Lab 3: Reducing Crime (DRAFT: Stage 1)

N. Akkineni, A. Thorp, K. Hanna

November 27, 2018

Contents

Introduction (Stage 1: Draft Report)

Our team has been hired to provide research for a local political campaign to help the campaign 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))

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

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

Exploratory Data Analysis

Data Summary

We were provided with a dataset that included crime statistics from the North Carolina Department of Corrections prison and probation files, demographic statistic taken from census, police data computed using 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.

Data Clean Up

Null Rows

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

```
# Delete the 6 empty observations at the end, including the row with the apostrophe.
# We can use complete.cases to do this as these 6 observations are the only incomplete observations.
crime = crime[complete.cases(crime), ]

# Fix prbconv which is a factor rather than numeric due to the apostrophe
# Convert from factor to numeric
crime$prbconv = as.numeric(as.character(crime$prbconv))
```

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

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

Concerns about 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 be a coding error, it is a ratio and not a probability.

prbconv (Probability of Conviction)

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

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.

- 29 - Camden County
- 31 - Carteret County
- 43 - Clay County
- 73 - Gates County
- 75 - Graham County
- 97 - Iredell County
- 103 - Jones County
- 121 - Mitchell County
- 177 - Tyrrell County
- 199 - 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.

Univariate Analysis

```
quick_uni_analysis = function(variable, description, roundto = 8) {
  hist(variable, xlab = paste(tools::toTitleCase(description),
        paste('\n Shapiro:',
        round(as.numeric(shapiro.test(variable)[2]), roundto)
        )), main = "")

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

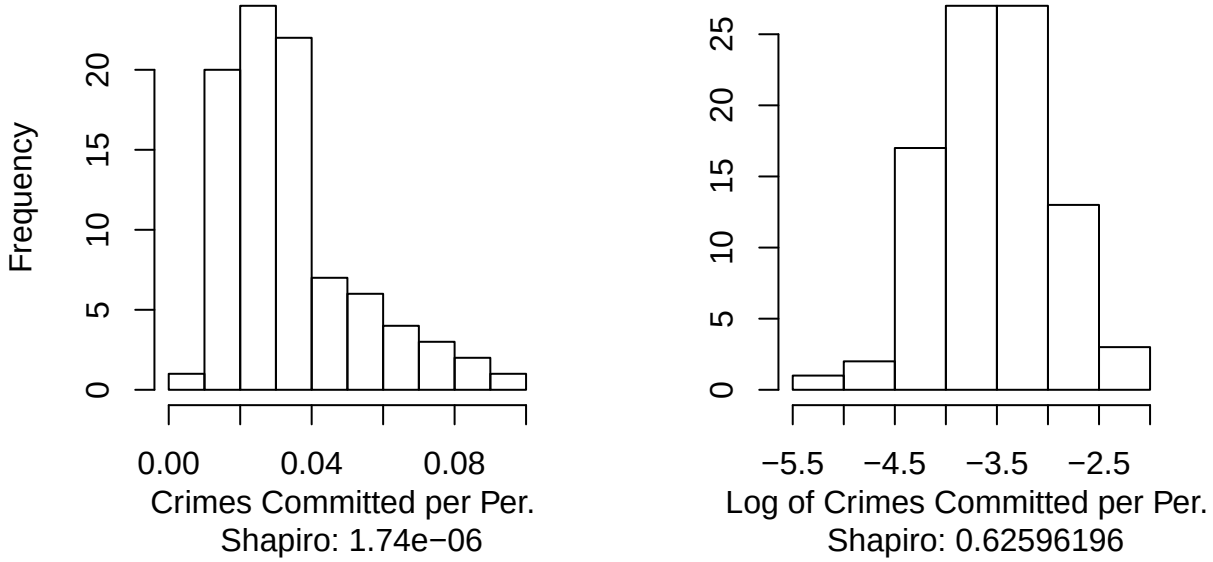
Key Variable

Crimes Committed Per Person

Campaign Significance: Crime rate is a politicized and effects economy

This is the key variable we will be explaining in terms of other variables in most of our modeling.

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



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.

Explanatory

Diagrams of Key Variables With and Without Log Transformations

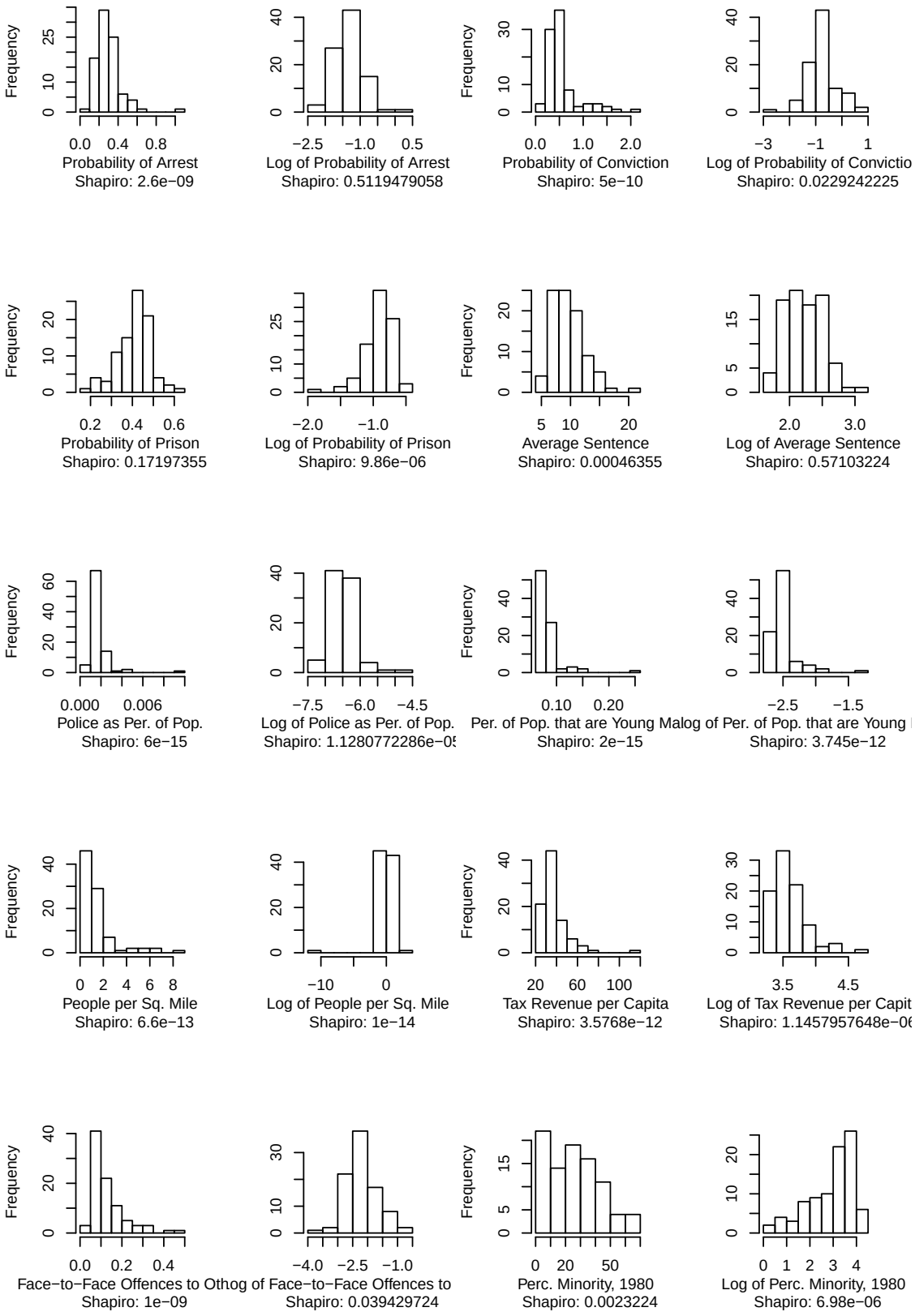
```
par(mfrow=c(5,4))
quick_uni_analysis(crime$prbarr, 'Probability of Arrest', roundto = 10)
quick_uni_analysis(crime$prbconv, 'Probability of Conviction', roundto = 10)

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

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

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

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


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
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 transformed variable is preferable for modelling.

Probability of Conviction

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

Probability of Prison

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

Average Sentence

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

Police as a Percentage of Population

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

The number of police can conceivably be either the cause of or result of crime rates. However, there is a fair amount of research showing increasing the police workforce causes a reduction in crime according to Vollaard & Hamad [(1)] . Without timeseries data including changes to Police as a Percentage of Population we cannon determine the true effects. However, many studies show that increasing the size of a police force does reduce crime.

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

[(1)] The Journal of Law & Economic https://www.jstor.org/stable/10.1086/666614

Percentage of Population That Are Young Males

This variable has a strong positive skew, using a natural log transformation results in a still skewwed distirbution, however, it is closer to normal.

People per Square Mile

———— ** Naga and Alex ** ————

This variable becomes less normal after log transformation, shifts skew to the negative, and more extreme. I don't think we should be using the log of this. We can change this before handing it in, after, or not at all. I'll update this later before handing in.

Tax Revenue per Capita

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

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

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

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

The transformed variable is preferable for modelling.

Percentage Minority, 1980

Model Analysis

Transforming the variables and storing them in the data frame

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

Model1 - Minimum Specification

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

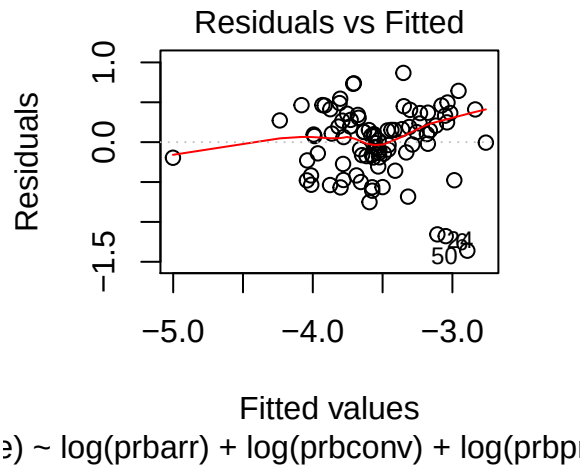
$$log(crmrte) = \beta_0 + \beta_1 log(prbarr) + \beta_2 log(prbconv) + \beta_3 log(prbpris) + \beta_4 log(avgsen) + u$$

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

####Checking if our Multiple Linear Regression Assumptions are valid in the model-

####Assumption - Linear model

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



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

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

####Assumption - Multicollinearity

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

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

We are not seeing any obvious signs of multicollinearity. Running some additional tests.

```
vif(model1)
```

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

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

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

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

```
bptest(model1)
```

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

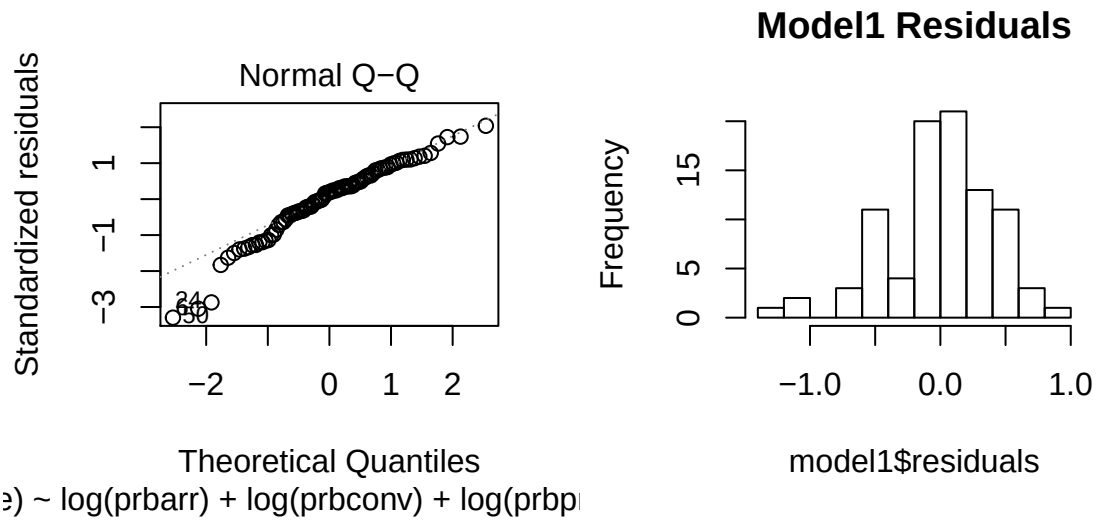
```
ncvTest(model1)
```

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

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

####Normality of Residuals

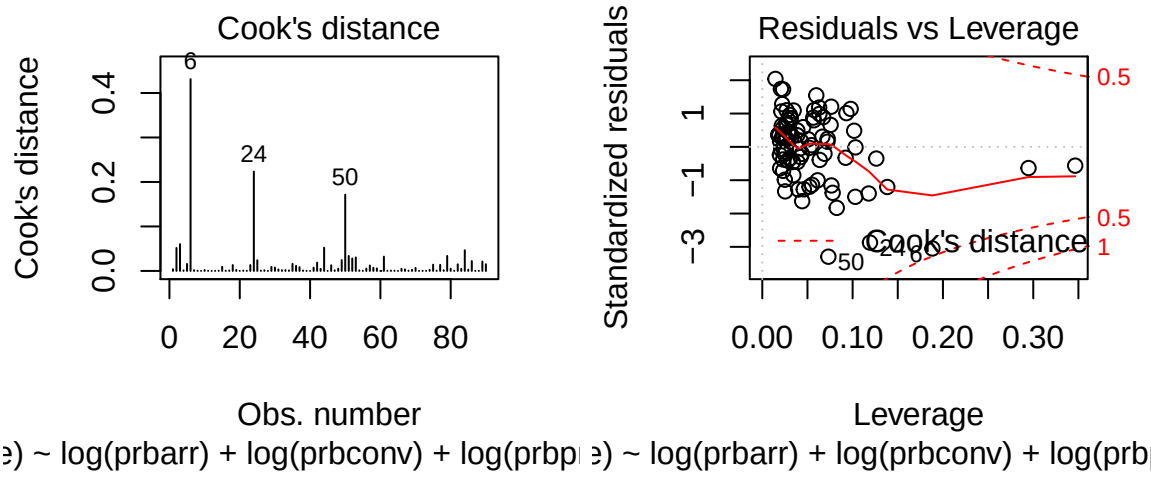
```
plot(model1, which=2)
hist(model1$residuals,main="Model1 Residuals")
```



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

####Cook's Distance:

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



There are some influential values however cook's distance is within the bounds. If we remove the outliers, we can see a very improved R-Square.

Calculating AIC

```
AIC(model1)
```

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

The AIC for this model is 110.0643

Model2 - Optimal Specification

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

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

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

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

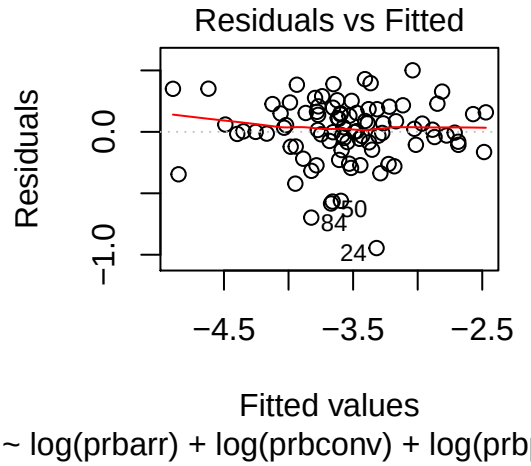
$$log(crmrte) = \beta_0 + \beta_1 log(prbarr) + \beta_2 log(prbconv) + \beta_3 log(prbpris) + \beta_4 log(avgsen) + \beta_5 log(polpc) + \beta_6 log(taxpc) + \beta_7 log(density) + \beta_8 log(pctymle) + \beta_9 log(pctmin80) + u$$

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

Checking if our Assumptions are valid in the model-

####Assumption - Linear model

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



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

####Assumption - Random Sampling As we are using the same data set - From Model1, it is a VALID Assumption.

####Assumption - Multicollinearity

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

```
##               log_crmrte   log_prbarr log_prbconv log_prbpris   log_avgsen
## log_crmrte    1.00000000 -0.435753889 -0.37249610  0.06960729  0.023417173
## log_prbarr   -0.43575389  1.000000000 -0.20235541 -0.02801506  0.003832922
## log_prbconv  -0.37249610 -0.202355412  1.00000000  0.01053523  0.011629954
## log_prbpris   0.06960729 -0.028015058  0.01053523  1.00000000 -0.123723380
## log_avgsen    0.02341717  0.003832922  0.01162995 -0.12372338  1.000000000
## log_polpc     0.28453961  0.054242849 -0.13863932 -0.01046407  0.395079641
## log_taxpc     0.33984322 -0.111996061 -0.20622995 -0.05671890  0.077900054
## log_density   0.49364251 -0.354779226 -0.07258589  0.41571758  0.159651910
## log_pctymle   0.31175403 -0.282255476 -0.22382225 -0.02486072  0.103327744
## log_pctmin80  0.39729097 -0.051223276  0.07171367  0.02897773 -0.164025122
##                log_polpc   log_taxpc log_density  log_pctymle log_pctmin80
## log_crmrte    0.28453961  0.33984322  0.49364251  0.311754030  0.397290965
## log_prbarr    0.05424285 -0.11199606 -0.35477923 -0.282255476 -0.051223276
## log_prbconv  -0.13863932 -0.20622995 -0.07258589 -0.223822249  0.071713668
## log_prbpris  -0.01046407 -0.05671890  0.41571758 -0.024860719  0.028977726
## log_avgsen    0.39507964  0.07790005  0.15965191  0.103327744 -0.164025122
## log_polpc     1.00000000  0.38298724  0.02304206  0.176404361 -0.161898455
## log_taxpc     0.38298724  1.00000000  0.10798692 -0.073609432  0.121027860
## log_density   0.02304206  0.10798692  1.00000000  0.175156111 -0.018819680
## log_pctymle   0.17640436 -0.07360943  0.17515611  1.000000000  0.008048105
## log_pctmin80 -0.16189846  0.12102786 -0.01881968  0.008048105  1.000000000
```

We are not seeing any obvious signs of multicollinearity. Runnig some additional tests

```
vif(model2)
```

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

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

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

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

```
bptest(model2)
```

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

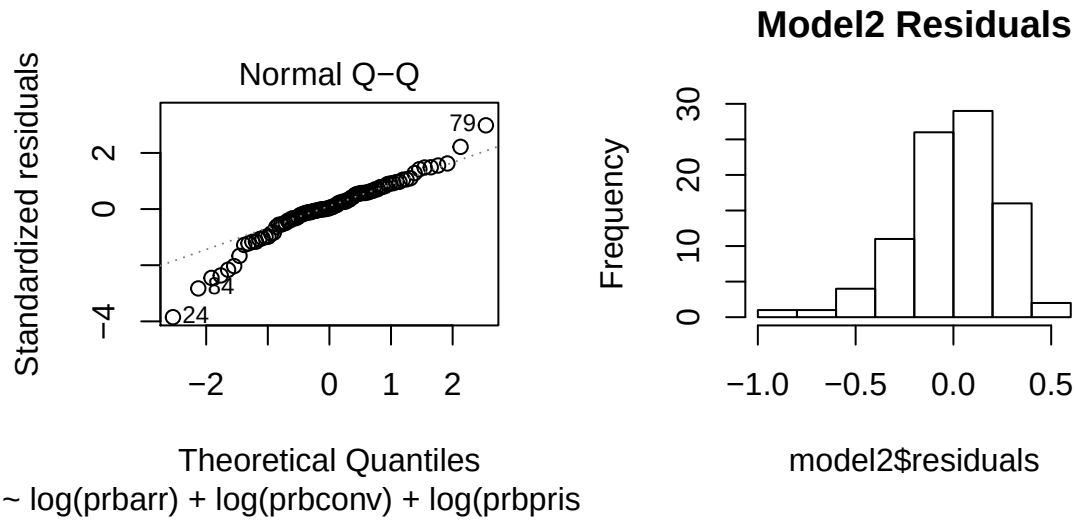
```
ncvTest(model2)
```

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

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

####Normality of Residuals

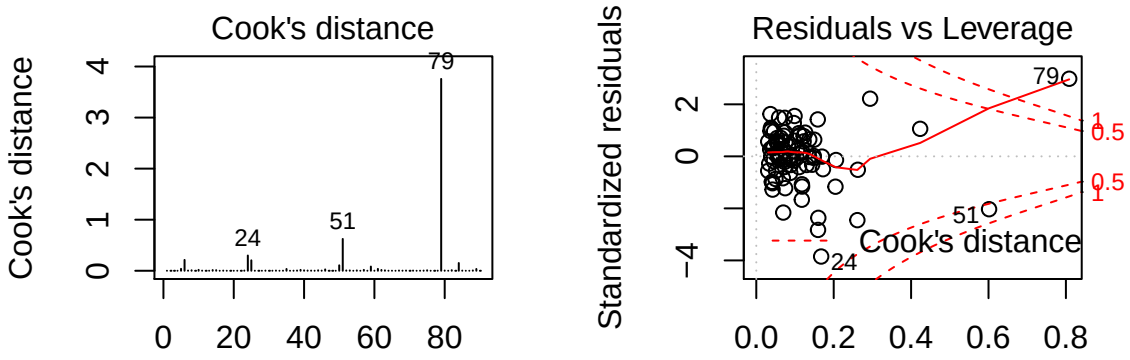
```
plot(model2, which=2)
hist(model2$residuals,main="Model2 Residuals")
```



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

####Cook's Distance:

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



There are some influential values [24,51,79] however cook's distance is within the bounds

```
AIC(model2)
```

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

The AIC for this model is 30.50572

Model3 - Optimal-2 Specification [Best-Fit model]

After obsercing Model2 - We found that tax and percent male doesn't influence crime extensively. So, removing those in our Best-Fit model

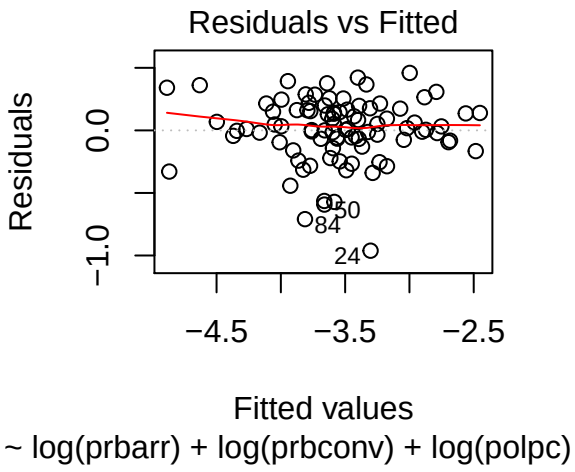
$$log(crmrte) = \beta_0 + \beta_1 log(prbarr) + \beta_2 log(prbconv) + \beta_3 log(prbpris) + \beta_4 log(avgsen) + \beta_5 log(polpc) + \beta_6 log(density) + \beta_7 log(pctmin80) + u$$

```
model3 = lm(log(crmrte) ~ log(prbarr) + log(prbconv)+ log(polpc) +
              log(density)  + log(pctmin80) +log(prbpris) + log(avgsen) , data = crime)
```

Checking if our Assumptions are valid in the model-

####Assumption - Linear model

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



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

####Assumption - Random Sampling As we are using the same data set - From Model1, it is a VALID Assumption.

####Assumption - Multicollinearity

```
cor_model3 <- data.matrix(subset(crime,
                                 select = c("log_crmrte","log_prbarr","log_prbconv",
                                            "log_avgsen","log_polpc","log_density","log_pctmin80")))
cor(cor_model3)
```

```
##               log_crmrte   log_prbarr log_prbconv   log_avgsen   log_polpc
## log_crmrte    1.00000000 -0.435753889 -0.37249610  0.023417173  0.28453961
## log_prbarr   -0.43575389  1.000000000 -0.20235541  0.003832922  0.05424285
## log_prbconv  -0.37249610 -0.202355412  1.00000000  0.011629954 -0.13863932
## log_avgsen    0.02341717  0.003832922  0.01162995  1.000000000  0.39507964
## log_polpc     0.28453961  0.054242849 -0.13863932  0.395079641  1.00000000
## log_density   0.49364251 -0.354779226 -0.07258589  0.159651910  0.02304206
## log_pctmin80  0.39729097 -0.051223276  0.07171367 -0.164025122 -0.16189846
##               log_density log_pctmin80
## log_crmrte    0.49364251   0.39729097
## log_prbarr   -0.35477923  -0.05122328
## log_prbconv  -0.07258589   0.07171367
## log_avgsen    0.15965191  -0.16402512
## log_polpc     0.02304206  -0.16189846
## log_density   1.00000000  -0.01881968
## log_pctmin80 -0.01881968   1.00000000
```

We are not seeing any obvious signs of multicollinearity. Running some additional tests

```
vif(model3)
```

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

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

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

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

```
bptest(model3)
```

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

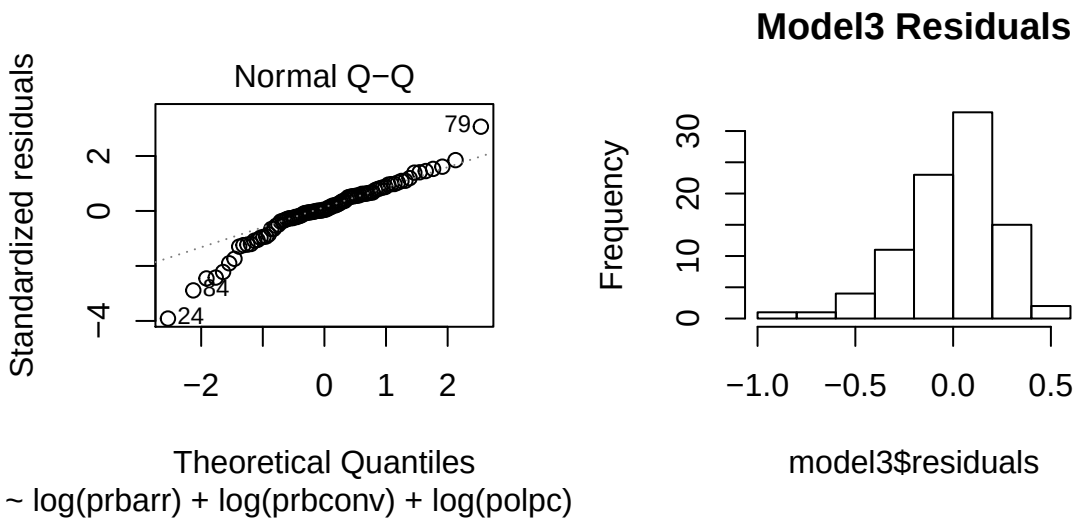
```
ncvTest(model3)
```

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

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

####Normality of Residuals

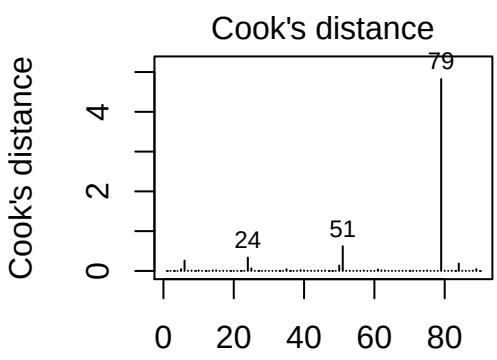
```
plot(model3, which=2)
hist(model3$residuals,main="Model3 Residuals")
```

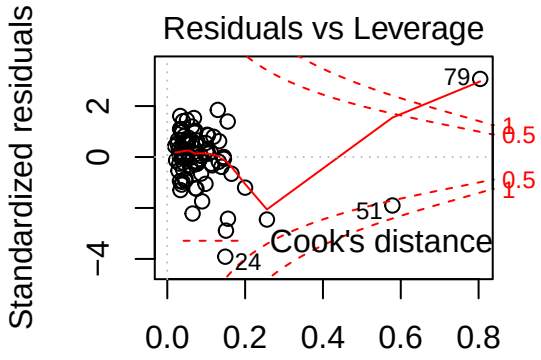


Other than a few outliers, the distribution is relatively normal for our given sample size. So, it is a VALID assumption.

####Cook's Distance:

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





There are some influential values [6,51,79] however cook's distance is within the bounds If we remove the outliers, we see a better and improved R-Square value ####Calculating AIC

```
AIC(model3)
```

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

The AIC for this model is 26.92602

Model4 - Using all variables Specification

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

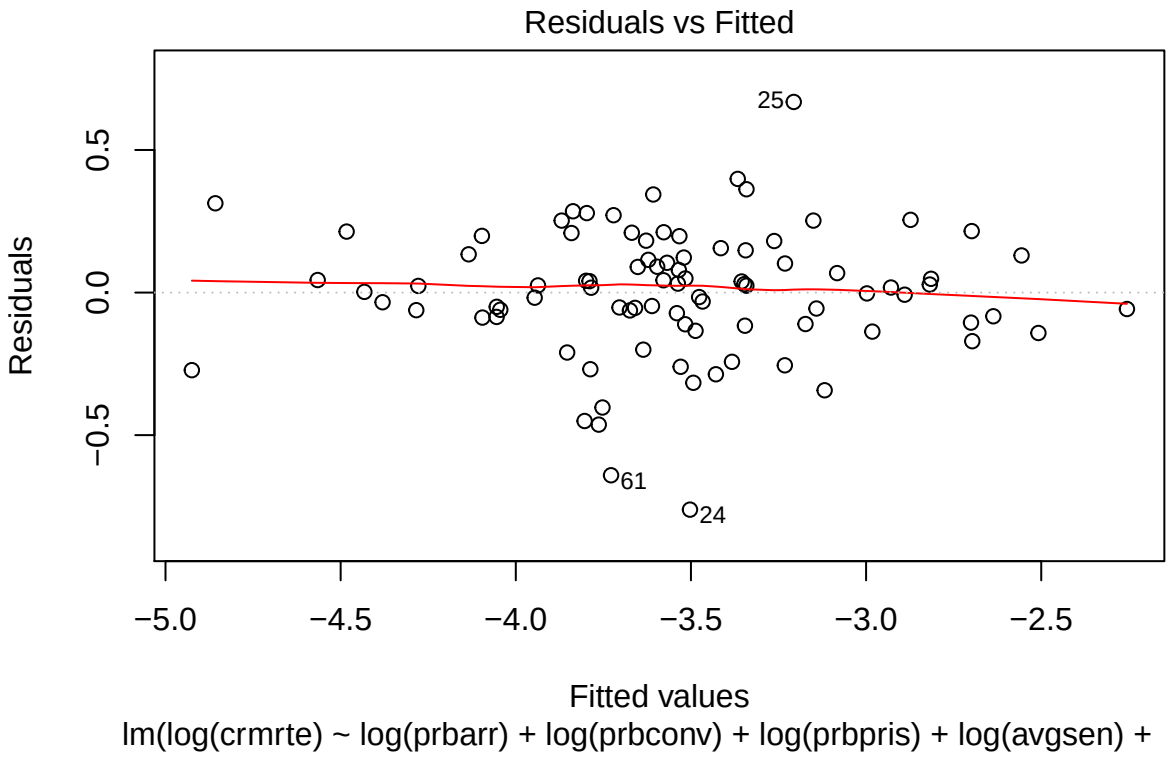
$$
\begin{aligned}
log(crmrte) = \beta_0 + \beta_1 log(prbarr) + \beta_2 log(prbconv) + \beta_3 log(prbpris) + \\
\beta_4 log(avgsen) + \beta_5 log(polpc) + \beta_6 log(taxpc) + \beta_7 log(west) + \\
\beta_8 log(central) + \beta_9 log(urban) + \beta_1 0 log(pctmin80) + \beta_1 1 log(wcon) + \beta_1 2 og(wtuc) + \\
\beta_1 3 log(wtrd) + \beta_1 4 log(wfir) + \beta_1 5 log(wser) + \beta_1 6 log(wmfg) + \beta_1 7 log(wfed) + \\
\beta_1 8 log(wsta) + \beta_1 9 log(wloc) + \beta_2 0 log(mix) + \beta_2 1 log(density) + \beta_2 2 log(pctymle) + u
\end{aligned}
$$

```
model4 = lm(log(crmrte) ~ log(prbarr) + log(prbconv)+ log(prbpris) +
              log(avgsen)  + log(polpc) +log(density) + log(taxpc) +
              west+central + urban + log(pctmin80) + wcon + wtuc +
              wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
              log(mix) + log(pctymle), data = crime)
```

Checking if our Assumptions are valid in the model-

####Assumption - Linear model The specified model has the dependent variable linear with explanatory variables

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



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

####Assumption - Random Sampling As we are using the same data set - From Model1, it is a VALID Assumption.

####Assumption - Multicollinearity Running tests for Multicollinearity

```
vif(model4)
```

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

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

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

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

```
bptest(model4)
```

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

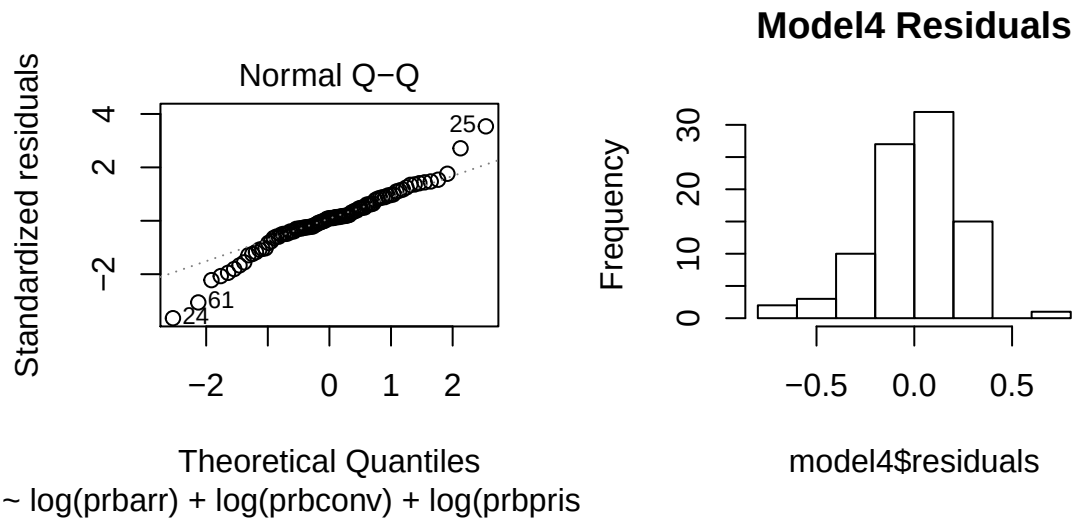
```
ncvTest(model4)
```

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

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

####Normality of Residuals

```
plot(model4, which=2)
hist(model4$residuals,main="Model4 Residuals")
```



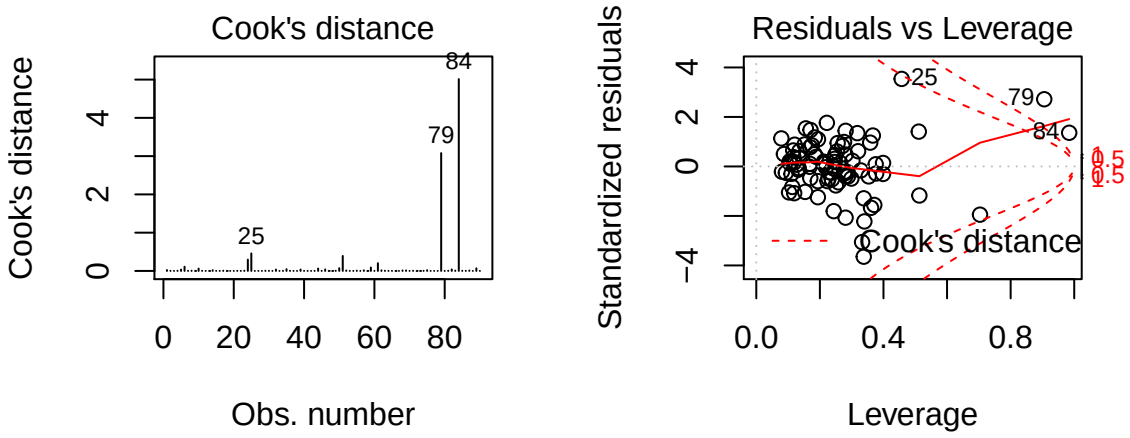
Other than a few outliers, the distribution is relatively normal for our given sample size. So, it is a VALID assumption.

####Cook's Distance:

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

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

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



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

####Calculating AIC

```
AIC(model4)
```

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

The AIC for this model is 30.78825

#Model Analysis

```
stargazer(model1, model2,model3, model4, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "n"),
          omit.table.layout = "n",
          column.labels=c("Not good","Good","Better","Using All"),
          dep.var.caption  = "Measuring Crime Rate",
          dep.var.labels   = "Crime Rate")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 26, 2018 - 12:06:40 PM

#Model1 - Only 41.7% of crime rate is being explained by our model. prbarr, prbconv has negative effect on crime. While other independent variables prbpris and avgsen has postive effect on crime. From the P-Values, crime rate fluctuates more with prbarr, prbconv

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

#Model 2 - 78.2% of crime rate is being explained by this model. Polpc, Density and Pctmin80 has a positive effect on crime. While other independent variables prbarr,prbconv,prbpris,avgsen,taxpc,pctymle has negative effect on crime. From the P-Values, crime rate fluctuates more with prbarr, prbconv, density, pctmin80 and polpc and has some impact with prbpris and avgsen

#Model 3 - 78.1% of crime rate is being explained by this model.

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

#Model 4 - 83.6% of crime rate is being explained by this model Using All variables, we got a better AIC compared to Model 1 and very close to Model 2- But it is not a best fit model

#Best-Fit Model Based on our analysis - we found that our Model3 is the best fit model with low AIC value 26.92602.

Table 1: Linear Models Predicting Crime Rate

	Measuring Crime Rate			
	Crime Rate			
	Not good	Good	Better	Using All
	(1)	(2)	(3)	(4)
log(prbarr)	−0.724	−0.522	−0.506	−0.533
log(prbconv)	−0.473	−0.403	−0.391	−0.298
log(prbpris)	0.160	−0.332	−0.320	−0.301
log(avgsen)	0.076	−0.235	−0.232	−0.308
log(polpc)		0.553	0.527	0.480
log(taxpc)		−0.068		0.018
west				0.052
central				−0.077
urban				0.071
log(density)		0.164	0.162	0.139
log(pctymle)		−0.066		0.101
log(pctmin80)		0.266	0.261	0.255
wcon				0.001
wtuc				0.0001
wtrd				0.001
wfir				−0.001
wser				−0.0004
wmfg				−0.0001
wfed				0.001
wsta				−0.00003
wloc				0.0001
log(mix)				0.062
Constant	−4.868	−1.417	−1.615	−1.984
Observations	90	90	90	90
R^2	0.416	0.784	0.783	0.835

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:

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

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 ealth inequality within a given observation, education, social cohesion, weapon availability, cost of living, or the low skill unemployment rate are operationalized within this dataset.

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

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

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

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