Lab 3: Reducing Crime (FINAL: Stage 3)

N. Akkineni, K. Hanna, A. Thorp

December 10, 2018

Contents

Introduction **1**

Exploratory Data Analysis **2**
Data Summary . . . 2
Data Clean Up . . . 3
Concerns about the data . . . 4
Univariate Analysis . . . 5
Key Variable . . . 5
Explanatory Variables . . . 5
Relationships . . . 9
Correlation Matrix . . . 9

Model Analysis **10**
Model 1 - Minimum Specification . . . 10
Intuition of the effects of potential data issues in county 115 on the model . . . 10
Testing the validity of the 6 CLM assumptions . . . 11
Model 2 - Optimal Specification . . . 15
Intuition of the Effects of County 71 Existing in Both Central and West . . . 17
Intuition of the effects of potential data issues in county 115 on the model . . . 17
Testing the Validity of the 6 CLM assumptions . . . 17
Model 3 - Sub-optimal Specification . . . 22
Intuition of the effects of potential data issues in county 115 on the model . . . 22
Testing the Validity of the 6 CLM assumptions . . . 22

Omitted Variables **29**

Conclusion **30**

Introduction

Our team has been hired to provide research for a local political campaign, which would like to understand the determinants of crime rates in order 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 three regression models and find that estimators related to variables used to operationalize the concept of certainty of punishment are directionally consistent and statistically siginificant across modeling specifications. 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.

####Load Data and Package Dependencies

```r
#load package dependencies
library(knitr)
library(kableExtra)
suppressMessages(library(car))
suppressMessages(library(stargazer))
suppressMessages(library(lmtest))
suppressMessages(library(corrplot))
library(sandwich)

#load data and codebook explaining the meaning of variables
crime <- read.csv('crime_v2.csv',header = TRUE, sep = ",")
codebook <- read.csv('codebook.csv')
```

Exploratory Data Analysis

Data Summary

We were provided with a dataset that includes variables for 90 counties in North Carolina. This data appears to have been collated from: 1)crime statistics from the North Carolina Department of Corrections' prison and probation files; 2)demographic statistics taken from the decennial census; 3)police data derived from FBI police agency data; and 4)wage data from the North Carolina Employment Security commission.

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.

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
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

Table 1: Crime Data Codebook *(continued)*

Variable	Label
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

####Create Two New Variables

We create 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. Note, we do not use east in our core regression specifications, but rather to better understand 'west'.

```
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
```

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, if it was a true probability, would not be 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 true 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, by construction as such a ratio, it is plausible to have values greater than 1 as a single arrest can result in multiple convictions, this may reflect natural differences in time period as arrests in one year do not mean all cases are decided in the same year, 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 about the data 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 listed to the tenths digit at 1.5, 0.5 and 0.1 respectively. Typical values for these variables contain 6, 7 and 8 decimal places, making 3 values all rounding to the tenths digit highly improbable.

A small number of these issues is cause for concern. This many issues suggests the observations for county 115 may be erroneous. We test the impact of this observation below.

Extreme Outlier Single Data point - 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. After researching the county in question through other data sources and seeing that its per capita income is lower than state average and that its population is not especially small (mitigating concerns about small-n leading to a handful of individuals exerting large influence on the mean), we believe this value may be erroneous.

Univariate Analysis

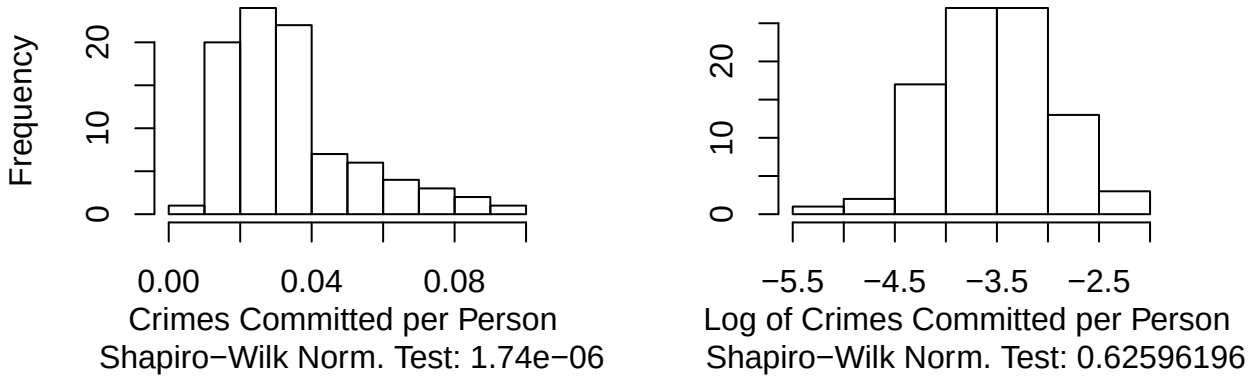
Key Variable

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

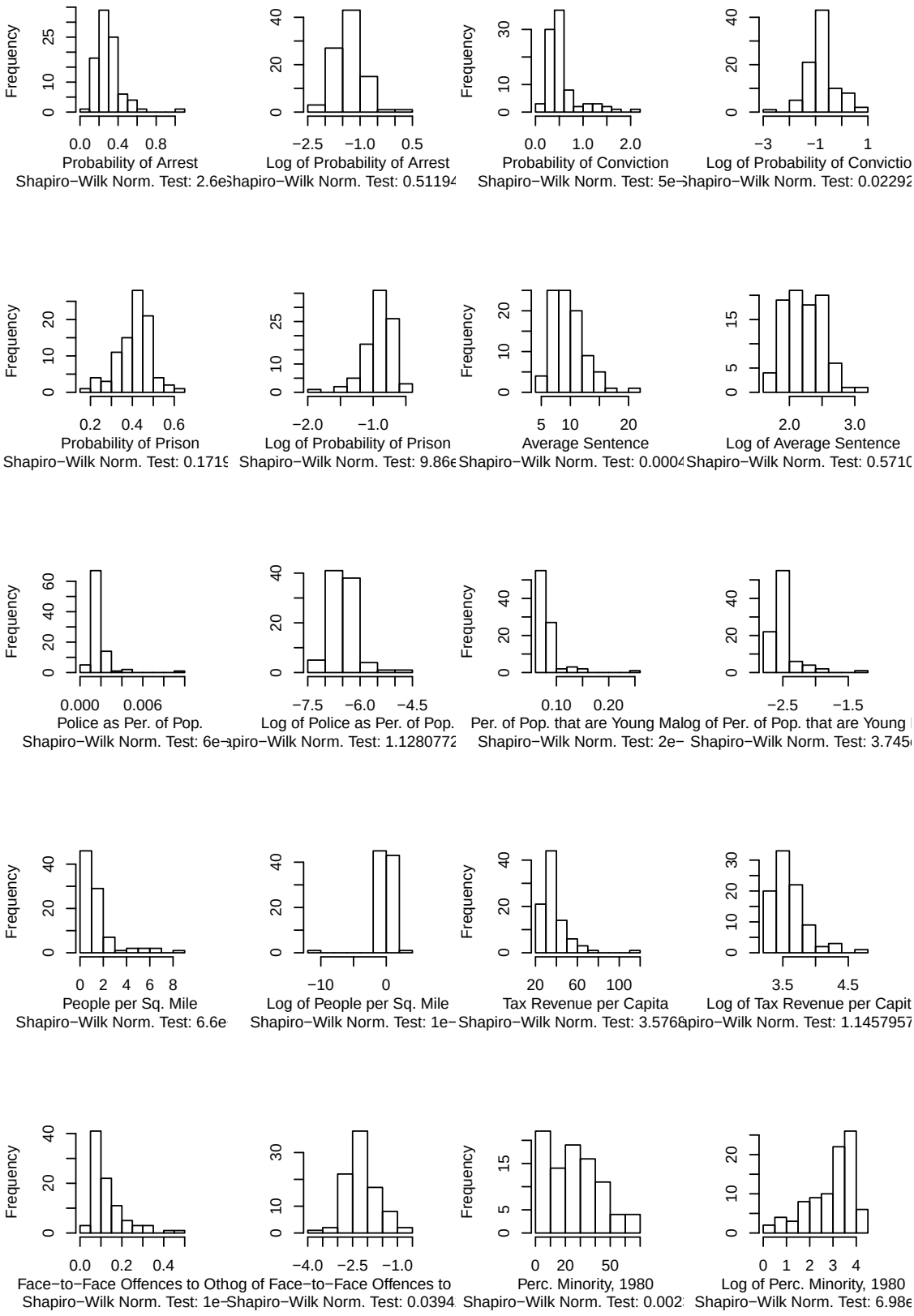
```
crime$log_crmrte <- log(crime$crmrte)
crmrte.outliers = boxplot(crime$log_crmrte, plot = FALSE)$out
```

Crime rate has 1 outliers, though it is not enough to cause concern.

Explanatory Variables

Diagrams of Variables With and Without Log Transformations

```r
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−Wilk Norm. Test: 2.6e−
Log of Probability of Arrest
Shapiro−Wilk Norm. Test: 0.5119
Probability of Conviction
Shapiro−Wilk Norm. Test: 5e−
Log of Probability of Convictio
Shapiro−Wilk Norm. Test: 0.0229
Probability of Prison
Shapiro−Wilk Norm. Test: 0.171
Log of Probability of Prison
Shapiro−Wilk Norm. Test: 9.86e
Average Sentence
Shapiro−Wilk Norm. Test: 0.000
Log of Average Sentence
Shapiro−Wilk Norm. Test: 0.571
Police as Per. of Pop.
Shapiro−Wilk Norm. Test: 6e−
Log of Police as Per. of Pop.
Shapiro−Wilk Norm. Test: 1.128077
Per. of Pop. that are Young Ma
Shapiro−Wilk Norm. Test: 2e−
Log of Per. of Pop. that are Young
Shapiro−Wilk Norm. Test: 3.745
People per Sq. Mile
Shapiro−Wilk Norm. Test: 6.6e
Log of People per Sq. Mile
Shapiro−Wilk Norm. Test: 1e−
Tax Revenue per Capita
Shapiro−Wilk Norm. Test: 3.576
Log of Tax Revenue per Capit
Shapiro−Wilk Norm. Test: 1.145795
Face−to−Face Offences to Oth
Shapiro−Wilk Norm. Test: 1e−
Log of Face−to−Face Offences to
Shapiro−Wilk Norm. Test: 0.0394
Perc. Minority, 1980
Shapiro−Wilk Norm. Test: 0.002
Log of Perc. Minority, 1980
Shapiro−Wilk Norm. Test: 6.98e

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

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 a 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 positive 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.

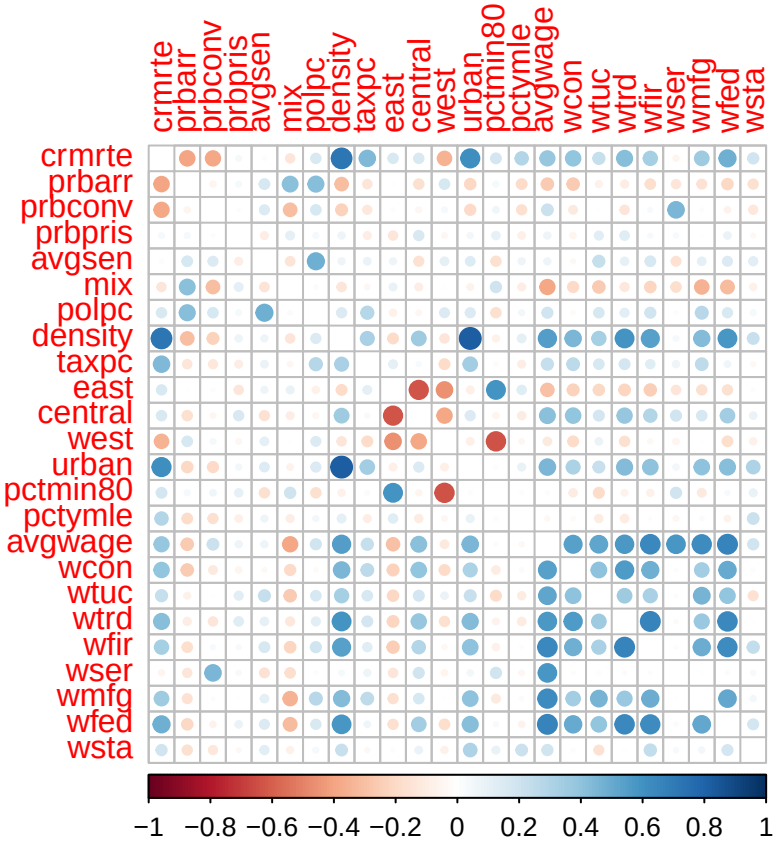
The mix of face-to-face crimes to other crimes has a positive skew, applying a natural log transformation creates a more symmetrical distribution. However, the resulting Shapiro-Wilk test would still reject the null hypothesis of normality. That said, the log transformation is preferable for modelling.

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

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 population 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)
```

```
##     value     numdf     dendf
##  4.693485  9.000000 80.000000
```

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

Model 1 - Minimum Specification

Based on the table above, we anticipate that variables operationalizing certainty of punishment (prbarr, prbconv, polpc) may have explanatory power for log(crime rate). We also expect that density may have a strong predictive power. As such, we include these variables in our baseline specification.

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

```
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
```

Intuition of the effects of potential data issues in county 115 on the 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, particularly for the variables operationalizing certainty of punishment.

Testing the validity of the 6 CLM assumptions

CLM 1: Linear model

The mode l 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.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 2: 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 could 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. In other words, we assume that the 10 missing observations are missing completely at random. We have no information at hand that would lead us to think otherwise.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 3: Multicollinearity

```
cor_model1 <- data.matrix(subset(crime,
 select = c("prbarr","prbconv", "polpc", "density", "log_polpc","log_density")))
cor_model1_out = round(cor(cor_model1), 4)
kable(cor_model1_out)
```

	prbarr	prbconv	polpc	density	log_polpc	log_density
prbarr	1.0000	-0.0558	0.4260	-0.3027	0.2162	-0.3479
prbconv	-0.0558	1.0000	0.1719	-0.2267	-0.0076	-0.0861
polpc	0.4260	0.1719	1.0000	0.1591	0.9058	-0.0563
density	-0.3027	-0.2267	0.1591	1.0000	0.3250	0.5785
log_polpc	0.2162	-0.0076	0.9058	0.3250	1.0000	0.0230
log_density	-0.3479	-0.0861	-0.0563	0.5785	0.0230	1.0000

We can see from the previous correlation matrix, Crime Rate is highly correlated with the variables in our model. We also see that probability of arrest and density are highly 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)
```

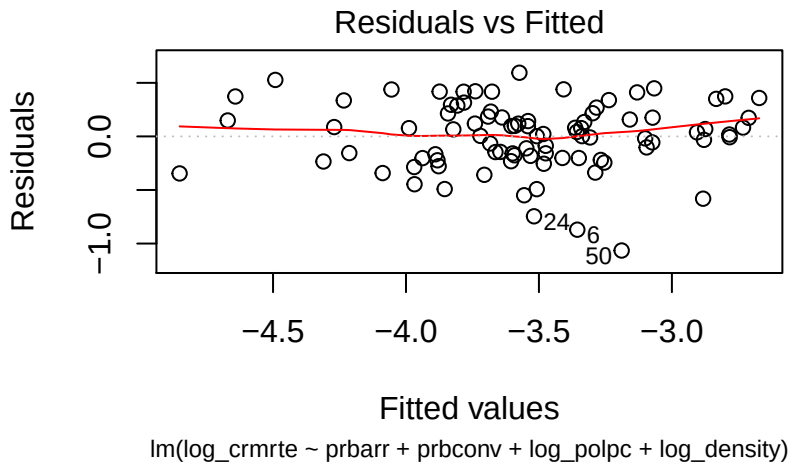
```
##      prbarr     prbconv   log_polpc log_density
##    1.217343    1.016318    1.061591    1.166073
```

Based on pairwise correlation in the independent variables prbarr, prbconv, log_polpc, and log_density, and the fact that no variance inflation factor nears 10, we do not detect evidence of multicollinearity negatively impacting our specification.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 4: Zero conditional mean

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



From the residuals vs fitted plot, the residuals appear largely centered on 0.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 5: Homoscedasticity

It is not easy to determine Homoscedasticity from the residuals vs fitted values plot alone, so we run some additional tests.

```
bptest(model1)
```

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

```
ncvTest(model1)
```

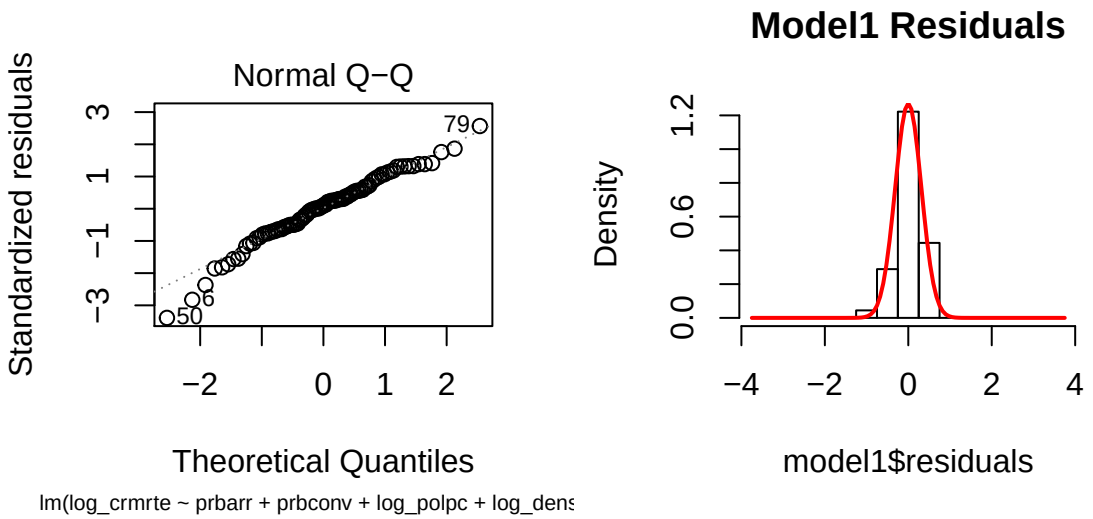
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0004349806, Df = 1, p = 0.98336
```

For both tests, we fail to reject the null hypothesis of homoscedasticity.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 6: Normality of Residuals

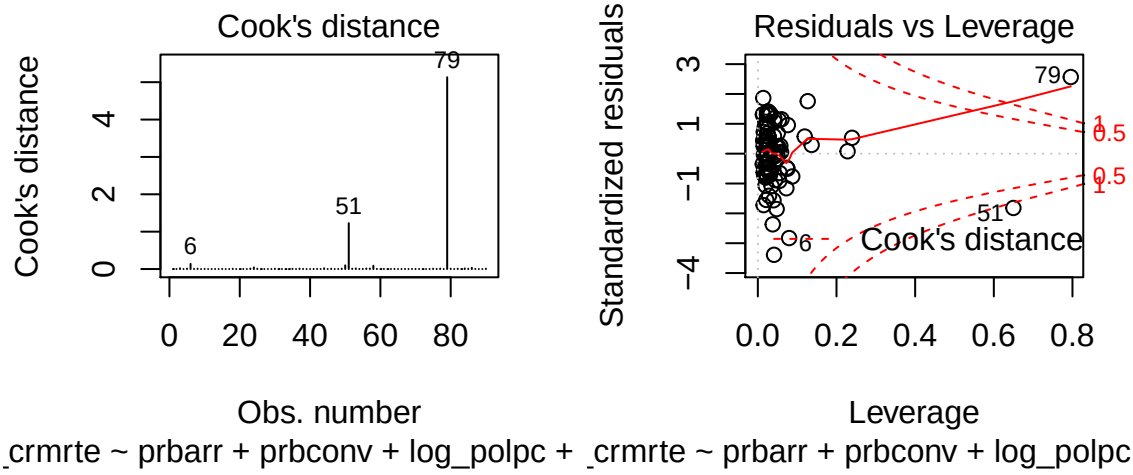
```
plot(model1, which=2, cex.sub=0.66)
hist(model1$residuals,main="Model1 Residuals", breaks = seq(-3.75, 3.75, 0.5), freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(model1$residuals)), col="red", lwd=2, add=TRUE)
```



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:

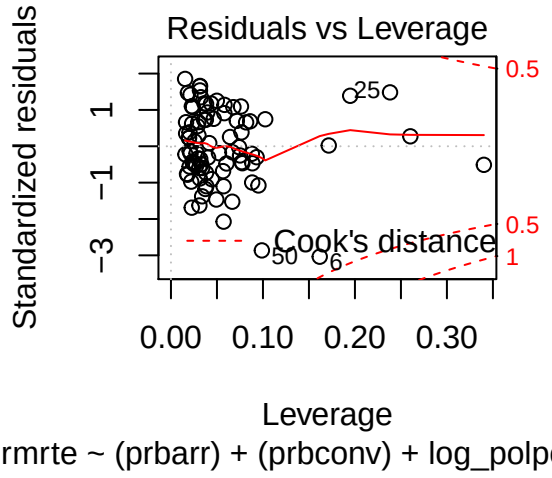
```
plot(model1, which=4)
plot(model1, which=5)
```



There are some influential values 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 - Density and polpc are quite low for 51 and 79, while the crime rate and probability of arrest are similar to other counties. Of note, these are both 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 ~ 0.55%",
                "For ~ 1 unit increase (100 people per square mile) in density,
                crime rate increases by ~ 0.111%")
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 ~ 0.55%
log_density	0.1113	For ~ 1 unit increase (100 people per square mile) in density, crime rate increases by ~ 0.111%

Most 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 1 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 variation in crime rate is explained by this model

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

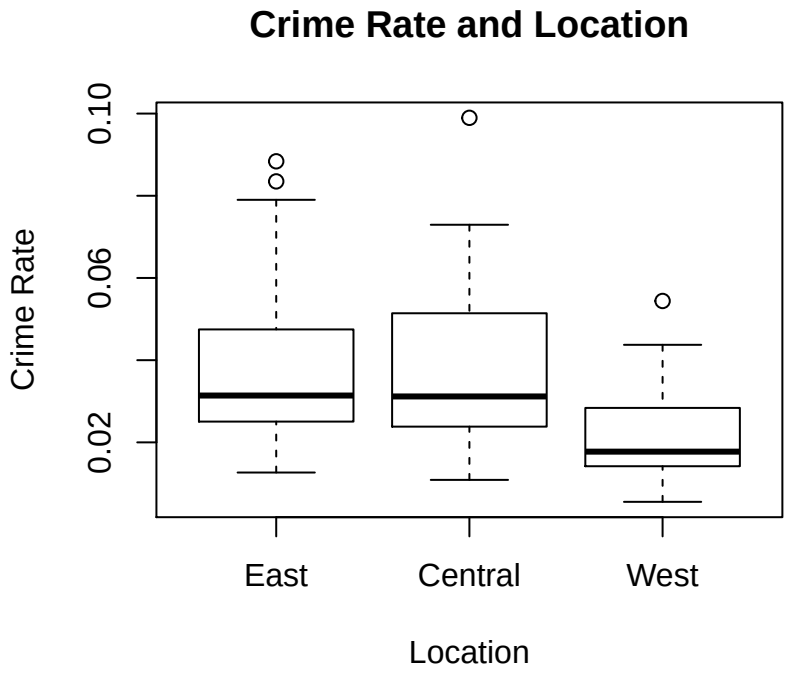
Model 2 - Optimal Specification

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 as seen in the below table, western counties have a higher concentration of police, but a lower crime rate on average. This is opposite to what we have observed in our #Model1.

```
boxplot(crime$crmrte[crime$east],crime$crmrte[crime$central],
crime$crmrte[crime$west], names=c("East","Central","West"),
main="Crime Rate and Location", xlab ="Location",ylab="Crime Rate")


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", "Mean Density", "Mean Police per Cap", "Mean 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	Mean Density	Mean Police per Cap	Mean 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, a relationship which is opposite to that observed in other regions and the overall model.

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

```
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
```

Intuition of the Effects of County 71 Existing in Both Central and West

This new model removes county 71 as the dataset we were provided show it is in both Western and Central North Carolina, this will allow us to understand the impacts of this error 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 small change in the coefficient of west, -1.81 - (-2.09) = 0.3 when county 71 is removed. This change of ~0.3 represents bias in west if county 71 is in fact in central NC.

Intuition of the effects of potential data issues in county 115 on the 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.

```
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 deosn't appear to have a very high impact on our proposed model. It is possible that adding a control for west and the interaction between west and polpc has mitigated this broader influence.

Testing the Validity of the 6 CLM assumptions

CLM 1: Linear model

Our views are identical to the previous model.

CLM 2: Random Sampling

Our views are identical to the previous model

CLM 3: Multicollinearity

```
cor_model2 <- data.matrix(subset(crime,
           select = c("prbarr","prbconv", "polpc", "density",
                      "log_polpc", "log_density", "pctmin80", "west")))
cor_model2_out = round(cor(cor_model2), 4)
kable(cor_model2_out)
```

	prbarr	prbconv	polpc	density	log_polpc	log_density	pctmin80	west
prbarr	1.0000	-0.0558	0.4260	-0.3027	0.2162	-0.3479	0.0491	0.1737
prbconv	-0.0558	1.0000	0.1719	-0.2267	-0.0076	-0.0861	0.0625	0.0473
polpc	0.4260	0.1719	1.0000	0.1591	0.9058	-0.0563	-0.1691	0.1514
density	-0.3027	-0.2267	0.1591	1.0000	0.3250	0.5785	-0.0746	-0.1358
log_polpc	0.2162	-0.0076	0.9058	0.3250	1.0000	0.0230	-0.1456	0.1088
log_density	-0.3479	-0.0861	-0.0563	0.5785	0.0230	1.0000	-0.0967	-0.2149
pctmin80	0.0491	0.0625	-0.1691	-0.0746	-0.1456	-0.0967	1.0000	-0.6336
west	0.1737	0.0473	0.1514	-0.1358	0.1088	-0.2149	-0.6336	1.0000

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

```
vif(model2)
```

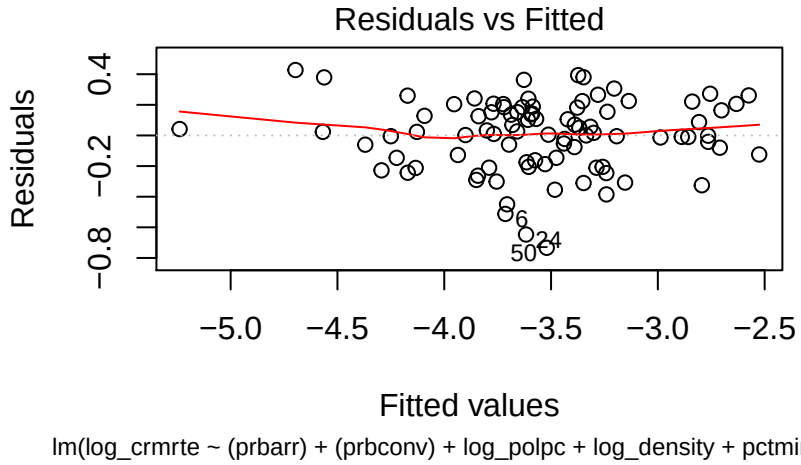
```
##          prbarr          prbconv       log_polpc     log_density         pctmin80
##        1.538158        1.086812        1.893781        1.364137        1.946049
##            west log_polpc:west
##      350.110767      337.569876
```

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.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 4: Zero conditional mean

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



From the residuals vs fitted plot, the residuals are largely centered on 0 except few values. **We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 5: Homoscedasticity

It is not easy to determine Homoscedasticity from the residuals vs fitted values plot alone, so we run some additional tests.

```
bptest(model2)
```

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

```
ncvTest(model2)
```

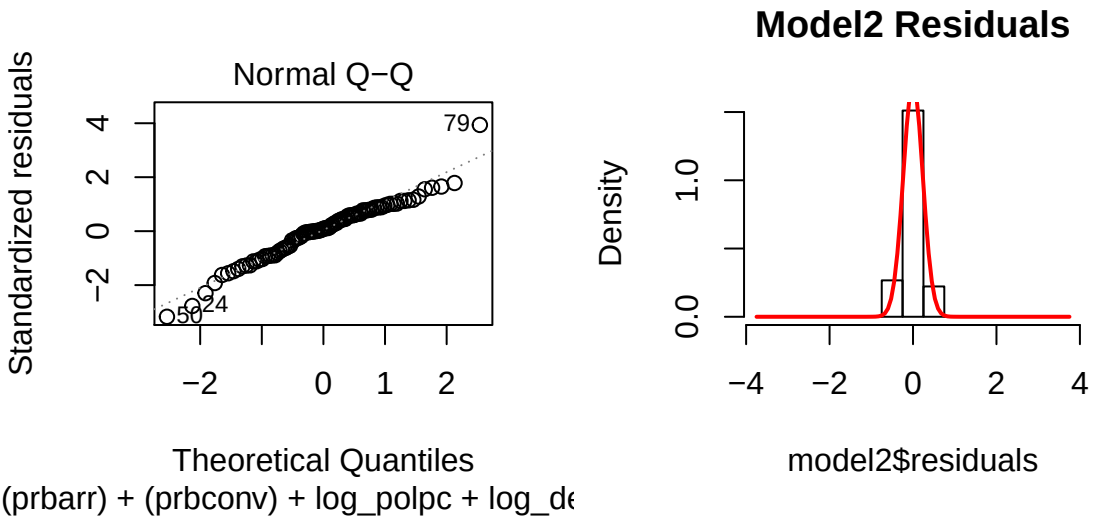
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2935739, Df = 1, p = 0.58794
```

For both tests, we fail to reject the null hypothesis of homoscedasticity.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 6: Normality of Residuals

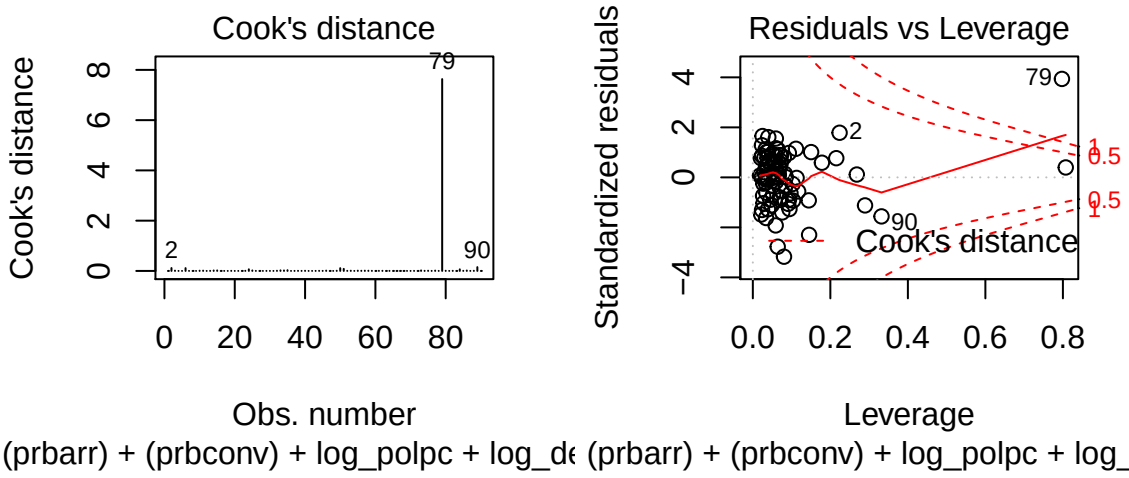
```
plot(model2, which=2)
hist(model2$residuals,main="Model2 Residuals", breaks = seq(-3.75, 3.75, 0.5), freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(model2$residuals)), col="red", lwd=2, add=TRUE)
```



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. **We have not uncovered evidence that would lead us to question the validity of assumption.*

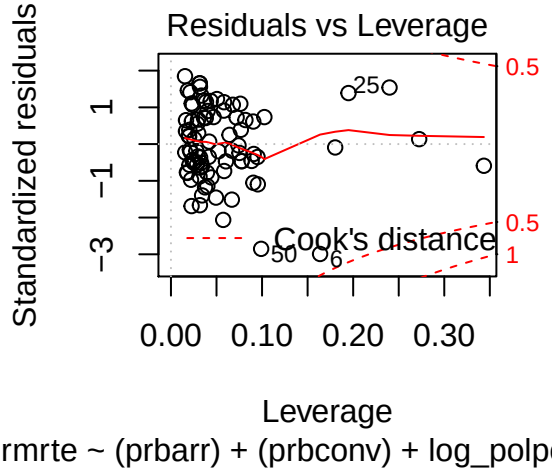
Cook's Distance:

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



Observation 79 is influential. 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)
```




```
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 ~ 0.73%",
                "For ~ 1 unit increase (100 people per square mile) in density,
                crime rate increases by ~ 0.104%",
                "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 ~ 0.73%
log_density	0.1044	For ~ 1 unit increase (100 people per square mile) in density, crime rate increases by ~ 0.104%
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%

Most of the coefficients are highly statistically significant except density when we look at heteroskedastic-robust errors: This indicates that, after controlling for the impact of being a western county, density doesn't have a statistically significant impact - thus reflecting a bias in our previous model specification.

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 variation of the dependent variable is 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.

Model 3 - Sub-optimal Specification

In addition to the variables included in model 2, we analyzed the effects of percentage of young males and tax revenue per capita on crime rate. Next, based on the analysis of wages above, we include wcon and wfed in our model due to their strong explainatory power as an example why they and other wages were not included in our optimal specification in model 2. We are not using any interaction terms here and want to check how crime rate varies across all regions for these variables.

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

```
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
```

Intuition of the effects of potential data issues in county 115 on the model

County 115 has high prbarr and low crime and low polpc and very low density this new model will allow us to see the effects on the model above.

```
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 above two coefficents summaries, we can observe that the county 115 doesn't have any big impact on our coefficients.

Testing the Validity of the 6 CLM assumptions

CLM 1: Linear model

Our views are identical to the previous model.

CLM 2: Random Sampling

Our views are identical to the previous model.

CLM 3: Multicollinearity

```
cor_model3 <- data.matrix(subset(crime,
 select = c("prbarr", "prbconv", "log_polpc", "log_density", "pctmin80",
            "log_pctymle", "log_taxpc", "wcon", "wfed")))
cor_model3_out = round(cor(cor_model3), 3)
kable(cor_model3_out)
```

	prbarr	prbconv	log_polpc	log_density	pctmin80	log_pctymle	log_taxpc	wcon	wfed
prbarr	1.000	-0.056	0.216	-0.348	0.049	-0.207	-0.131	-0.252	-0.208
prbconv	-0.056	1.000	-0.008	-0.086	0.062	-0.185	-0.133	-0.117	-0.061
log_polpc	0.216	-0.008	1.000	0.023	-0.146	0.176	0.383	0.103	0.314
log_density	-0.348	-0.086	0.023	1.000	-0.097	0.175	0.108	0.371	0.525
pctmin80	0.049	0.062	-0.146	-0.097	1.000	-0.012	0.029	-0.108	0.031
log_pctymle	-0.207	-0.185	0.176	0.175	-0.012	1.000	-0.074	0.008	0.004
log_taxpc	-0.131	-0.133	0.383	0.108	0.029	-0.074	1.000	0.271	0.102
wcon	-0.252	-0.117	0.103	0.371	-0.108	0.008	0.271	1.000	0.507
wfed	-0.208	-0.061	0.314	0.525	0.031	0.004	0.102	0.507	1.000

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 0.5066639 along with polpc and density. We examine how these variables affect 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 these variables are jointly significant. But the coefficients of wcon and wfed are close to zero.

```
vif(model3)
```

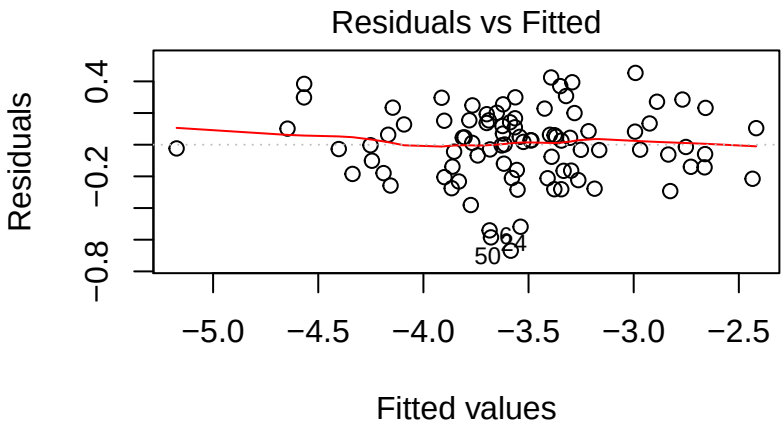
```
##      prbarr     prbconv   log_polpc log_density    pctmin80 log_pctymle
##    1.530657    1.146751    2.055567    1.659918    1.158265    1.391718
##   log_taxpc        wcon        wfed
##    1.622171    1.564755    2.273048
```

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

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 4: Zero conditional mean

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



From the residuals vs fitted plot, the residuals are largely centered on 0.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

CLM 5: Homoscedasticity

It is not easy to determine Homoscedasticity from the residuals vs fitted values plot alone, so we run some additional tests.

```
bptest(model3)
```

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

```
ncvTest(model3)
```

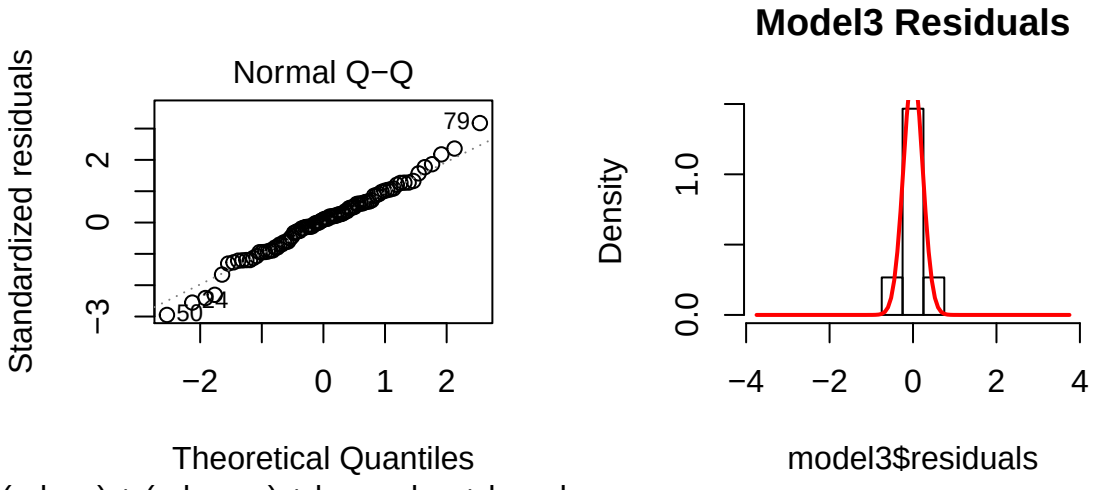
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.002576534, Df = 1, p = 0.95952
```

For both tests, we fail to reject the null hypothesis of homoscedasticity.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

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)
```

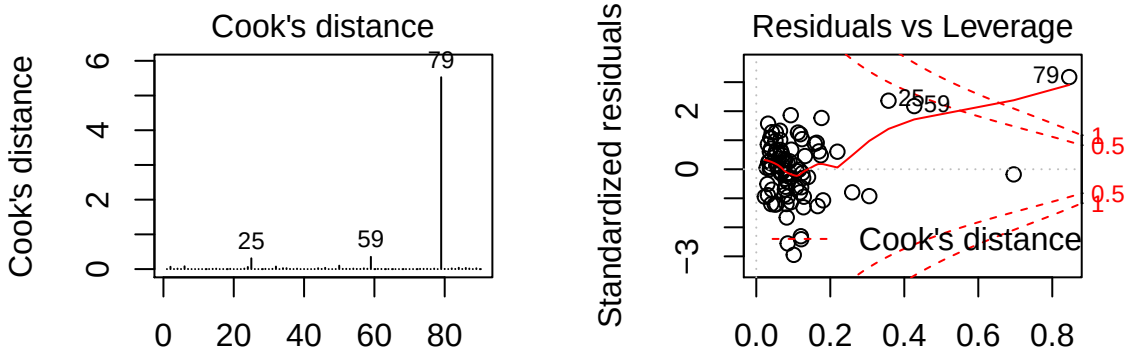


Other than a few outliers, the distribution is relatively normal for our given sample size. We do see some outliers in the Q-Q plot indicating that there is some skew because of the outlier values at the ends.

**We have not uncovered evidence that would lead us to question the validity of assumption.*

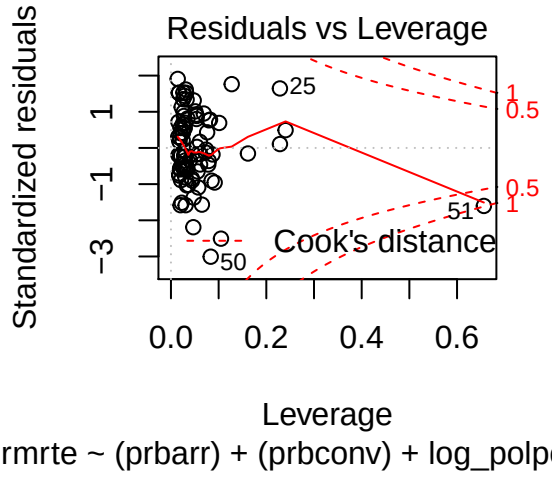
Cook's Distance:

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



There are some influential values - values 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 59 - This is an eastern county with low crime rate and low polpc, differing from many other observations in our model. We do not think that this observation merits exclusion.

```
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 0.562%",
"For ~ 1 unit increase (100 people per square mile)
in density,crime rate increases by 0.09%",
"For ~ 1 unit increase of percent minority, crime rate increases by ~ 1.12%",
"For ~ 1% increase in pctymle, crime rate increases by ~ 0.07%",
"For ~ 1% increase in taxpc, crime rate increases by ~ 0.009%",
"For ~ 1 unit increase in wcon, crime rate increases by ~ .01%",
"For ~ 1 unit increase in wfed, crime rate increases by ~ 0.01%")
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 0.562%
log_density	0.0905	For ~ 1 unit increase (100 people per square mile) in density,crime rate increases by 0.09%
pctmin80	0.0119	For ~ 1 unit increase of percent minority, crime rate increases by ~ 1.12%
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 ~ 0.009%
wcon	0.0010	For ~ 1 unit increase in wcon, crime rate increases by ~ .01%
wfed	0.0010	For ~ 1 unit increase in wfed, crime rate increases by ~ 0.01%

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 don'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
```

Model3 conclusion -

Probability of arrest is consistently estimated to have an effect on crime rate.

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

The adjusted R^2 for this model is 0.8096141% which means a lot of the the variation in the dependent variable is explained by this model. That said, we have hit the point of diminishing returns from adding more variables as these values are quite close to those for the more parsimoneous model 2.

We run a Wald Test to see if these additional variables has any signifiance.

```
waldtest(model3, model1, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: log_crmrte ~ (prbarr) + (prbconv) + log_polpc + log_density +
##     pctmin80 + log_pctymle + log_taxpc + wcon + wfed
## Model 2: log_crmrte ~ prbarr + prbconv + log_polpc + log_density
##   Res.Df Df      F    Pr(>F)
## 1     80
## 2     85 -5 12.904 3.314e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wald test reveals that variables added to model 3 jointly have no effect on crime rate.

Now let's compare the coefficients of all models and also adding standard errors to the models:

```
se.model1 <- sqrt(diag(vcovHC(model1)))
se.model2 <- sqrt(diag(vcovHC(model2)))
se.model3 <- sqrt(diag(vcovHC(model3)))
```

```
stargazer(model1,model2,model3,
  title="Linear Models Predicting Crime Rate",
  column.labels=c("Model 1","Model 2","Model 3"),
  se = list(se.model1, se.model2, se.model3),
  single.row=TRUE,
  star.cutoffs = c(0.05, 0.01, 0.001),
  dep.var.caption  = "Measuring Crime Rate",
  keep.stat = c("adj.rsq","n"),
  dep.var.labels   = "Crime Rate",
  no.space = TRUE
)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Dec 09, 2018 - 6:58:19 PM

Table 3: Linear Models Predicting Crime Rate

	Measuring Crime Rate		
	Crime Rate		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
prbarr	−1.924** (0.615)	−1.821*** (0.383)	−1.878*** (0.326)
prbconv	−0.692*** (0.158)	−0.690*** (0.142)	−0.700*** (0.117)
log_polpc	0.556** (0.188)	0.736*** (0.199)	0.562*** (0.153)
log_density	0.111 (0.130)	0.104 (0.152)	0.091 (0.160)
pctmin80		0.010*** (0.002)	0.012*** (0.002)
west		−1.810 (1.283)	
log_polpc:west		−0.260 (0.196)	
log_pctymle			0.074 (0.267)
log_taxpc			0.010 (0.156)
wcon			0.001 (0.001)
wfed			0.001 (0.001)
Constant	1.008 (1.260)	1.919 (1.409)	0.186 (1.391)
Observations	90	90	90
Adjusted R^2	0.658	0.807	0.810
Note:			*p<0.05; **p<0.01; ***p<0.001

This table demonstrates the following:

- The key Variables `prbarr` and `prbconv` have robust estimates in all models. Co-efficient of `prbarr` is always further from zero than `prbconv`.
- The coefficient for `density` decreases as we add more variables that likely interact with it. Eg. `west`. From the correlation matrix, we do see west is correlated with density.
- `polpc` maintains its statistically significant coefficient with low standard error and low coefficient in model 3.
- `pctmin80` is also robust as its coefficient stays statistically significant in models 2 and 3.
- All other variables in model3 are not statistically significant. We can conclude these variables are not robust enough to introduce them in our main model.

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. 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 limited 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:

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

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 ratio of arrests to crimes, and 3) the ratio 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 reinforced 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 education 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 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.

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.