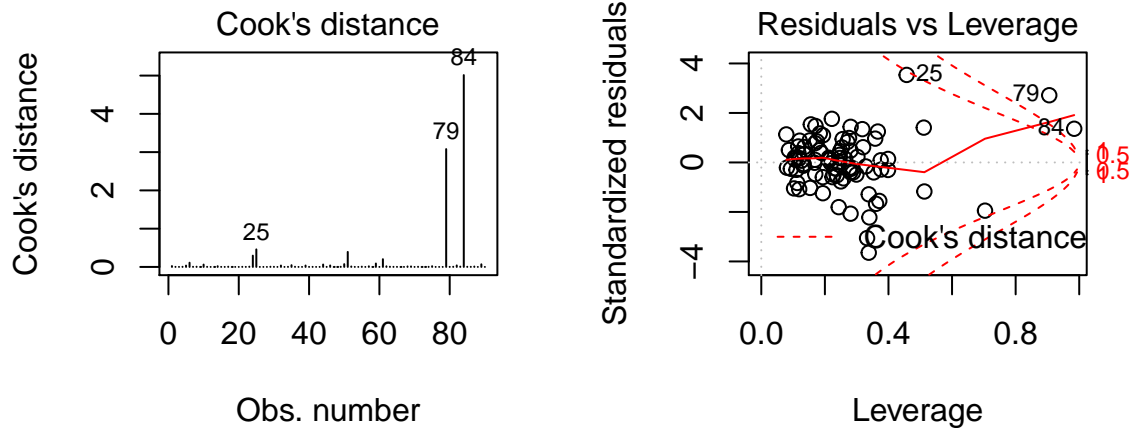
Other than a few outliers, the distribution is relatively normal for our given sample size. So, it is a VALID assumption.

Cook's Distance:

```
plot(model4, which=4)
plot(model4, which=5)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



~ log(prbarr) + log(prbconv) + log(prbpris) ~ log(prbarr) + log(prbconv) + log(prbpris) There are some influential values [25,79,84] however cook's distance is within the bounds

```
AIC(model4)
```

```
## [1] 31.94569
```

The AIC for this model is 30.78825

```
summary(model4)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##   log(avgsen) + log(polpc) + log(density) + log(taxpc) + west +
##   central + urban + log(pctmin80) + wcon + wtuc + wtrd + wfir +
##   wser + wmfg + wfed + wsta + wloc + log(mix) + log(pctymle),
##   data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76126 -0.10102  0.02046  0.13307  0.66840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.984e+00  1.285e+00  -1.544  0.127414
## log(prbarr)   -5.328e-01  9.088e-02  -5.863  1.53e-07 ***
## log(prbconv)  -2.978e-01  7.000e-02  -4.255  6.66e-05 ***
## log(prbpris)  -3.011e-01  1.517e-01  -1.984  0.051311 .
## log(avgsen)   -3.079e-01  1.259e-01  -2.447  0.017053 *
## log(polpc)     4.804e-01  1.158e-01   4.150  9.60e-05 ***
## log(density)   1.388e-01  3.501e-02   3.964  0.000181 ***
## log(taxpc)     1.754e-02  1.482e-01   0.118  0.906160
## westTRUE       5.178e-02  1.304e-01   0.397  0.692526
## centralTRUE    -7.671e-02  7.730e-02  -0.992  0.324588
## urbanTRUE       7.099e-02  1.368e-01   0.519  0.605513
## log(pctmin80)  2.552e-01  5.525e-02   4.620  1.79e-05 ***
## wcon           7.324e-04  8.474e-04   0.864  0.390539
## wtuc           1.443e-04  4.713e-04   0.306  0.760457
```

```
## wtrd          1.480e-03  1.444e-03   1.025 0.308892
## wfir          -1.416e-03  8.488e-04  -1.668 0.100032
## wser          -3.798e-04  1.533e-04  -2.477 0.015770 *
## wmfg          -1.344e-04  4.409e-04  -0.305 0.761500
## wfed          1.483e-03  8.295e-04   1.788 0.078245 .
## wsta          -2.507e-05  8.421e-04  -0.030 0.976337
## wloc          6.192e-05  1.458e-03   0.042 0.966262
## log(mix)       6.205e-02  7.446e-02   0.833 0.407597
## log(pctymle)   1.013e-01  1.791e-01   0.565 0.573668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2565 on 67 degrees of freedom
## Multiple R-squared:  0.8355, Adjusted R-squared:  0.7815
## F-statistic: 15.47 on 22 and 67 DF,  p-value: < 2.2e-16
```

Looking at the summary of model2, 78.26% of crime rate is being explained by our model. Looking at the co-efficient values -> Polpc, Density and Pctmin80 has a positive effect on crime. While other independent variables prbarr,prbconv,prbpris,avgsen has negative effect on crime. Looking at the  $\Pr(>|t|)$  values -> crime rate fluctuates more with prbarr, prbconv, density, pctmin80 and polpc and has some impact with prbpris and avgsen

This says that crimerate decreases if more people are arrested, convicted, sentenced in prison With better policing in place more crimes would be identified. And as density increases crimerates tend to go up. Also, Minority has a positive impact on crime.

## Omitted Variables

In order to make valid policy recommendations, we need confidence that our estimated coefficients for policy-relevant variables are unbiased, statistically significant, and practically significant. Statistical software makes it quite easy to determine if there is a relationship between a given variable and the dependent variable that is statistically significantly different from zero - an area of analysis that we will expand upon in follow-ups to this piece. Practical significance of our estimates requires just one extra step to interpret the meaning of the estimate for each variable under consideration. Accounting for elements which could bias our estimates is more difficult and, to some degree, not a solvable problem.

We only have observational data available. Moreover, we are not able to design or even infer experiments for our data generating process. As such, we are left to reason about counterfactuals, rather than conduct experiments to verify the implications of our model. Additionally, we have a flawed data collection process, which we also have no ability to correct for. Our desired population variables are by-in-large not included in the dataset we were provided. Some of these desired variables are practically or ethically unobservable. Others were operationalized in a flawed manner, with a negative impact on our ability to model relationships with a causal interpretation. We address some of these issues here.

Our ideal model of the causes of the crime rate would be something like:

$$\begin{aligned} \text{crime\_rate} = & \beta_0 + \beta_1 \text{crtty\_punish} + \beta_2 \text{svrty\_punish} + \beta_3 \text{wealth\_inequality} + \\ & \beta_4 \text{educ} + \beta_5 \text{social\_cohesion} + \beta_6 \text{weapon\_availability} + \beta_7 \text{real\_wage} + \\ & \beta_8 \text{low\_skill\_unemployment\_rate} + \beta_9 \text{age\_15\_to\_30\_proportion\_population} + \\ & \beta_{10} \text{gender\_imbalance} + \dots + \text{error} \end{aligned}$$

Unfortunately, we are unable to observe virtually all of these concepts.

Some concepts have been operationalized in our dataset. For example, certainty of punishment has been operationalized through three variables: 1) the percent of the population which are police, 2) the proportion of arrests to crimes, and 3) the proportion of convictions to arrest. This is among the most effective operationalizations in this dataset. Severity of punishment is also operationalized through 1) the proportion of convictions that result in a prison sentence and 2) the average length of a prison sentence. Nominal wages are operationalized in the dataset with average wages for certain industry groupings. None of wealth inequality within a given observation, education, social cohesion, weapon availability, cost of living, or the low skill unemployment rate are operationalized within this dataset.

Moreover, certain variables which are included in our dataset are likely correlated with many of our desired variables, but actually measure something distinct - introducing the possibility for model estimates based on those variables to be biased and thus misleading. For example, the `pctmin80` variable measures the percent of a county that was minority in 1980 - 7 years prior to our other observations. Setting the time divergence aside and extrapolating from national trends in the U.S. in the 1980s, the percentage of a county which is minority is likely negatively correlated with education. It may also exhibit a parabolic relation with wealth inequality and social cohesion. If we were to include `pctmin80` in our regression, we would expect the model estimate to be biased as we have not adjusted for the impacts of education, wealth inequality, or social cohesion. Examining the impact of education alone on the estimator for `pctmin80` - as education was likely negatively correlated with `pctmin80`, and we expect educated to be negatively related to the crime rate, the model's estimate of the impact of the percent of a county which was minority in 1980 would be upwardly biased. In other words, the estimator for `pctmin80` in the underspecified model would imply a much larger relationship between `pctmin80` and crime rate than actually exists.

Similarly, our dataset contains a variable `density` which is likely correlated with two of our desired but unobserved explanatory variables: social cohesion and wealth inequality. In practice, in the U.S. in the 1980s, we would expect social cohesion to be negatively correlated with density, while wealth inequality would be positively correlated with density. We expect the beta for social cohesion to crime rate to be negative, while the beta for wealth inequality to crime rate is expected to be positive. The impact of both of these omitted



variables is that the model's estimate for density is likely upwardly biased. As with `pctmin80`, the model would again overestimate the impact of density on crime rate.

Our ability to interpret the variable `polpc` in our dataset is also compromised by omitted variable bias. While we understand the idea that increased police presence should increase the certainty of punishment (more likely to be detected and more likely to be caught) *ceteris paribus*, in our current dataset, we do not have the ability to use `polpc` in this way. We are unable to observe the counterfactual of the same location with the same characteristics at the same point in time having more or less police. Rather, the variable in our dataset is the current level of police as a percent of the population. Given that we expect local governments to respond to increased crime by highering more police, our model is more likely to reflect that higher crime rate locations also have higher police concentrations. Given an alternate work environment where we could retrieve more data, we might think about attempting to compensate for this by locating police concentration and crime rate statistics for previous years, then using them to create variables for the percentage point change in police concentration, which we could use to explain a newly created variable for the percentage point change in crime rate for a given location. However, in their current single point in time forms, our model is likely to estimate the relationship between police percentage and crime rate as positive, thus providing a misleading estimate for the relationship we would actually like to observe.

Finally, our dataset contains several variables with nominal wages for certain industries. Including these in our model is likely to be somewhat misleading, producing biased estimators because these measures are not adjusted for cost of living. Said in other terms, each of the nominal wage indicators is likely positively correlated with our desired explanatory variable - real wages. Conceptually, we expect the relationship between real wages and crime rate to be negative, while the relationship between real wages and nominal wages is positive. As such our model's estimator for wages is likely to understate the impact of wages on crime rate. As such, these nominal wage variables are an imperfect proxy for the desired variable real wages

## Conclusion

We examined several models of crime rate and found a directionally consistent, statistically significant negative relationship for the probability of arrest and the probability of conviction on crime rate. As such, policies adopted should focus on increasing the certainty of punishment for committing crimes. One such policy could focus on improving information flow from local communities to police and judicial officials. A good model to build off of is community policing, where police focus on developing ties to the local community to build trust and thereby promote flow of needed information.

That said, our ability to draw policy prescriptions from our models is limited due to notable omitted variable bias, which leads our model's estimators to be biased. These omitted variable biases are not possible to overcome while limited to the current data collection process. Should more work requiring causal inference be desired on these relationships in the future, we would seek input into the data collecting process in order to correct for some of our omitted variable biases.

## Appendix A: Codebook

Table 1: Crime Data Codebook

Variable	Label	Notes
county	county identifier	
year	1987	
crmrte	crimes committed per person	
prbarr	'probability' of arrest	County 115 has a value of 1.09, which is not a possible probability. There are 10 observations greater than 1, which is not a possible probability.
prbconv	'probability' of conviction	
prbpris	'probability' of prison sentence	
avgsen	avg. sentence, days	
polpc	police per capita	
density	people per sq. mile	
taxpc	tax revenue per capita	
west	=1 if in western N.C.	
central	=1 if in central N.C.	
urban	=1 if in SMSA	
pctmin80	perc. Minority, 1980	
wcon	weekly wage, construction	
wtuc	wkly wge, trns, util, commun	
wtrd	wkly wge whlesle, retail, trade	
wfir	wkly wge, fin, ins, real est	
wser	wkly wge, service industry	
wmfg	wkly wge, manufacturing	
wfed	wkly wge fed employees	
wsta	wkly wge state employees	
wloc	wkly wge local gov emps	
mix	offense mix: face-to-face/other	
pctymle	percent young male	