Lab 3: Reducing Crime (DRAFT: Stage 1)

N. Akkineni, K. Hanna, A. Thorp

November 27, 2018

Contents

Introduction

Our team has been hired to provide research for a local political campaign, which would like to understand the determinants of crime rates and to provide policy suggestions that are applicable to local government. We examine the provided dataset to determine if a model with causal interpretation is feasible. After examining the data, we detail four regression models and find that estimators related to variables used to operationalize the concept of certainty of punishment are directionally consistent and statistically siginificant. From this we draw a limited policy recommendation to adopt a model of community policing to improve trust and information flow to law enforcement. However, our policy recommendations are limited because omitted variable bias confounds our estimators. Should local officials desire more robust conclusions, we recommend involving data scientists in the data collection process to improve our ability to draw causal inference from our modeling process and thus be able to make robust policy recommendations.

```
library(knitr)
library(kableExtra)
suppressMessages(library(car))
suppressMessages(library(stargazer))
suppressMessages(library(lmtest))
suppressMessages(library(corrplot))
library(sandwich)

crime <- read.csv('crime_v2.csv',header = TRUE, sep = ",")
```

```r
codebook <- read.csv('codebook.csv')

crime$county = as.factor(crime$county)
crime$west = as.logical(crime$west)
crime$central = as.logical(crime$central)
crime$urban = as.logical(crime$urban)

# Create a variable for counties not in west or central.
crime$east <- !(crime$west | crime$central)

# Average of all weekly wage variables.
crime$avgwage = (crime$wcon + crime$wtuc + crime$wtrd + crime$wfir +
                   crime$wser + crime$wmfg + crime$wfed + crime$wsta +
                   crime$wloc)/9
```

Exploratory Data Analysis

Data Summary

We were provided with a dataset that includes crime statistics from the North Carolina Department of Corrections' prison and probation files, demographic statistics taken from the decennial census, police data derived from FBI police agency data and wage data from the North Carolina Employment Security commission. In all we were provided with 25 variables and 90 counties.

Some of the values in this dataset were calculated from other datasets and we found some characteristics in the dataset which may bring its veracity in to question and we have addressed those below and in our analyses.

Additionally, we added two new variables based on data provided to aid in our analysis. 'avgwage' is the mean of all the weekly wages included in the dataset, and 'east' is all the counties that are not in central or west. It is possible this is not a safe assumption, however that distinction will not become relevant in this analysis.

Table 1: Crime Data Codebook

Variable	Label
county	county identifier
year	1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
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

Table 1: Crime Data Codebook *(continued)*

Variable	Label
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
east	Created: observations not in west or central
avgwage	Created: mean of all wages

Data Clean Up

Removing Null Rows

The dataset contained an apostrophe 6 rows after the data which caused the csv reader to create 6 invalid rows. We removed these rows as they contain no data.

```
# Delete the 6 empty observations at the end, including the row with the apostrophe.
crime <- crime[!is.na(crime$county) & !is.na(crime$crmrte), ]
```

Removing Duplicate County

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

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

Convert prbconv to numeric

```
# Convert prbconv to numeric
crime$prbconv = as.numeric(as.character(crime$prbconv))
```

Concerns about the data

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 either be a labeling error, as it is a ratio used to approximate a probability, or erroneous, as we'll explain later, there are several concerns about the observations for county 115.

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 and persons can be convicted in absentia. Similar to the prbarr, we beleive this to be a labeling error, as it is a ratio used to approximate a probability, rather than a true probability.

Omitted Counties Creating Bias

The dataset contained 90 counties, and there are 100 couties in North Carolina. The observation id labeled 'county' in our data set appears to contain FIPS codes, if this assumption is correct the following are the missing counties: Camden County, Carteret County, Clay County, Gates County, Graham County, Iredell County, Jones County, Mitchell County, Tyrrell County and Yancey County

These missing counties will introduce a slight clustering bias into our analyses at a minimum, and possibly a more significant bias if they were ommitted based on specific criteria whether deliberate or not.

Mislabeled Region County 71

County 71 is both part of the West region and Central region. It's possible to straddle the border however, we'd expect more instances if it was done in this manner, thus we expect one of these to be erroneous and should belong in only one region. This is the only case where this occurs.

Suspicious values for county 115

There are several causes for concern for county 115. First the percentage of police per person is 0.009, the highest value in the dataset and twice that of the second highest value for that variable. It also has the lowest crime rate in the dataset, a variable we show later is actually positively correlated with police per person. The probability of arrest is greater than 1 as we mentioned above, the only variable in the dataset greater than 1. Lastly, the values for probability of conviction, probability of prison and crime mix are in the tenths, 1.5, 0.5 and 0.1 respectively. Typical values for these variables are 6, 7 and 8 decimal places, making 3 round values highly improbable.

A small number of these issues is cause for concern. This many issues suggests the observations for county 115 are erroneous.

Extreme Outlier County 185 Service Industry Wage

The value for the mean weekly wage for the service industry is 2177, nearly 8 times greater than the mean, and 5.5 times greater than the second highest value for that variable. It is likely erroneous.

Univariate Analysis

Key Variable

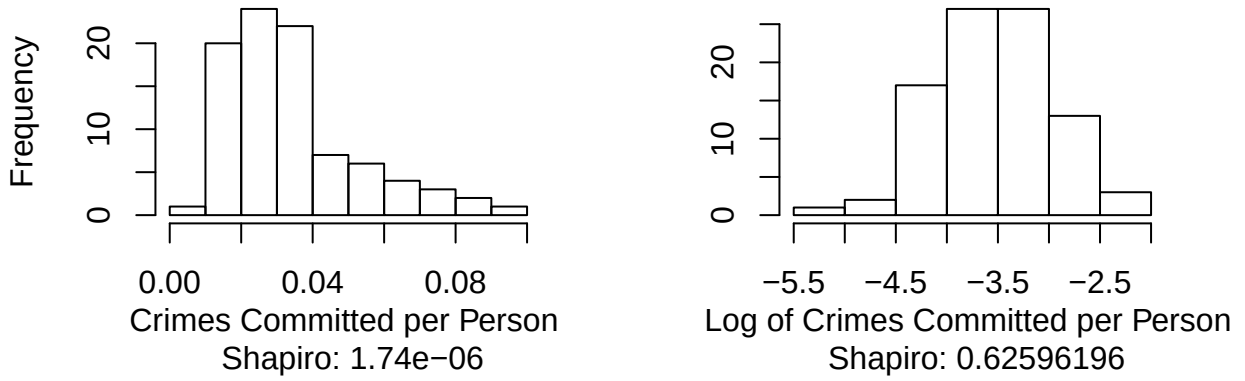
Crimes Committed Per Person

Campaign Significance: The political campaign which hired us is interested in policy prescriptions derived from causal analysis of crime rates. This is the key variable our models will attempt to explain.

```
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 = '')
}
```

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



Crimes committed per capita has a fairly strong 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.

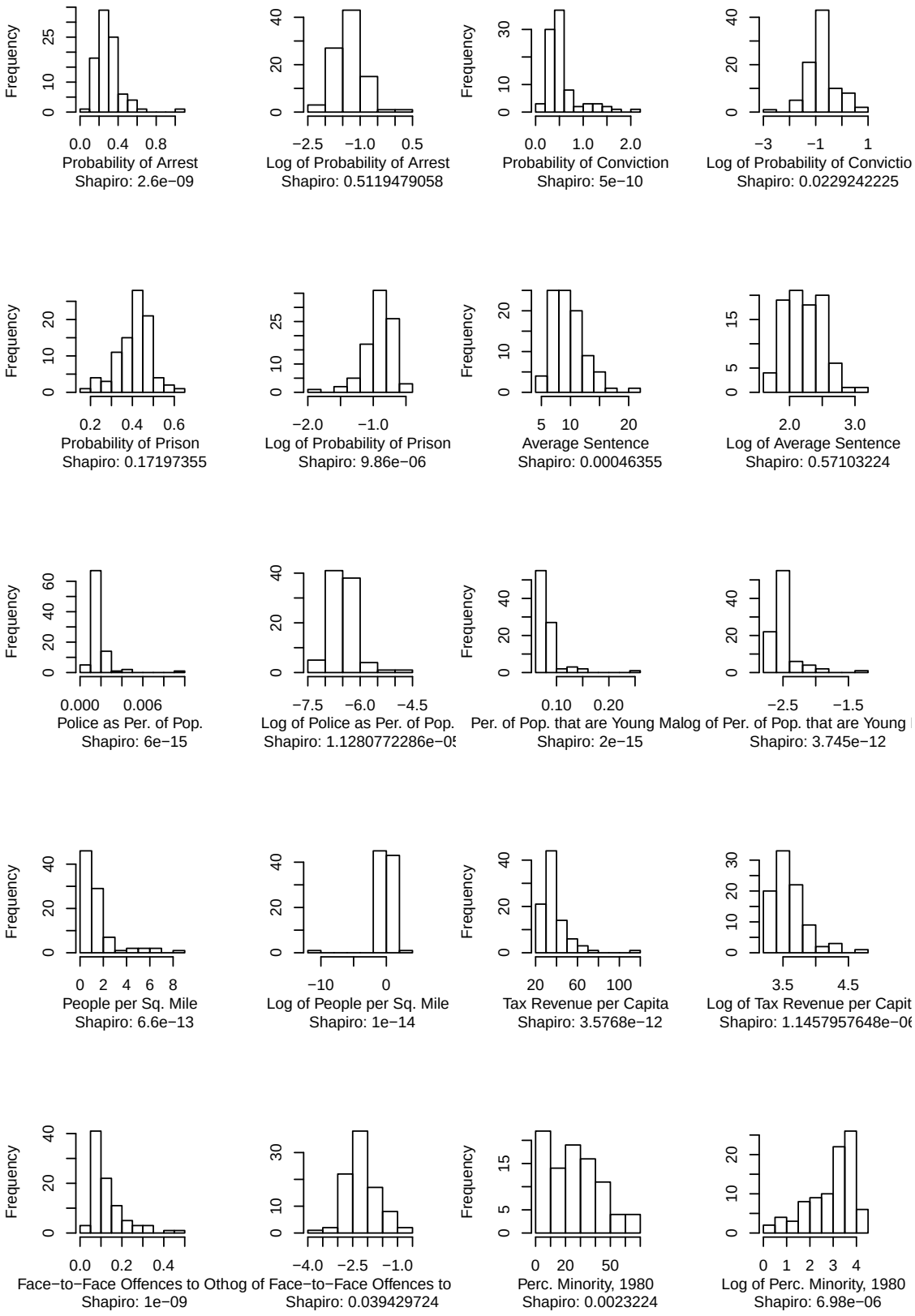
```
crime$log_crmrte <- log(crime$crmrte)
crmrte.outliers = boxplot(crime$log_crmrte, plot = FALSE)$out
```

Crime rate has 1 outliers, none that are extreme or causing concern.

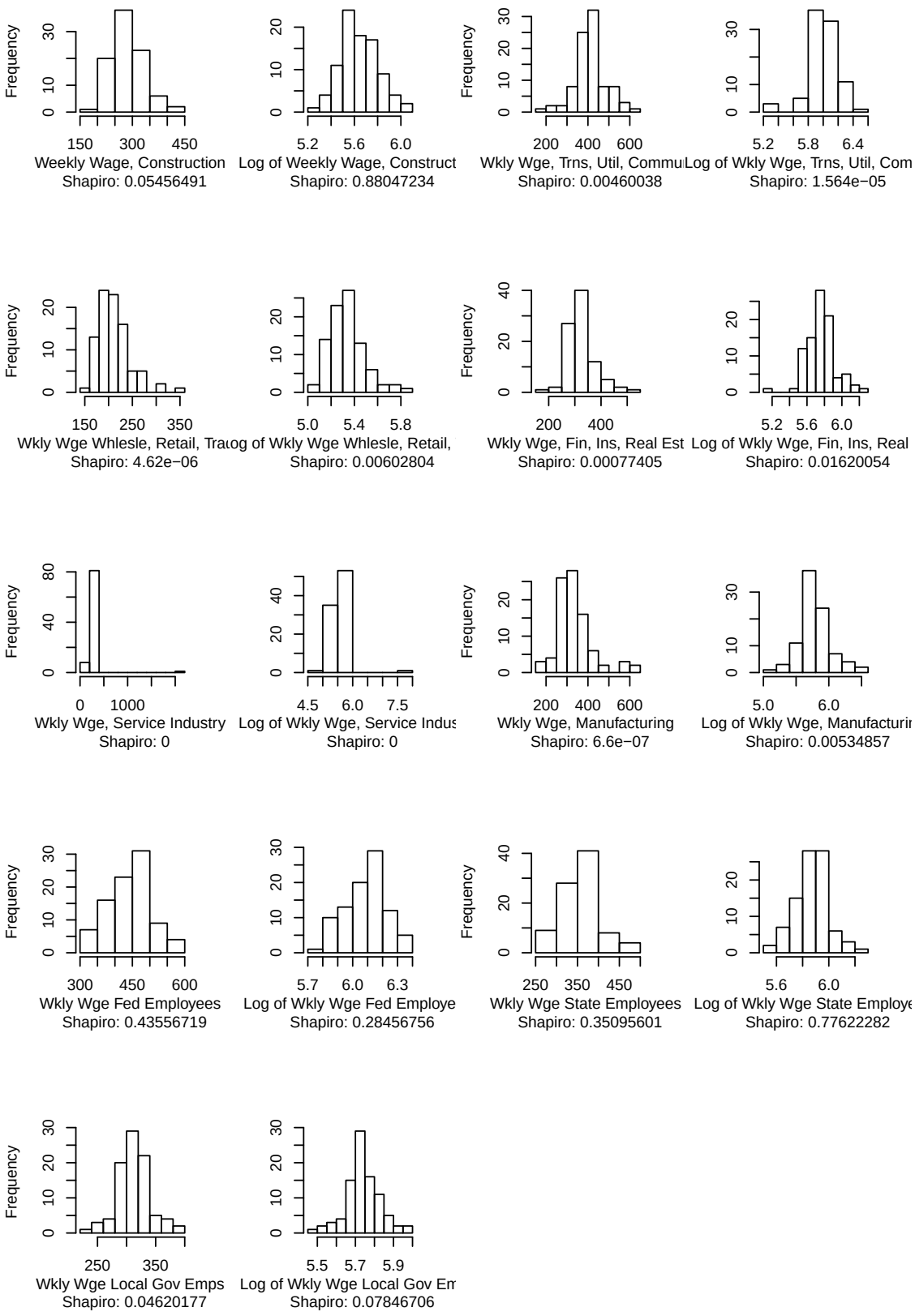
Explanatory Variables

Diagrams of 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')
```


Frequency
Probability of Arrest
Shapiro: 2.6e−09
Log of Probability of Arrest
Shapiro: 0.5119479058
Probability of Conviction
Shapiro: 5e−10
Log of Probability of Convictio
Shapiro: 0.0229242225
Probability of Prison
Shapiro: 0.17197355
Log of Probability of Prison
Shapiro: 9.86e−06
Average Sentence
Shapiro: 0.00046355
Log of Average Sentence
Shapiro: 0.57103224
Police as Per. of Pop.
Shapiro: 6e−15
Log of Police as Per. of Pop.
Shapiro: 1.1280772286e−0
Per. of Pop. that are Young Ma
Shapiro: 2e−15
Log of Per. of Pop. that are Young
Shapiro: 3.745e−12
People per Sq. Mile
Shapiro: 6.6e−13
Log of People per Sq. Mile
Shapiro: 1e−14
Tax Revenue per Capita
Shapiro: 3.5768e−12
Log of Tax Revenue per Capit
Shapiro: 1.1457957648e−0
Face−to−Face Offences to Oth
Shapiro: 1e−09
og of Face−to−Face Offences to
Shapiro: 0.039429724
Perc. Minority, 1980
Shapiro: 0.0023224
Log of Perc. Minority, 1980
Shapiro: 6.98e−06

```r
quick_uni_analysis(crime$wcon, 'weekly wage, construction')
quick_uni_analysis(crime$wtuc, 'wkly wge, trns, util, commun')
quick_uni_analysis(crime$wtrd, 'wkly wge whlesle, retail, trade')
quick_uni_analysis(crime$wfir, 'wkly wge, fin, ins, real est')
quick_uni_analysis(crime$wser, 'wkly wge, service industry')
quick_uni_analysis(crime$wmfg, 'wkly wge, manufacturing')
quick_uni_analysis(crime$wfed, 'wkly wge fed employees')
quick_uni_analysis(crime$wsta, 'wkly wge state employees')
quick_uni_analysis(crime$wloc, 'wkly wge local gov emps')

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


Frequency
Weekly Wage, Construction
Shapiro: 0.05456491
Log of Weekly Wage, Construct
Shapiro: 0.88047234
Wkly Wge, Trns, Util, Commu
Shapiro: 0.00460038
Log of Wkly Wge, Trns, Util, Com
Shapiro: 1.564e−05
Wkly Wge Whlesle, Retail, Tra
Shapiro: 4.62e−06
og of Wkly Wge Whlesle, Retail,
Shapiro: 0.00602804
Wkly Wge, Fin, Ins, Real Est
Shapiro: 0.00077405
Log of Wkly Wge, Fin, Ins, Real
Shapiro: 0.01620054
Wkly Wge, Service Industry
Shapiro: 0
Log of Wkly Wge, Service Indus
Shapiro: 0
Wkly Wge, Manufacturing
Shapiro: 6.6e−07
Log of Wkly Wge, Manufacturi
Shapiro: 0.00534857
Wkly Wge Fed Employees
Shapiro: 0.43556719
Log of Wkly Wge Fed Employe
Shapiro: 0.28456756
Wkly Wge State Employees
Shapiro: 0.35095601
Log of Wkly Wge State Employe
Shapiro: 0.77622282
Wkly Wge Local Gov Emps
Shapiro: 0.04620177
Log of Wkly Wge Local Gov Em
Shapiro: 0.07846706

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 null hypothesis of normality. The transformed variable is preferable for modelling.

Probability of Conviction

The log transform 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

Police as a percentage of population has a possitive skew, performing a natural log is preferable for modeling.

```
crime$log_polpc <- log(crime$polpc)
```

Percentage of Population That Are Young Males

This variable has a strong positive skew, using a natural log transformation results in a distribution that is still skewed, however, it is closer to normal and more consistent with other precentages.

```
crime$log_pctymle <- log(crime$pctymle)
```

People per Square Mile

There is one extreme outlier for county 173, with 0.000023 people per square mile. This will affect our modeling, specifically the cooks distance. Without that outlier, this variable is more normal with a log transformation.

```
crime$log_density <- log(crime$density)
```

Tax Revenue per Capita

Tax revenue per capita has a strong positive skew, using a natural log transformation results in a distribution that is still skewed, however, it is closer to normal.

```
crime$log_taxpc <- log(crime$taxpc)
```

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.

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

The 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 still reject the null hypothesis of normality. That said, the log transformation is preferable for modelling.

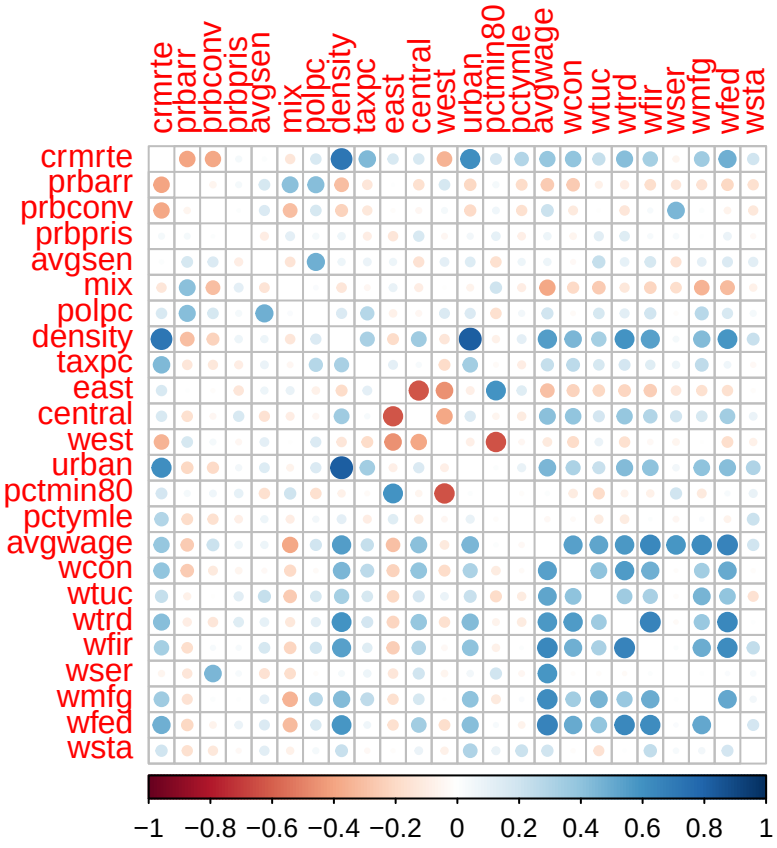
```
crime$log_mix <- log(crime$mix)
```

Percentage Minority, 1980

Relationships

Correlation Matrix

```
# Used for displaying subsets of variables.
columns_logical_order = c("county", "year", "crmrte", "prbarr", "prbconv",
                          "prbpris", "avgsen", "mix",  "polpc", "density",
                          "taxpc", "east", "central", "west", "urban",
                          "pctmin80", "pctymle", "avgwage", "wcon", "wtuc", "wtrd",
                          "wfir", "wser", "wmfg","wfed", "wsta", "wloc")

corrmatrix <- cor(crime[,columns_logical_order[3:26]])
corrplot(corrmatrix, cl.pos = "b", diag = FALSE)
```



Using the correlation matrix above we can see there are some strong correlations between Crime Rate (crmrte) and Population Density (density), Crime Rate and Urban, Other (likely east) counties and Percentage Minority from 1980 (pctmin80) and West and Percentage Minority, interestingly opposite correlations in

those last two. We also see strong correlation with Populations Density and the Average Wage (avgwage) and several wages, specifically trade, financial, insurance real estate and federal employees.

Checking signifance of wage variables

```
crmrate_wage_model <- lm(log(crmrte)~wcon + wtuc + wtrd + wfir + wser + wmfg
                         +wfed + wsta + wloc, data=crime)
(crmrate_wage_model$coefficients)
```

```
##   (Intercept)          wcon          wtuc          wtrd          wfir
## -6.0765119907  0.0024350006 -0.0003371284  0.0025645560 -0.0021917914
##          wser          wmfg          wfed          wsta          wloc
## -0.0003059448  0.0006737030  0.0038141894  0.0018627518 -0.0011344846
```

```
f <- summary(crmrate_wage_model )$fstatistic
```

As all coefficients are close to 0 -> we can say none of the wage variables are independently significant. But, the F-stat p-value 5.2040089×10^{-5} shows that all the wage variables are jointly significant. wcon and wfed are somewhat important.

Model Analysis

Model1 - Minimum Specification

Variables of Interest -> prbconv, prbarr, density, polpc. Based on the table above, We anticipate that the crime rate depends on certainty of punishment and severity of punishment along with police per capita and density. As such, our base model includes variables that attempt to operationalize those concepts

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

```
model1 = lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density, data = crime)
(model1$coefficients)
```

```
## (Intercept)      prbarr     prbconv   log_polpc log_density
##   1.0078551  -1.9244220  -0.6923862   0.5559296   0.1112556
```

Checking if county 71 that has both west=central=1 has any impact on our model.

```
model1.a = lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density, data = crime[crime$county != 71,])
(model1.a$coefficients)
```

```
## (Intercept)      prbarr     prbconv   log_polpc log_density
##   1.0138374  -1.9250387  -0.6932570   0.5566923   0.1115396
```

We don't observe any major changes to our coefficients here, it could be because we didn't add any region variable in our model.

Checking if county 115 that has high prbarr and low crime and low polpc and very low density has any impact on our model.

```
model1.b = lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density, data = crime[crime$county != 115,])
(model1.b$coefficients)
```

```
## (Intercept)      prbarr     prbconv   log_polpc log_density
##   1.4845485  -1.4606859  -0.5934248   0.6575126   0.1238930
```

From the coefficents summary, we can observe that the county 115 has a very high impact on our proposed model. This county can impact our CLM assumptions.

Testing the validity of the 6 assumptions of the CLM

CLM1 - Linear model

The model is specified such that the dependent variable is a linear function of the explanatory variables. We assume linearity in the dependent variable vs independent variables by default. Is the assumption valid? Yes

CLM2 - Random Sampling

Since we got 90% of records from the total counties in North Carolina, we expect this assumption is valid. But to point out, the counties we didn't have can have high density or high policing and low probability of arrest or probability of conviction. This can change our interpretation of the crime rate. Since, we see that crime rate is normally distributed in the given data set, we anticipate that this would be same if we consider the entire population.

We can see that west has less number of counties compared to east or central. This can introduce some bias.But, if we look at the map of North Carolina, west side is narrower compared to central and east. So, this confirms that the bias is not a problem here.

Is the assumption valid? Highly Likely

CLM3 - Multicollinearity

We can see from the above correlation matrix, Crime Rate is highly correlated with the variables in our model. We also see that probability of arrest and density are correlated -0.3027029. Let's also check how probability of arrest and density alone are jointly affecting crime rate.

```
model1.c = lm(log_crmrte ~ (prbarr) +  log_density, data = crime)
f <- summary(model1.c )$fstatistic
```

The p-value of the entire model 9.0164669×10^{-9} indicates that both these variables are jointly significant.

```
vif(model1)
```

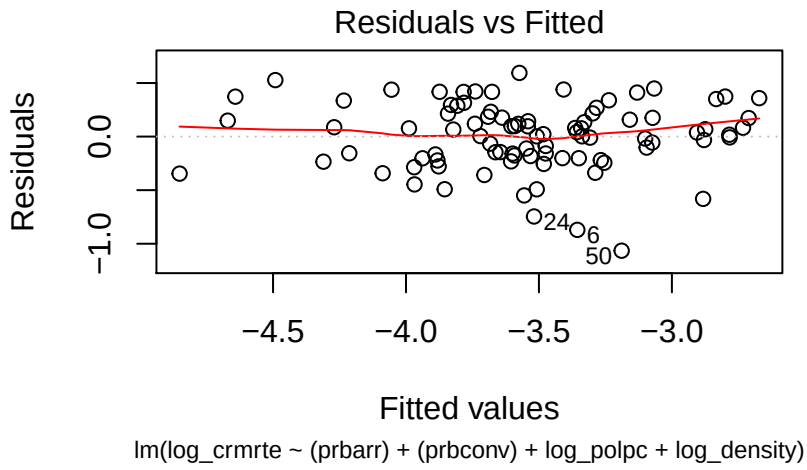
```
##      prbarr     prbconv   log_polpc log_density
##    1.217343    1.016318    1.061591    1.166073
```

Based on pairwise correlation in the dependent variables prbarr, prbconf, polpc, density, log_polpc and log_density, and no variance inflation factors near 10, we do not detect evidence of multicollinearity negatively impacting our specification.

Is the assumption valid? YES

CLM 4 - Zero conditional mean

```
plot(model1, which=1, cex.sub=0.75)
```



From the residuals vs fitted plot, the residuals are centered on 0 except with three values. outside of that all values are close to zero. Is the assumption valid? Highly Likely but not 100% sure

CLM 5 - Homoscedasticity

So, it is not easy to determine Homoscedasticity from the residuals vs fitted values plot alone. Running some additional tests.

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 7.186, df = 4, p-value = 0.1264
```

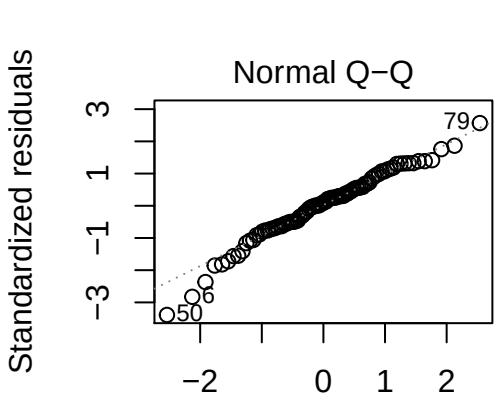
```
ncvTest(model1)
```

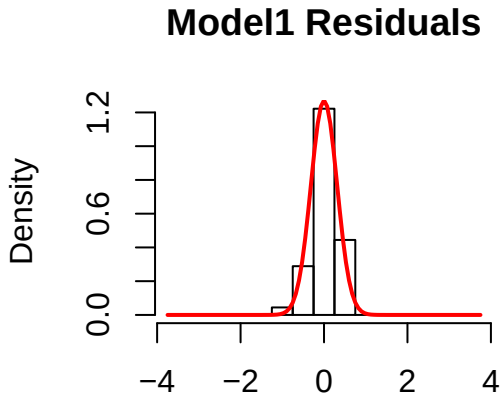
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0004349806, Df = 1, p = 0.98336
```

Both tests are showing small p-values showing that we have to reject the hypothesis. Homoscedasticity does not appear to be a valid assumption here indicating that our standard errors may not be used for inference. Is the assumption valid? NO

CLM 6 - Normality of Residuals

```
plot(model1, which=2)
hist(model1$residuals,main="Model1 Residuals", breaks = seq(-3.75, 3.75, 0.5), freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(model1$residuals)), col="red", lwd=2, add=TRUE)
```

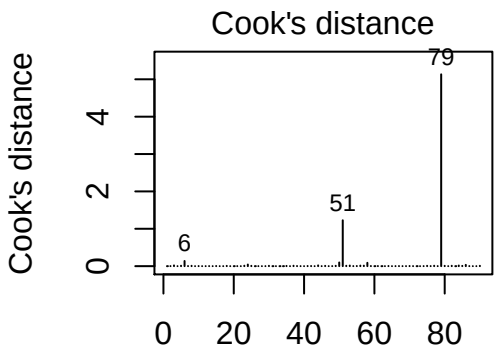


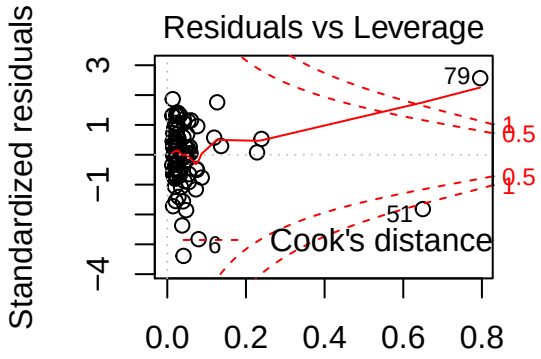


Other than a few outliers, the distribution is relatively normal for our given sample size. We do see some outliers in the Q-Q plot indicating that there is some skew because of the outlier values at the ends. Is the assumption valid? Higly likely but not 100% sure

Cook's Distance:

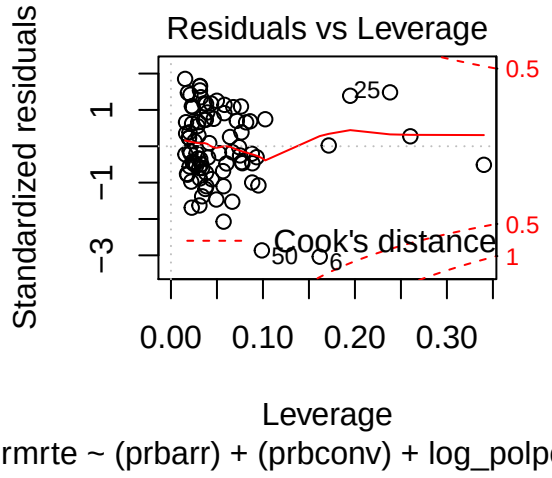
```
plot(model1, which=4)
plot(model1, which=5)
```





There are some influential values however cook's distance is within the bounds. At value 51 and 79. ####if we remove those values and plot the graph .

```
model1.d <- lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density, data = crime[c(-79,-51),])
plot(model1.d, which=5)
```



Index 51 and 79 - The density and polpc is very low -> where as the crime rate and prb arr and prb conv is similar to other counties. Both these are western counties. Now, we could see Cook's is within the bounds.

```
model1_intrepreation <-  c("","For ~ 1 unit increase in probability of arrest,
                crime rate decreases by 1.92%",
                "For ~ 1 unit increase in probability of conviction,
                crime rate decreases by 0.69%",
                "For ~ 1% increase in polpc,
                crime rate increases by 55.5%",
                "For ~ 1 unit increase (100 people per square mile) in density,
                crime rate increases by 11.1%")
model1_coefficients <- data.frame("Model 1 Coefficients" = round(model1$coefficients, 4),
                    "Interpretation" = model1_intrepreation)
kable(model1_coefficients, booktabs = TRUE) %>%
  kable_styling(font_size = 8, full_width = FALSE) %>%
      column_spec(3, width = "35em")
```

	Model.1.Coefficients	Interpretation
(Intercept)	1.0079	
prbarr	-1.9244	For ~ 1 unit increase in probability of arrest, crime rate decreases by 1.92%
prbconv	-0.6924	For ~ 1 unit increase in probability of conviction, crime rate decreases by 0.69%
log_polpc	0.5559	For ~ 1% increase in polpc, crime rate increases by 55.5%
log_density	0.1113	For ~ 1 unit increase (100 people per square mile) in density, crime rate increases by 11.1%

All of the coefficients are highly statistically significant when we look at heteroskedastic-robust errors:

coefficient-significance < Heteroskedastic-Robust Errors >

```
coeftest(model1, vcov = vcovHC, level = 0.05)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.00786    1.26028  0.7997 0.426110
## prbarr      -1.92442    0.61493 -3.1295 0.002400 **
## prbconv     -0.69239    0.15783 -4.3870 3.28e-05 ***
## log_polpc    0.55593    0.18755  2.9642 0.003937 **
## log_density  0.11126    0.13045  0.8529 0.396129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model conclusion -

Probability of arrest has more impact on crime rate - it is easier to get arrested than convicted. From the arrests, we can see even 50% are not convicted.

The adjusted R^2 for this model is 0.6579144% which means a lot of the the variations are explained by this model

The Akaike information criterion indicates that the relative quality of predicting crime rate based on our variables is 57.7000864 when only little information is lost.

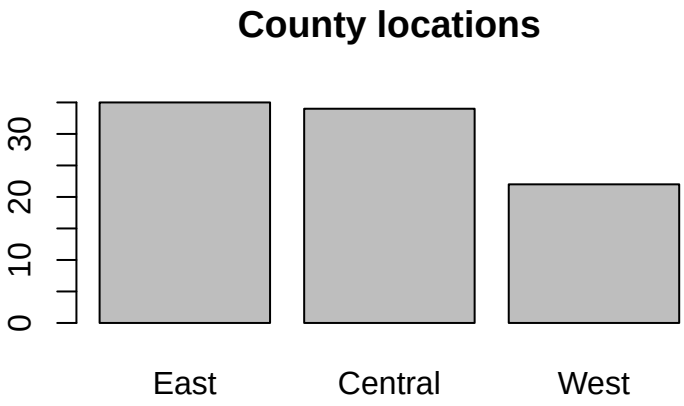
Model2 - Optimal Specification

In additional to the explanatory variables introduced in our #Model1, we have decided to include the following variables in the model. `west`, `pctmin80`, and an interaction of `west` and `polpc`. This is because from the below table, west some how higher polpc and low crime rate. which is opposite to what we have observed in our model1. Also, we would ike to pctmin80 -> how crime rate changes with percent minority.

```
barplot(c(sum(crime$east), sum(crime$central), sum(crime$west)),
       names.arg = c("East", "Central", "West"), main = "County locations")

crime$region <- ifelse(crime$west == 1, "West",
                       ifelse(crime$central == 1, "Central",
                              ifelse(crime$east == 1, "East",        "other")))

region = aggregate(density ~ region, data = crime, mean)
region$polpc = aggregate(polpc ~ region, data = crime, mean)[2]
region$crmrte = aggregate(crmrte ~ region, data = crime, mean)[2]
colnames(region) = c("Region", "Density", "Police per Cap", "Crime Rate")

kable(region, "latex", longtable = TRUE, booktabs = TRUE, caption = "Regions") %>%
    kable_styling(full_width = TRUE, latex_options =
                    c("HOLD_position", "striped", "repeat_header"),
                  row_label_position = 1)
```

Table 2: Regions

Region	Density	Police per Cap	Crime Rate
Central	2.047960	0.001637046	0.03699627
East	1.085503	0.001609983	0.03739491
West	1.074319	0.001970259	0.02209975



There are three regions, two provided in the data set; West and Central, and those not in those two regions which we have determined to be East.

West shows a higher mean police per capita and a lower crime rate, which is likely opposing our initial expectations.

$$log(crmrte) = \beta_0 + \beta_1(prbarr) + \beta_2(prbconv) + \beta_3 log(polpc) + \beta_4 log(density) + \beta_5(pctmin80) + \beta_6 west + \beta_7 west * log(polpc) + u$$

```
model2 = lm(log_crmrte ~ (prbarr) + (prbconv)+ log_polpc +
              log_density + pctmin80 +
              west + west * log_polpc, data = crime)
(model2$coefficients)
```

```
##        (Intercept)             prbarr            prbconv
##         1.91895608        -1.82080087        -0.69040694
##          log_polpc        log_density           pctmin80
##         0.73628656         0.10439659         0.01001377
##           westTRUE log_polpc:westTRUE
##        -1.81032462        -0.26035345
```

####checking if county 71 that has both west=central=1 has any impact on our model.

```
model2.a = lm(log_crmrte ~ (prbarr) + (prbconv)+ log_polpc +
               log_density + pctmin80 +
              west + west * log_polpc, data = crime[crime$county != 71,])
(model2.a$coefficients)
```

```
##        (Intercept)             prbarr            prbconv
##        1.968389066       -1.784011302       -0.677641311
##          log_polpc        log_density           pctmin80
##        0.744786409        0.099275321        0.009661545
##           westTRUE log_polpc:westTRUE
##       -2.090092375       -0.299829159
```

There is a very small change in the coefficient of west, 2.09-1.81 ~ 0.3 - but almost all other coefficients doesn't change much

####checking if county 115 that has high prbarr and low crime and low polpc and very low density has any impact on our model.

```
model2.b = lm(log_crmrte ~ (prbarr) + (prbconv)+ log_polpc +
                log_density + pctmin80 +
              west + west * log_polpc, data = crime[crime$county != 115,])
(model2.b$coefficients)
```

```
##        (Intercept)             prbarr            prbconv
##         1.93987440        -1.89308556        -0.70741807
##          log_polpc        log_density           pctmin80
##         0.73576203         0.10010645         0.01018997
##           westTRUE log_polpc:westTRUE
##        -2.17568786        -0.31697141
```

From the coefficents summary, we can observe that the county 115 deosnt have a very high impact on our proposed model. but this changed alot in our previous model. This could be due to low crime rate and high polpc in western regions.

###Testing the validity of the 6 assumptions of the CLM #### CLM1 - Linear model

our views are identical to the previous model

CLM2 - Random Sampling

our views are identical to the previous model

CLM3 - Multicollinearity

We can see from the above correlation matrix, Crime Rate is highly correlated with the variables in our model.

```
vif(model2)
```

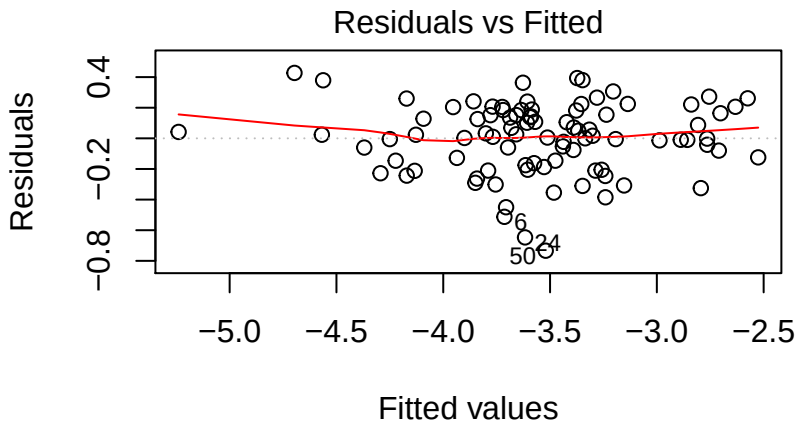
```
##         prbarr        prbconv      log_polpc    log_density       pctmin80
##       1.538158       1.086812       1.893781       1.364137       1.946049
##           west log_polpc:west
##     350.110767     337.569876
```

Based on the low pairwise correlation between the independent variables and variance inflation factors all well below 10, we do not detect evidence of multicollinearity negatively impacting our specification.

Is the assumption valid? YES

CLM 4 - Zero conditional mean

```
plot(model2, which=1, cex.sub=0.75)
```



From the residuals vs fitted plot, the residuals are centered on 0 except few values values. outside of that all values are close to zero. Is the assumption valid? Highly Likely but not 100% sure

CLM 5 - Homoscedasticity

So, it is not easy to determine Homoscedasticity from the residuals vs fitted values plot alone. Running some additional tests.

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 11.268, df = 7, p-value = 0.1273
```

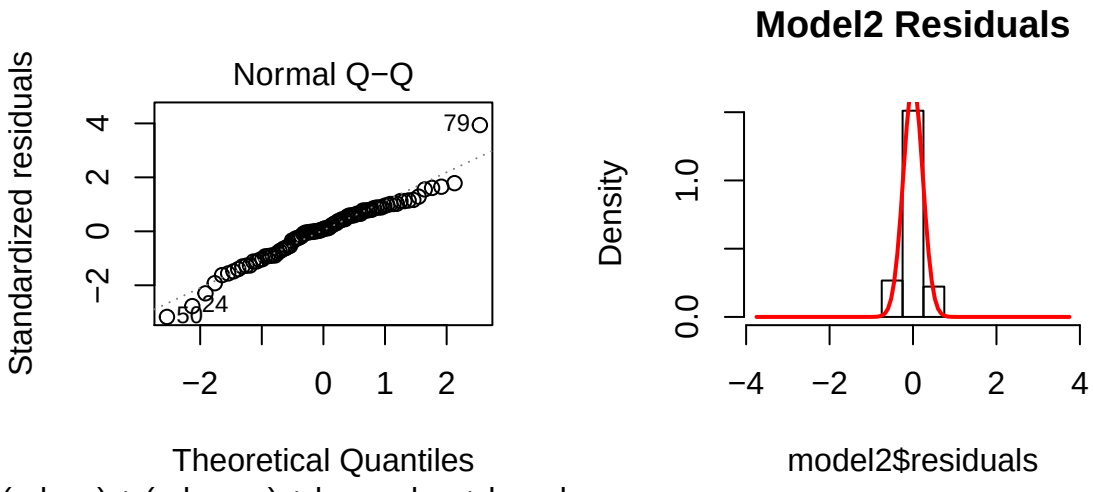
```
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2935739, Df = 1, p = 0.58794
```

Both tests are showing small p-values showing that we have to reject the hypothesis. Homoscedasticity does not appear to be a valid assumption here indicating that our standard errors may not be used for inference. Is the assumption valid? NO

CLM 6 - Normality of Residuals

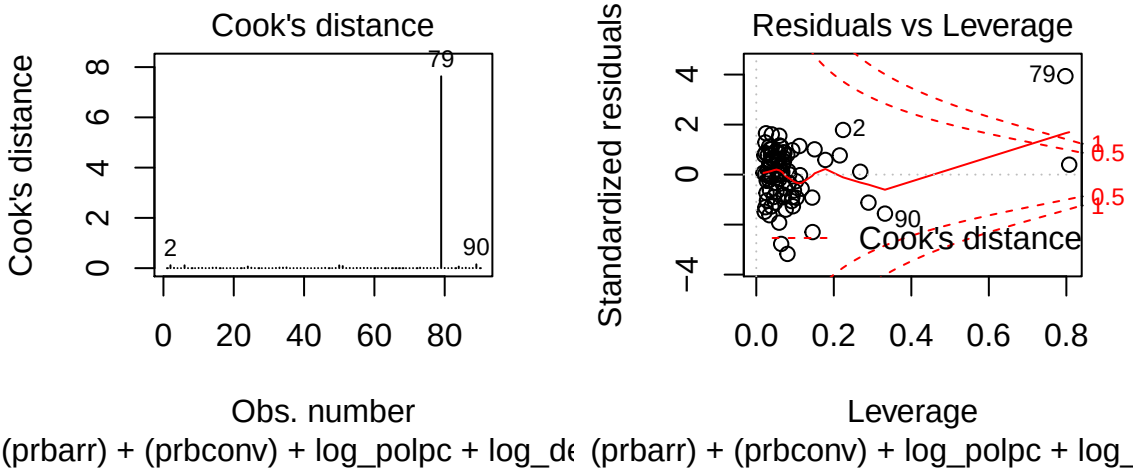
```
plot(model2, which=2)
hist(model2$residuals,main="Model2 Residuals", breaks = seq(-3.75, 3.75, 0.5), freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(model2$residuals)), col="red", lwd=2, add=TRUE)
```



Other than a few outliers, the distribution is relatively normal for our given sample size. We do see some outliers in the Q-Q plot indicating that there is some skew because of the outlier values at the ends. Is the assumption valid? Higly likely but not 100% sure

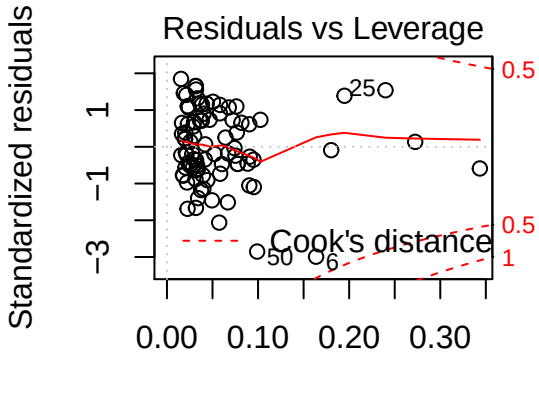
Cook's Distance:

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



There are some influential values however cook's distance is within the bounds. At value 90, 51 and 79. if we remove those values and plot the graph, .

```
model2.c <- lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density, data = crime[c(-51,-79,-90),])
plot(model2.c, which=5)
```



Index 51, 79 and 90 - The density and polpc is very low -> where as the crime rate and prb arr and prb conv and pctmin80 is similar to other counties. ANd all are western counties.

```
model2_intrepreation <-  c("","For ~ 1 unit increase in probability of arrest,
                crime rate decreases by 1.82%",
                "For ~ 1 unit increase in probability of conviction,
                crime rate decreases by 0.69%",
                "For ~ 1% increase in polpc,
                crime rate increases by 73%",
                "For ~ 1 unit increase (100 people per square mile) in density,
                crime rate increases by 10.4%",
                "For ~ 1 unit increase in percent minority,
                crime rate increases by 1%",
                "For all western counties crime rate decreases by
                1.81% ~ compared to other counties",
                "For all western counties,
                an ~ 1% increase in polpc reduces the crime rate by 0.26%")
model2_coefficients <- data.frame("Model 2 Coefficients" = round(model2$coefficients, 4),
                    "Interpretation" = model2_intrepreation)
kable(model2_coefficients, booktabs = TRUE) %>%
  kable_styling(font_size = 8, full_width = FALSE) %>%
      column_spec(3, width = "35em")
```

	Model.2.Coefficients	Interpretation
(Intercept)	1.9190	
prbarr	-1.8208	For ~ 1 unit increase in probability of arrest, crime rate decreases by 1.82%
prbconv	-0.6904	For ~ 1 unit increase in probability of conviction, crime rate decreases by 0.69%
log_polpc	0.7363	For ~ 1% increase in polpc, crime rate increases by 73%
log_density	0.1044	For ~ 1 unit increase (100 people per square mile) in density, crime rate increases by 10.4%
pctmin80	0.0100	For ~ 1 unit increase in percent minority, crime rate increases by 1%
westTRUE	-1.8103	For all western counties crime rate decreases by 1.81% ~ compared to other counties
log_polpc:westTRUE	-0.2604	For all western counties, an ~ 1% increase in polpc reduces the crime rate by 0.26%

All of the coefficients are highly statistically significant except density when we look at heteroskedastic-robust errors: This indicates that density doesn't have much impact for western counties.

coefficient-significance < Heteroskedastic-Robust Errors >

```
coeftest(model2, vcov = vcovHC, level = 0.05)
```

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          1.9189561  1.4091365  1.3618 0.1769920
## prbarr              -1.8208009  0.3830028 -4.7540 8.409e-06 ***
## prbconv             -0.6904069  0.1415946 -4.8759 5.227e-06 ***
## log_polpc            0.7362866  0.1988147  3.7034 0.0003846 ***
## log_density          0.1043966  0.1519063  0.6872 0.4938687
## pctmin80             0.0100138  0.0021665  4.6220 1.398e-05 ***
## westTRUE            -1.8103246  1.2834530 -1.4105 0.1621711
## log_polpc:westTRUE -0.2603535  0.1955745 -1.3312 0.1868047
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model conclusion - Probability of arr has more impact on crime rate - it is easier to get arrested than convicted. From the arrests, we can see even 50% are not convicted.

The adjusted R^2 for this model is 0.8068387% which means a lot of the the variations are explained by this model:

The Akaike information criterion (AIC) indicates that the relative quality of predicting crime rate based on our variables is 9.0279987 when only little information is lost.

Model3 - Sub-optimal Specification

Other than the variables we added in our model2, we would like to see how pctymle and taxpc has any impact on the crime rate. Also, from our wage variable analysis above, we would like to add wcon and wfed to our model. We are not using any interaction terms here and want to check how crimerate is varying across all regions for these variables.

$$log(crmrte) = \beta_0 + \beta_1 log(prbarr) + \beta_2 log(prbconv) + \beta_3 log(polpc) + \beta_4 log(density) + \beta_5 (pctmin80) + \beta_6 log(pctymle) + \beta_7 log(taxpc) + \beta_8 (wcon) + \beta_9 (wfed) + u$$

```
model3 = lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density + pctmin80
            + log_pctymle + log_taxpc + wcon + wfed, data = crime)
(model3$coefficients)
```

```
##   (Intercept)        prbarr       prbconv     log_polpc   log_density
##  0.1861357787 -1.8781188626 -0.7001301544  0.5623523453  0.0905239339
##      pctmin80   log_pctymle     log_taxpc          wcon          wfed
##  0.0118914982  0.0743049895  0.0098133796  0.0009643015  0.0009523985
```

Checking if county 115 that has high prbarr and low crime and low polpc and very low density has any impact on our model.

```
model3.a = lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density + pctmin80
            + log_pctymle + log_taxpc + wcon + wfed, data = crime[crime$county != 115,])
(model3.a$coefficients)
```

```
##   (Intercept)        prbarr       prbconv     log_polpc   log_density
##  0.2611456276 -1.8452057210 -0.6930713953  0.5721926350  0.0917819362
##      pctmin80   log_pctymle     log_taxpc          wcon          wfed
##  0.0118372394  0.0733529718  0.0048324894  0.0009534671  0.0009432399
```

From the coefficents summary, we can observe that the county 115 doesn't have any big impact on our coefficients.

###Testing the validity of the 6 assumptions of the CLM #### CLM1 - Linear model

our views are identical to the previous model

CLM2 - Random Sampling

our views are identical to the previous model

CLM3 - Multicollinearity

We can see from the above correlation matrix, Crime Rate is highly correlated with the variables in our model. We do see strong correleation between wfed and wcon 'r cor(crime$wcon, crime$wfed)' along with polpc and density. Let's also check how these variables alone are affecting crime rate.

```
model3.b = lm(log_crmrte ~ (wcon) + (wfed) + log_density + log_polpc, data = crime)
(model3.b$coefficients)
```

```
##  (Intercept)         wcon         wfed  log_density    log_polpc
## -3.078138710  0.001638614  0.002040546  0.127538735  0.282328074
```

```
f <- summary(model1.c )$fstatistic
```

The p-value of the entire model 9.0164669×10^{-9} indicates that both these variables are jointly significant. But the co-efficients of wcon and wfed are so-close to zero.

```
vif(model3)
```

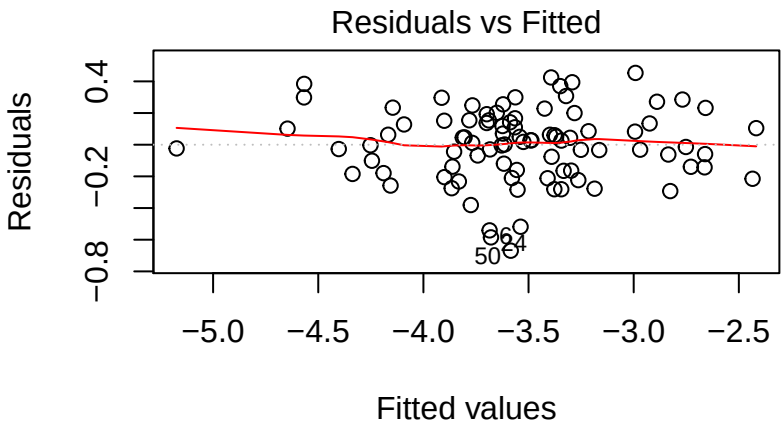
```
##      prbarr     prbconv   log_polpc log_density    pctmin80 log_pctymle
##    1.530657    1.146751    2.055567    1.659918    1.158265    1.391718
##   log_taxpc        wcon        wfed
##    1.622171    1.564755    2.273048
```

Based on pairwise correlation in the dependent variables and no variance inflation factors near 10, we do not detect evidence of multicollinearity negatively impacting our specification.

Is the assumption valid? YES

CLM 4 - Zero conditional mean

```
plot(model3, which=1, cex.sub=0.75)
```



From the residuals vs fitted plot, the residuals are centered on 0 except with few values values. outside of that all values are close to zero. Is the assumption valid? Highly Likely but not 100% sure

CLM 5 - Homoscedasticity

So, it is not easy to determine Homoscedasticity from the residuals vs fitted values plot alone. Running some additional tests.

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 22.417, df = 9, p-value = 0.007648
```

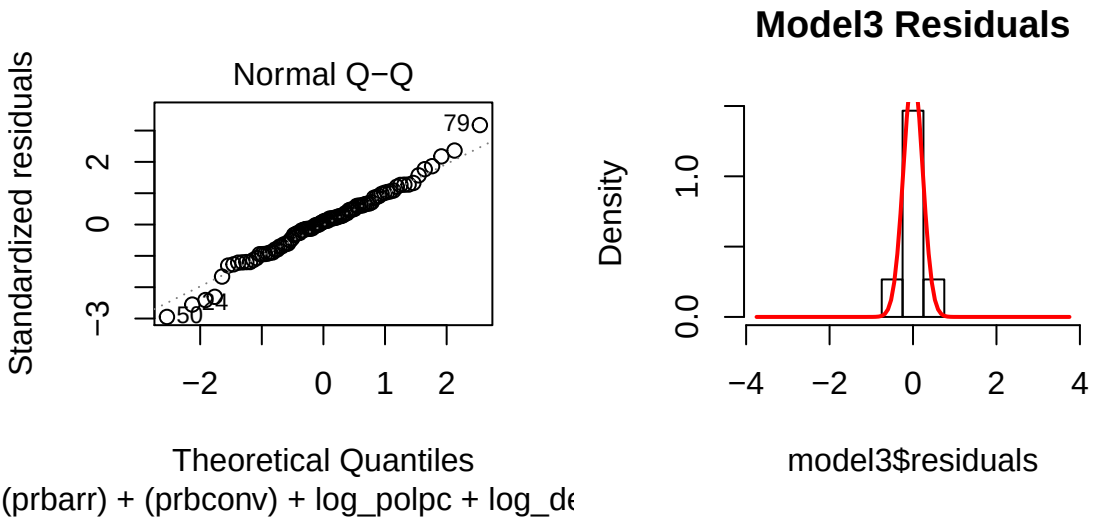
```
ncvTest(model3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.002576534, Df = 1, p = 0.95952
```

Both tests are showing small p-values showing that we have to reject the hypothesis. Homoscedasticity does not appear to be a valid assumption here indicating that our standard errors may not be used for inference. Is the assumption valid? NO

CLM 6 - Normality of Residuals

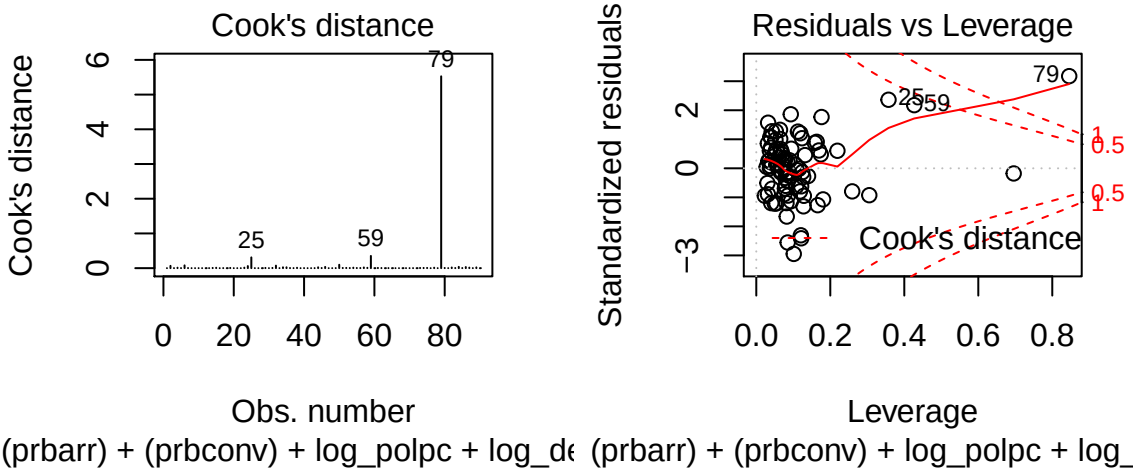
```
plot(model3, which=2)
hist(model3$residuals,main="Model3 Residuals", breaks = seq(-3.75, 3.75, 0.5), freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(model3$residuals)), col="red", lwd=2, add=TRUE)
```



Other than a few outliers, the distribution is relatively normal for our given sample size. We do see some outliers in the Q-Q plot indicating that there is some skew because of the outlier values at the ends. Is the assumption valid? Higly likely but not 100% sure

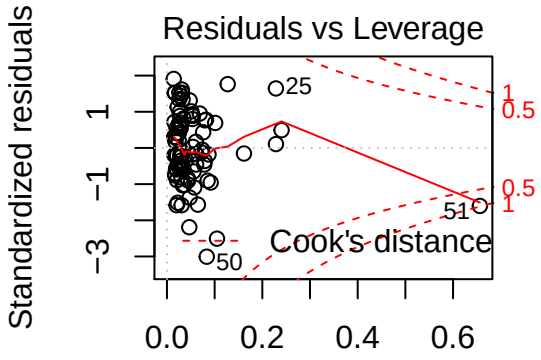
Cook's Distance:

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



There are some influential values however cook's distance is within the bounds. At value 59 and 79. ####if we remove those values and plot the graph, we could see Cook's distance is within the bounds.

```
model3.c <- lm(log_crmrte ~ (prbarr) + (prbconv)+
              log_polpc +  log_density, data = crime[c(-79,-59),])
plot(model3.c, which=5)
```



Index 51 and 79 - The density and polpc is very low -> where as the crime rate and prb arr and prb conv is similar to other counties. Both these are western counties Index 59 - This is an Eastern county with low crime rate and low polpc. But it has high pctmin80 that is affecting our accuracy.

```
model3_intrepreation <-  c("","For ~ each 1 unit
increase in probability of arrest,
                crime rate decreases by 1.87%",
"For ~ each 1 unit increase in
probability of conviction,crime rate decreases by 0.7%",
"For ~ 1% increase in polpc, crime rate increases by 56.2%",
"For ~ 1 unit increase (100 people per square mile)
in density,crime rate increases by 9%",
"For ~ 1 unit increase of percent minority, crime rate increases by 1%",
"For ~ 1% increase in pctymle, crime rate increases by 0.07%",
"For ~ 1% increase in taxpc, crime rate increases by 0.9%",
"For ~ 1 unit increase in wcon, crime rate increases by .09%",
"For ~ 1 unit increase in wfed, crime rate increases by 0.09%")
model3_coefficients <- data.frame("Model 3 Coefficients" = round(model3$coefficients, 4),
                    "Interpretation" = model3_intrepreation)
kable(model3_coefficients, booktabs = TRUE) %>%
  kable_styling(font_size = 8, full_width = FALSE) %>%
      column_spec(3, width = "35em")
```

	Model.3.Coefficients	Interpretation
(Intercept)	0.1861	
prbarr	-1.8781	For ~ each 1 unit increase in probability of arrest, crime rate decreases by 1.87%
prbconv	-0.7001	For ~ each 1 unit increase in probability of conviction,crime rate decreases by 0.7%
log_polpc	0.5624	For ~ 1% increase in polpc, crime rate increases by 56.2%
log_density	0.0905	For ~ 1 unit increase (100 people per square mile) in density,crime rate increases by 9%
pctmin80	0.0119	For ~ 1 unit increase of percent minority, crime rate increases by 1%
log_pctymle	0.0743	For ~ 1% increase in pctymle, crime rate increases by 0.07%
log_taxpc	0.0098	For ~ 1% increase in taxpc, crime rate increases by 0.9%
wcon	0.0010	For ~ 1 unit increase in wcon, crime rate increases by .09%
wfed	0.0010	For ~ 1 unit increase in wfed, crime rate increases by 0.09%

All of the coefficients are highly statistically significant for prbarr, prbconv, polpc, density and pctmin80 when we look at heteroskedastic-robust errors. But taxpc, pctymle, wcon and wfed doesn't have much signifance. But All these variables are jointly significant:

coefficient-significance < Heteroskedastic-Robust Errors >

```
coeftest(model3, vcov = vcovHC, level = 0.05)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.18613578  1.39110035  0.1338  0.893893
## prbarr      -1.87811886  0.32608991 -5.7595 1.505e-07 ***
## prbconv     -0.70013015  0.11734215 -5.9666 6.293e-08 ***
## log_polpc    0.56235235  0.15252989  3.6868  0.000412 ***
## log_density  0.09052393  0.15972295  0.5668  0.572467
## pctmin80     0.01189150  0.00179810  6.6134 3.885e-09 ***
## log_pctymle  0.07430499  0.26662725  0.2787  0.781206
## log_taxpc    0.00981338  0.15625430  0.0628  0.950079
## wcon         0.00096430  0.00071112  1.3560  0.178902
## wfed         0.00095240  0.00121401  0.7845  0.435061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mode3 conclusion -

Probability of arr has more impact on crime rate - it is easier to get arrested than convicted. From the arrests, we can see even 50% are not convicted.

The adjusted R^2 for this model is 0.8096141% which means a lot of the the variations are explained by this model. But is almost close to our model2.

The Akaike information criterion (AIC) indicates that the relative quality of predicting crime rate based on our variables is 9.5031292 when only little information is lost.

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:

$$\begin{aligned} crime\_rate = \beta_0 + \beta_1 crty\_punish + \beta_2 svrty\_punish + \beta_3 poverty\_rate + \\ \beta_4 educ + \beta_5 social\_cohesion + \beta_6 weapon\_availability + \beta_7 real\_wage + \\ \beta_8 low\_skill\_unemployment\_rate + \beta_9 age\_15\_to\_30\_proportion\_population + \\ \beta_{10} percent\_of\_population\_previously\_committed\_crime + \\ \beta_{11} percent\_of\_population\_previously_i mprisoned + ... + error \end{aligned}$$

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 operationlizations 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 poverty rate, education, social cohesion, weapon availability, cost of living, or the low skill unemployment rate are operationalized within this dataset.

Moreover, certain variables that 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 that belongs to a minority class could be the result of *red lining*, a segregating practice which led to poverty and low-quality education, both positively correlated with crime rate. It may also exhibit a parabolic relation with 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, poverty, 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 poverty. In practice, in the U.S. in the 1980s, we would expect social cohesion to be negatively correlated with density, while poverty may be positively correlated with density. We expect the beta for social cohesion to crime rate to be negative, while the beta for poverty 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 misleadign 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.

For future analyses, the availability of poverty rate, number of years of education and percentage of convicts, and the availability of these data as a regularly collected time series would aid in removing the above biases from the analysis and allow for more concrete policy recommendations.