Lab 3, Stage 3: Reducing Crime

W203: Statistics for Data Science

Jeff Li, Vasanth Ramani, Richard Ryu

07 August 2019

Contents

Abstract

We apply data analysis, statistical methods, and linear regression to data in the goal of modeling crime in the North Carolina area. The objective is to identify key determinants of crime and potential policy prescriptions that can be applied to the issue of crime. Through our analysis, we believe that policy options of increasing the efficacy of police and prosecutorial efficacy will be the most effective factors in reducing crime in the North Carolina area.

Introduction

The consulting group Significant Effects, comprised of data scientists Jeff Li, Vasanth Ramani and Richard Ryu, has been tasked by a political campaign to identify determinants of crime. The political campaign is interested in understanding crime statistics of selected North Carolina counties, so that they can propose better policies.

Significant Effects was provided with a dataset directly from the campaign. It's important to note that this dataset is sourced from another study by Cornwell and Trumball (https://www.jstor.org/stable/2109893), redacted and modified by the campaign. Despite the data challenges, Significant Effects has been tasked to make the best out of the data.

Methodology

Our approach to investigating and modeling the data is as follows:

1) Perform exploratory data analysis, identify any anomalous attributes or mis-entered data. If necessary, transform the data so that it can be modeled.
2) Identify and create a model using the strongest predictor of the crime rate in each county, and no other covariates.
3) Identify and create a second model using what we believe to be the strongest model that provides clear determinants on crime. This model will be used to prescribe policy initiatives. We believe this model strikes a proper balance between accuracy and parsimony and reflects our best understanding of the determinants of crime.
4) Identify and create a 3rd model using all of the data that was provided by the political campaign. The purpose of this model is to demonstrate the robustness of our results to the model specification.
5) Evaluate each model according to the 6 models of the Classical Linear Model (henceforth referred to as CLM) assumptions.
6) Evaluate and compare the fit of the 3 models.
7) Make our recommendations to the political campaign based off of our analysis.

EDA

This data was provided by the political campaign. After careful examination of the dataset, we have applied several procedures to clean up this data to remove any outliers. The key transformations of this data involve:

1) Removing null values
2) Removing duplicate entries

It is important to note that removal of outlier values for prbconv (values above 100%) was considered. However, we have decided to leave them in the dataset, as it may be possible for a single arrest to result in multiple convictions

Data Preparation

We have performed extensive EDA with the data provided. For the purposes of brevity, only the code for transforming our dataset is shown below. However, for a more extensive explanation on the reasoning behind our transformations, please refer to Appendix 1: Exploratory Data Analysis.

```
# Packages used for this study
library(car)
library(stargazer)
library(dplyr)
library(ggfortify)
library(corrplot)
library(corrgram)
library(ggplot2)
library(tidyr)
library(lmtest)
library(sandwich)
library(gridExtra)

# Reading the data
A = read.csv("crime_v2.csv", header=TRUE, sep=",")

# Creating a new dataframe without the suspect rows and reformatting prbconv
# so it can be read in R properly
B = A[0:91, ]
B$prbconv <- as.numeric(as.character(B$prbconv))

# Creating new dataframe C with unique values only.
C <- unique(B) #unique values
C = select(C, -county, -year) #removing static variables that will not be used
```

Description of Variables in Dataset

We have provided a table with the description of each variable in this dataset below.

Table 1: Variables in Dataset

Variable	Description
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

Variable	Description
wfir	wkly wge, fin, ins, real est
wser	wkly wge, service industry
wmfg	wkly wge, manufacturing
wfed	wkly wge, fed employees
wsta	wkly wge, state employees
wloc	wkly wge, local gov emps
mix	offense mix: face-to-face/other
pctymle	percent young male

Data Analysis

Summary Table

We have provided a summary table of the cleaned dataset for the reader's convenience.

```
# Summary table of all data provided
stargazer(C, header= F, title = "Summary Table of Data")
```

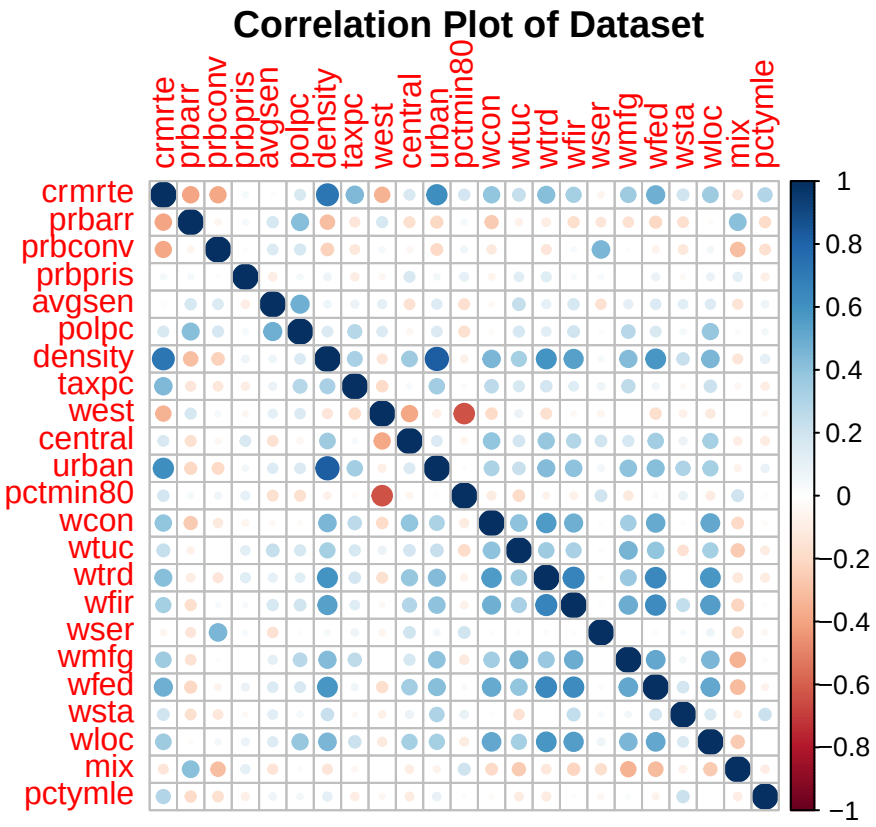
Table 2: Summary Table of Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
crmrte	90	0.034	0.019	0.006	0.021	0.040	0.099
prbarr	90	0.295	0.138	0.093	0.205	0.345	1.091
prbconv	90	0.551	0.354	0.068	0.344	0.585	2.121
prbpris	90	0.411	0.081	0.150	0.364	0.458	0.600
avgsen	90	9.689	2.834	5.380	7.375	11.465	20.700
polpc	90	0.002	0.001	0.001	0.001	0.002	0.009
density	90	1.436	1.522	0.00002	0.547	1.569	8.828
taxpc	90	38.161	13.112	25.693	30.735	41.010	119.761
west	90	0.244	0.432	0	0	0	1
central	90	0.378	0.488	0	0	1	1
urban	90	0.089	0.286	0	0	0	1
pctmin80	90	25.713	16.985	1.284	10.024	38.183	64.348
wcon	90	285.353	47.753	193.643	250.754	314.979	436.767
wtuc	90	410.907	77.355	187.617	374.331	440.679	613.226
wtrd	90	210.921	33.870	154.209	190.710	224.282	354.676
wfir	90	321.621	53.999	170.940	285.560	342.628	509.466
wser	90	275.338	207.396	133.043	229.338	277.650	2,177.068
wmfg	90	336.033	88.231	157.410	288.598	359.895	646.850
wfed	90	442.619	59.951	326.100	398.785	478.255	597.950
wsta	90	357.740	43.294	258.330	329.272	383.155	499.590
wloc	90	312.280	28.132	239.170	297.228	328.775	388.090
mix	90	0.129	0.082	0.020	0.081	0.152	0.465
pctymle	90	0.084	0.023	0.062	0.074	0.084	0.249

Correlation Plot of Dataset

Now that we have a clean dataframe, let's examine the dataset provided by the campaign with a correlation plot.

```
# Correlation plot
M <- cor(C)
corrplot(M, title = "
         Correlation Plot of Dataset", is.corr = TRUE, method = "circle",mar=c(0,0,2,0))
```



Overall, crime rate seems to have the strongest relationship with 'density' and 'urban' variable. 'Density' and 'urban' also appear to be highly correlated, which makes sense since urban areas would likely be more populated. Density appears to be slightly more correlated, so we will use this variable.

Additionally, 'prbconv', 'prbarr', 'pctmin80', and 'density', as well as some wage variables appear to all be correlated to 'crmrte', which all bear investigating in model 2.

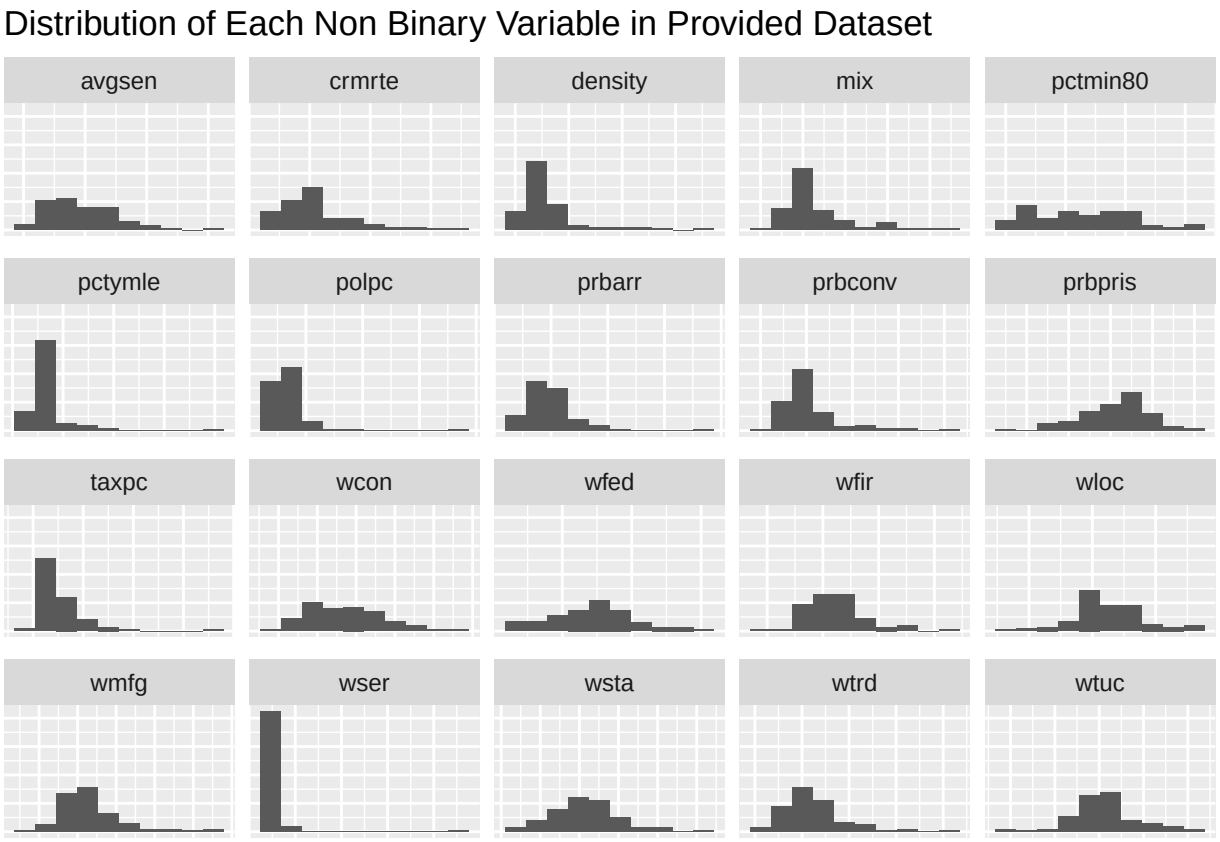
Distribution of Dataset

```
# Histograms for EDA
# Creating dataframe for continuous variables
C_cont_vars_hist = select(C, avgsen, crmrte, density, mix, pctmin80, pctymle,
polpc, prbarr, prbconv, prbpris, taxpc, wcon, wfed, wfir, wloc, wmfg, wser, wsta,
wtrd, wtuc)

# Gathering for aggregated ggplots
invisible(C_cont_vars_hist %>% gather() %>% head())

# Plot histograms for all variables in C_cont_vars_hist
ggplot(gather(C_cont_vars_hist), aes(value)) +
  geom_histogram(bins = 10) +
```

```
ggtitle("Distribution of Each Non Binary Variable in Provided Dataset") +
theme(axis.text.x=element_blank()) +
theme(axis.text.y=element_blank()) +
theme(axis.title.x=element_blank()) +
theme(axis.title.y=element_blank()) +
theme(axis.ticks.x=element_blank()) +
theme(axis.ticks.y=element_blank()) +
facet_wrap(~key, scales = 'free_x')
```



Upon inspection of the continuous variables in the dataset, it appears that the skewed variables in this dataset are: taxpc, pctymle, wser, pctmin80, polpc.

An examination of the distribution of the variables will help to evaluate which variables are key candidate for log transformation, a common variable transformation used in linear regression to obtain a more robust model.

It's also important to note that 'prbconv' variable and 'prbarr' variable had max values exceeding 1, despite being called 'probability'. Upon further inspection of the data origins, we will conclude that these variables are ratios, not probability. Therefore, we will keep them as they are.

Model 1

Overview

The goal of our initial model is to establish a baseline. From our review of Correlation Plot of Dataset, we have selected the following key variables:

- People per Sq. Mile (density) The density variable had one of the strongest correlations with the crmrte variable. Since there are many policy changes that could affect the density of an area, it makes sense for us to include density variable as an explanatory variable in our first model.
- Probability of Arrest (prbarr) Unlike the Density variable, the Probability of Arrest variable was negatively correlated with the crmrte variable but with less significant effect. We have decided to include the prbarr variable, since it does not blatantly violate collinearity with the density variable and provides interesting options for the model to evolve. From policy change perspective, there are many interventions to affect the probability of arrest proxied by the ratio of arrests to offenses from the FBI's Uniform Crime Report's single year data.

Furthermore, we have decided to log transform the crmrte variable and density variable for better fit of the model since both variables have no obvious maximum point and meaningful zero point. For crime rate, while we may be losing parsimony and intuitiveness for this model by logging crime rate, we discovered that log transforming this variable allowed us to meet the CLM assumption for zero conditional mean. For visualizations of the effects of log transformation, please refer to Appendix 2.

$$log(crmrte) = \beta_0 + \beta_1 log(density) + \beta_2(prbarr) + u$$

Let's go ahead and build our first model in R.

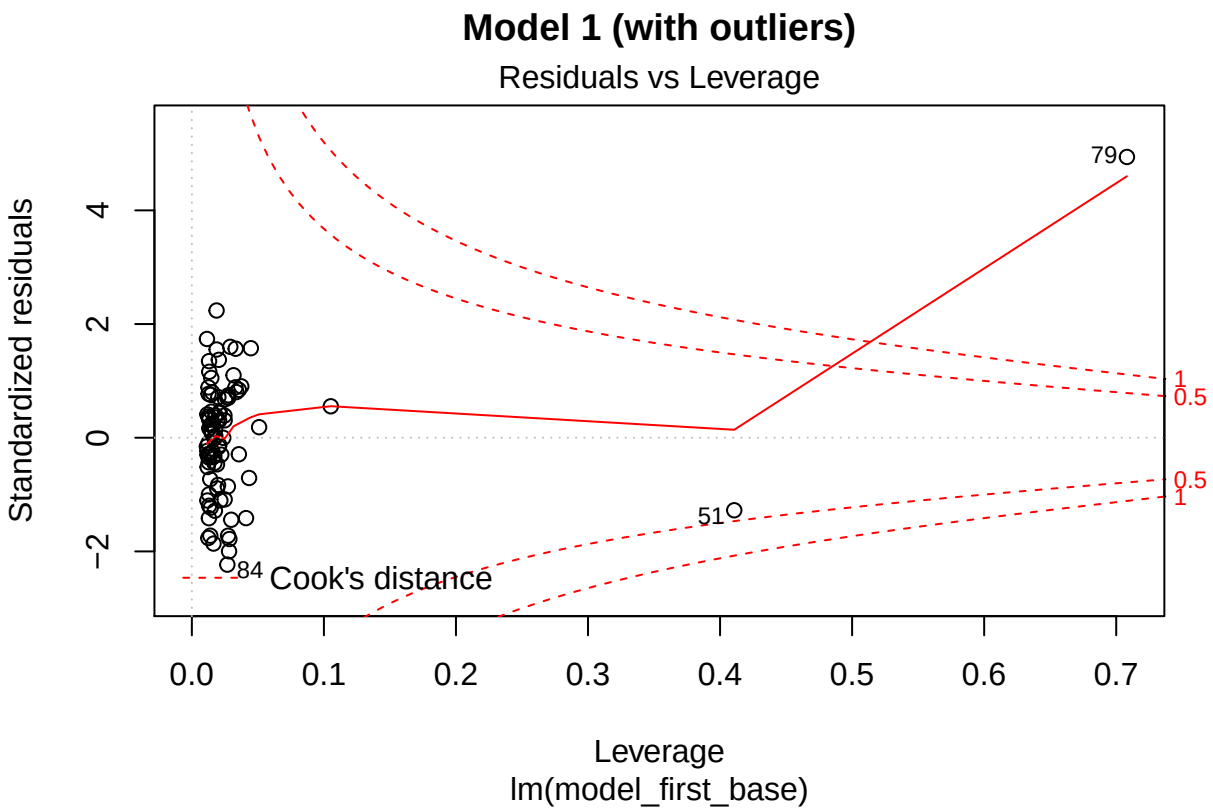
```
# Creating the first model
model_first_base = "log(crmrte)~log(density)+prbarr"

# Code to generate the first model
model_first_pre <- lm(model_first_base, data=C)
C$log_density <- log(C$density)
C$log_crmrte <- log(C$crmrte)
```

Outliers

Unfortunately, our first model had an outlier with significant leverage. The outlier (point 79) is evident in the Residuals vs Fitted chart below.

```
plot(model_first_pre, which=5, main = "Model 1 (with outliers)")
```

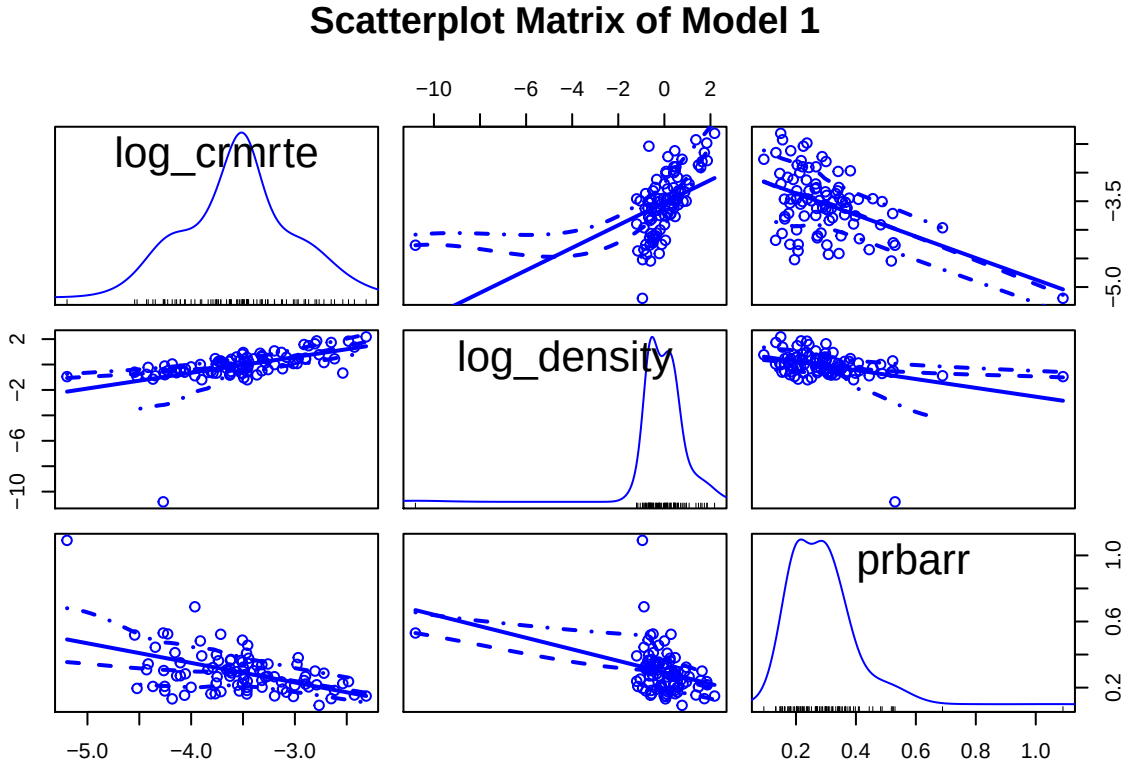


Since the outlier is beyond Cook's distance of 1, we will remove the outlier from the dataset. This will ensure that we satisfy the CLM assumption later on.

```
# Removing outlier points with high leverage/influence using cook's d
cooksd <- cooks.distance(model_first_pre)
influential <- as.numeric(names(cooksd)[(cooksd > 1)])
C_screen1 <- C[-influential, ]

# Rerunning model minus influential points
model_first <- lm(log(crmrte)~log(density)+prbarr, data = C_screen1)

# Scatterplot matrix to measure collinearity of our variables of interest
scatterplotMatrix(C[ , c("log_crmrte", "log_density","prbarr")],
                  main="Scatterplot Matrix of Model 1")
```

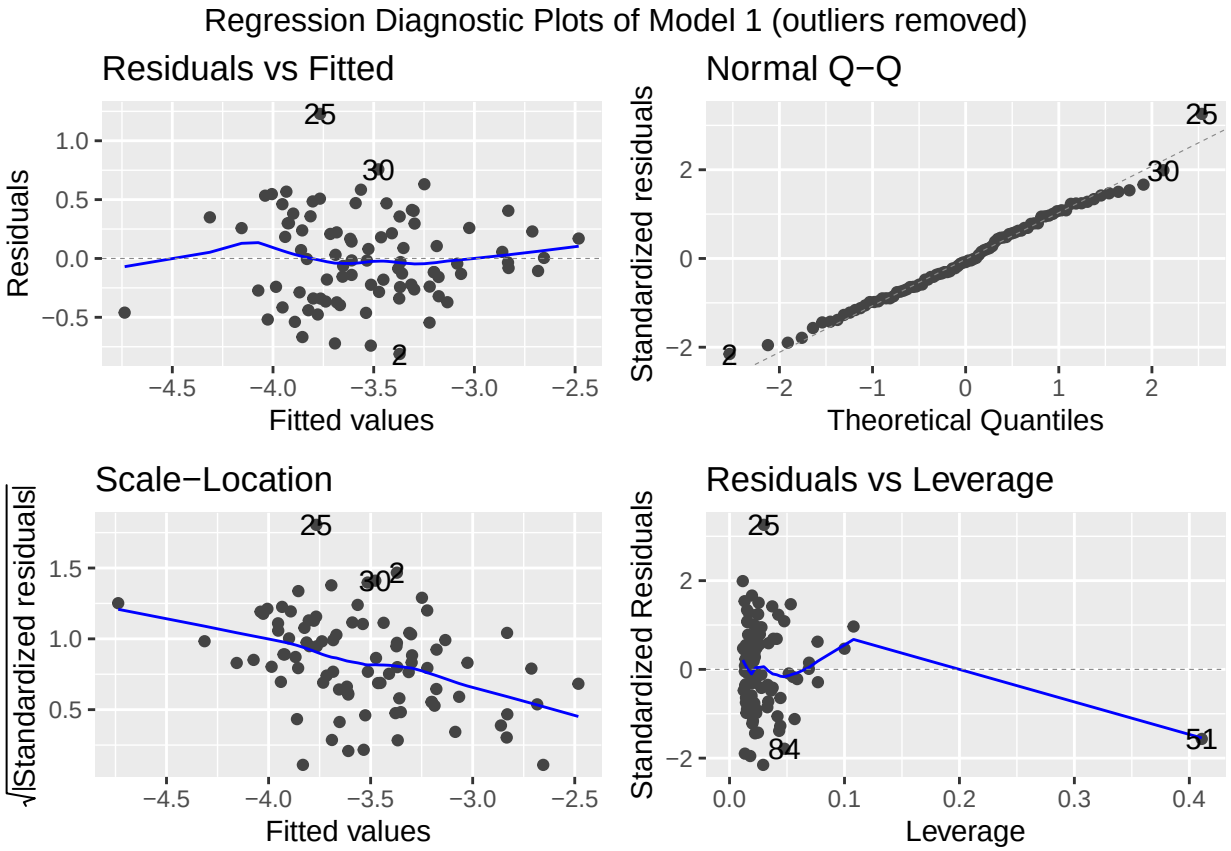


Based on the scatterplot matrix above, the log transformed variable of crime rate appears positively correlated with log transformed variable of density. However, we must note the negative correlation with prbarr. We will examine this further in the following section about CLM assumptions for this model.

Below are the various plots of our first model to assess CLM Assumptions.

```
p<-autoplot(model_first, top="Regression Diagnostic Plots of Model 1 (outliers removed)")

gridExtra::grid.arrange(grobs = p@plots,
                        top="Regression Diagnostic Plots of Model 1 (outliers removed)")
```



CLM Assumptions

CLM Assumption 1: The regression model is linear.

The regression model stated above is linear in the coefficients and the error term.

CLM Assumption 2: The data is from a random sample which is independently and identically distributed.

Without prior knowledge of how the data was generated and sourced, and considering that the data provided had errors and was modified from the original Cornwell and Trumball study, we had trouble satisfying this assumption at face value.

For the purposes of our OLS model, we will assume that the data provided by the campaign is independently and identically distributed. However, it is possible that this may not be the case. For instance, it is plausible that a county may under-report or over-report crime statistics in order to appear safer compared to other areas of North Carolina.

Additionally, we will assume that it is a random sample. A more extensive study on the determinants of crime, using individual level data rather than county aggregated statistics may yield different results.

We have considered subsampling our data to generate a random sample - however, given the relatively small amount of data (>100 observations), we did not feel comfortable doing so. For the purposes of our model, we will assume that this is true, however, it is possible that it may be false.

CLM Assumption 3: The initial model does not have multicollinearity

```
stargazer(vif(model_first), header=F, title = "VIF for Model 2")
```

Table 3: VIF for Model 2

log(density)	prbarr
1.149	1.149

Our Variance Inflation Factor test results were less than 4 for both key variables in the first model. We do not need to worry about large standard errors. We will assume that multicollinearity is not an issue for this model.

CLM Assumption 4: The initial model satisfies zero conditional mean of errors and exogeneity

According to the Residuals vs Fitted chart above, the mean line (highlighted in blue) approximately follows the slope of 0 throughout the curve, considering the fact that there is a slight bend in our dataset. Since there are no obvious outliers present in the chart, we can assume that the model meets the assumption of zero conditional mean.

CLM Assumption 5: Errors in the model are heteroscedastic

The residuals-fitted chart indicates a relatively even spread of data points in the middle. There are no obvious spreads available in the chart. To further confirm, let's take a look at it from Breusch-Pagan test's results.

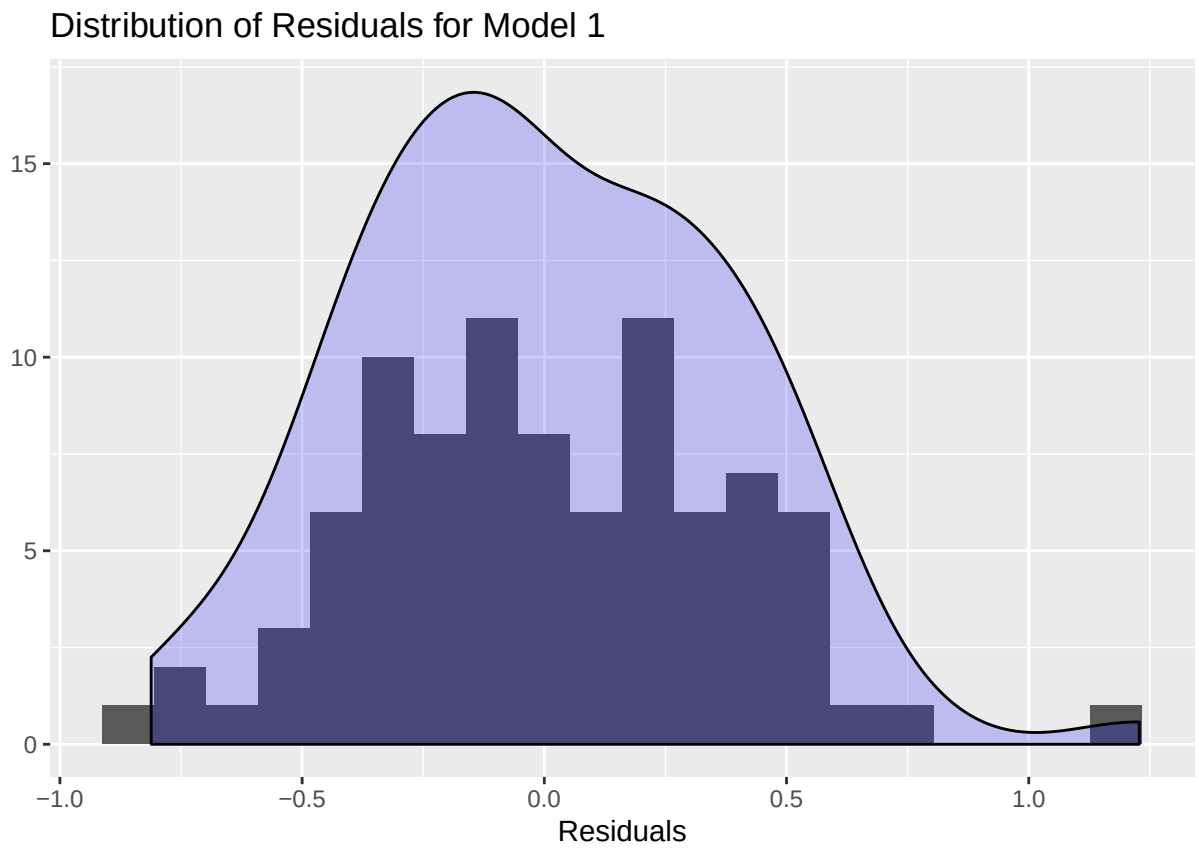
```
# H0: homoscedasticity
bptest(model_first)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_first
## BP = 9.1269, df = 2, p-value = 0.01043
```

The Breusch-Pagan test yielded a significant p value of 0.0104. Therefore, we reject the null hypothesis and can assume that there is heteroscedasticity. In order to mitigate this, we will need to use heteroscedastic robust errors.

CLM Assumption 6: Errors in the model have normal distribution

```
# Histogram of the residuals
ggplot(model_first, aes(model_first$residuals)) + geom_histogram(bins=20) +xlab("Residuals") +
ylab("")+ geom_density(alpha=.2, fill="blue",aes(y=.2 * ..count..))+
  ggtitle("Distribution of Residuals for Model 1")
```



According to the histogram, the residuals of our initial model appears to be normally distributed. To be sure, let's take a closer look by applying the Shapiro-Wilk test

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

```
##
##  Shapiro-Wilk normality test
##
## data:  model_first$residuals
## W = 0.98696, p-value = 0.5202
```

The null hypothesis is that these residuals are drawn from a population with a normal distribution. Since the p-value of Shapiro-Wilk test is greater than 0.05, we fail to reject the null hypothesis. The SW test allows us to assume normal distribution of errors.

Responding to violated assumptions.

Our review of the 6 CLM assumptions have revealed heteroscedasticity. To be safe, we will use robust standard errors as a form of mitigation.

Model interpretation

```
# Summary table of the first model
stargazer(model_first, header = F, type = "latex", omit.table.layout= "n",
          title="Regression Results for Model 1")
```

With the introduction of log transformation to crmrte variable and prbarr variable, our model results and coefficients changed. Our log transformed model resulted in a high F-statistic (46.662), which indicated

Table 4: Regression Results for Model 1

	Dependent variable:
	log(crmrte)
log(density)	0.425^{***}
	(0.057)
prbarr	-0.982^{***}
	(0.321)
Constant	-3.261^{***}
	(0.103)
Observations	89
R^2	0.520
Adjusted R^2	0.509
Residual Std. Error	0.383 (df = 86)
F Statistic	46.662^{***} (df = 2; 86)

high statistical significance. However, we need to be aware of the adjusted R-squared value of 0.510, which indicates medium practical significance. Below is the complete picture of the model with coefficients:

$$log(crmrte) = -3.2610 + 0.4248(log(density)) - 0.9819(prbarr) + u$$

We can confirm that both prbarr variable and log transformed density variable are good predictors for our initial model since they had low p-values with significant coefficients.

Positively Correlated Factors

If the density (number of people per square mile) goes up by 1 percent, we expect the crime rate to go up by approximately 0.425 percent.

Negatively Correlated Factors

If the probability of arrest in a county change by 1 percent, we expect the crime rate to drop by .982 percent.

With the baseline understanding that density and the probability of arrest in a county has a strong positive relationship with crime rate, let's continue to explore further by adding more predictors in our 2nd model.

Model 2

Overview

Our second model best balances parsimony and accuracy. We want to pick out the variables that we believe to be the best estimations on crime. Our first model provides a fairly good, broad assessment on some basic covariates that have a relationship to the crime rate on a county level. We want to build out a more robust version of the first model which has more explanatory power, but maintains a good degree of parsimony. The areas we want to cover are as follows:

How densely populated a county is

From our first model, we understand that the denser a particular area is, the higher the crime rate. Similar to the first model, we have decided to log transform this variable since it has no obvious maximum point and meaningful zero point.

How county addresses crime

We postulate that how tough an individual county is on crime will likely have an impact on the crime rate. How tough a county is on crime can be dictated by how aggressively they try to address crime via arrests (prbarr), or on a prosecutorial level via convictions (prbconv). While intuitively, these factors would seem to be correlated, an examination of the correlation plot and scatterplot matrix appears to indicate that these variables are not correlated with each other. As such, we will include both of these variables in our model.

Demographics

Furthermore, we will investigate the percentage of minorities in a county in the year 1980. In investigating the variables for this model, we found that the log transformed version of this variable helped to improve the explainability of model. It may be possible that counties with high levels of crime may have a high concentration of minorities. We decided to log transform this variable as well since we want to help secure normality and homoscedasticity for our model.

Dependent Variable

Lastly, similar to the first model, we will want to log transform our dependent variable, crime rate (crmrte). While we may be losing some parsimony in doing so, log transforming our dependent variable will help us fulfill the CLM assumptions (specifically, zero conditional mean). For a more detailed exploration on this topic, please refer to Appendix 2.

Our model will be as follows:

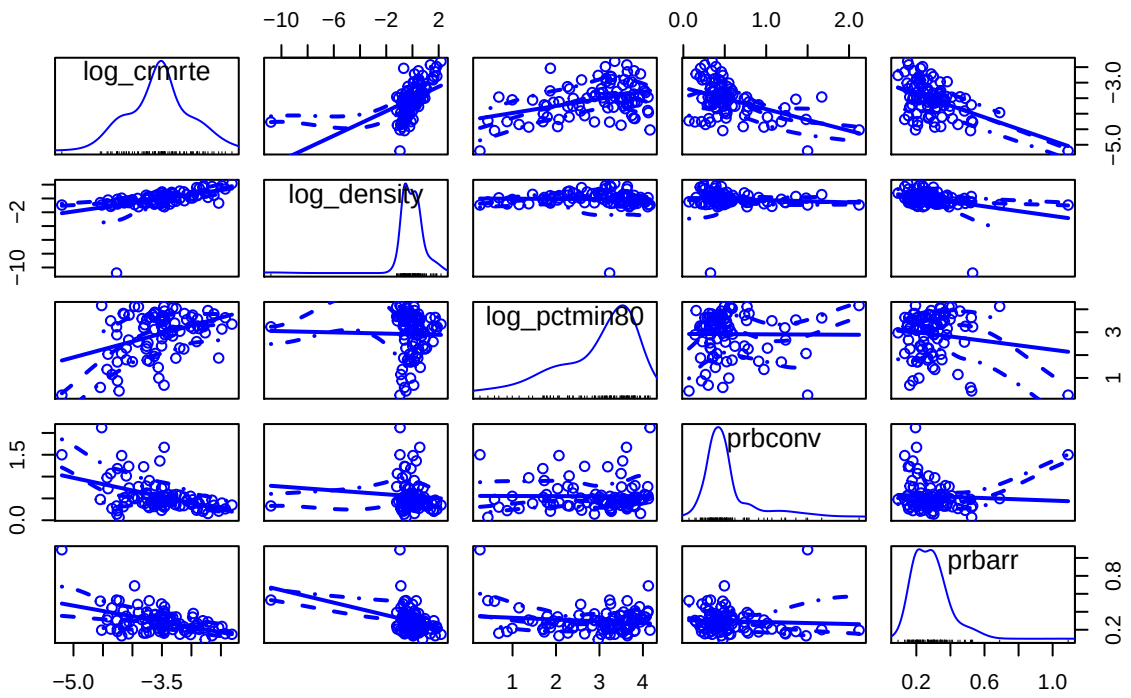
$$log(crmrte) = \beta_0 + \beta_1(prbarr) + \beta_2(prbconv) + \beta_3 log(density) + \beta_4 log(pctmin80) + u$$

```
# Creating the second model
model_second_base = "log(crmrte)~prbarr+prbconv+log(density)+log(pctmin80)"

# Code to generate second model
model_second_pre <- lm(model_second_base, data = C)

C$log_crmrte <- log(C$crmrte)
C$log_density <- log(C$density)
C$log_pctmin80 <- log(C$pctmin80)

# Scatterplot matrix to measure collinearity of our variables of interest
scatterplotMatrix(C[ , c("log_crmrte", "log_density","log_pctmin80", "prbconv", "prbarr")],
                  main="Scatterplot Matrix of Variables in Model 2 (no outliers removed)")
```

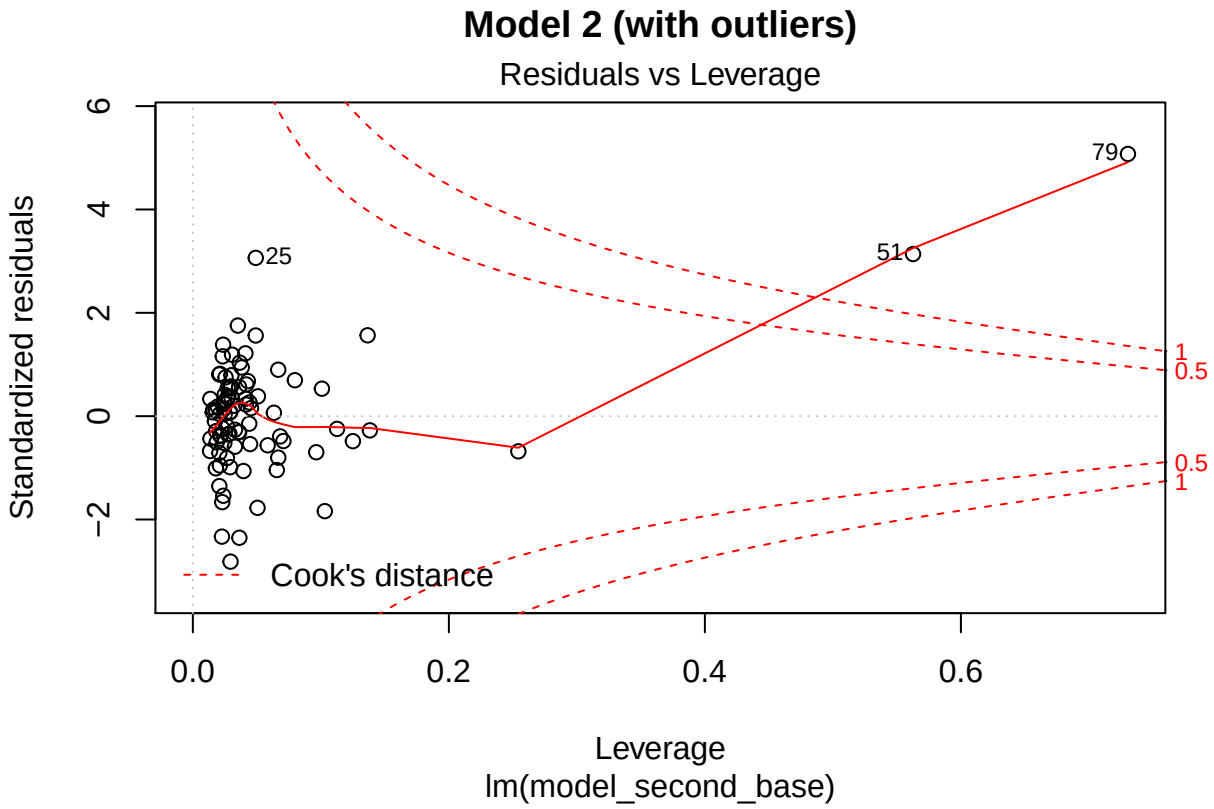


Based on the scatterplot matrix above, the logged variable of crime rate appears positively correlated with density and pctmin80, while being negatively correlated with prbconv and prbarr. For our remaining predictors, there does not appear to be an overwhelming positive or negative correlation, which is a good sign for multicollinearity. However, we will examine this assumption further in the CLM assumptions for this model.

Outliers

Before proceeding further, we want to address any outliers that may be influencing our model line. We will use a Cook's distance of 1 in order to properly remove lines that are highly influential and have high leverage. We will examine the model first by looking at the residuals vs leverage plot.

```
plot(model_second_pre, which=5, main="Model 2 (with outliers)")
```



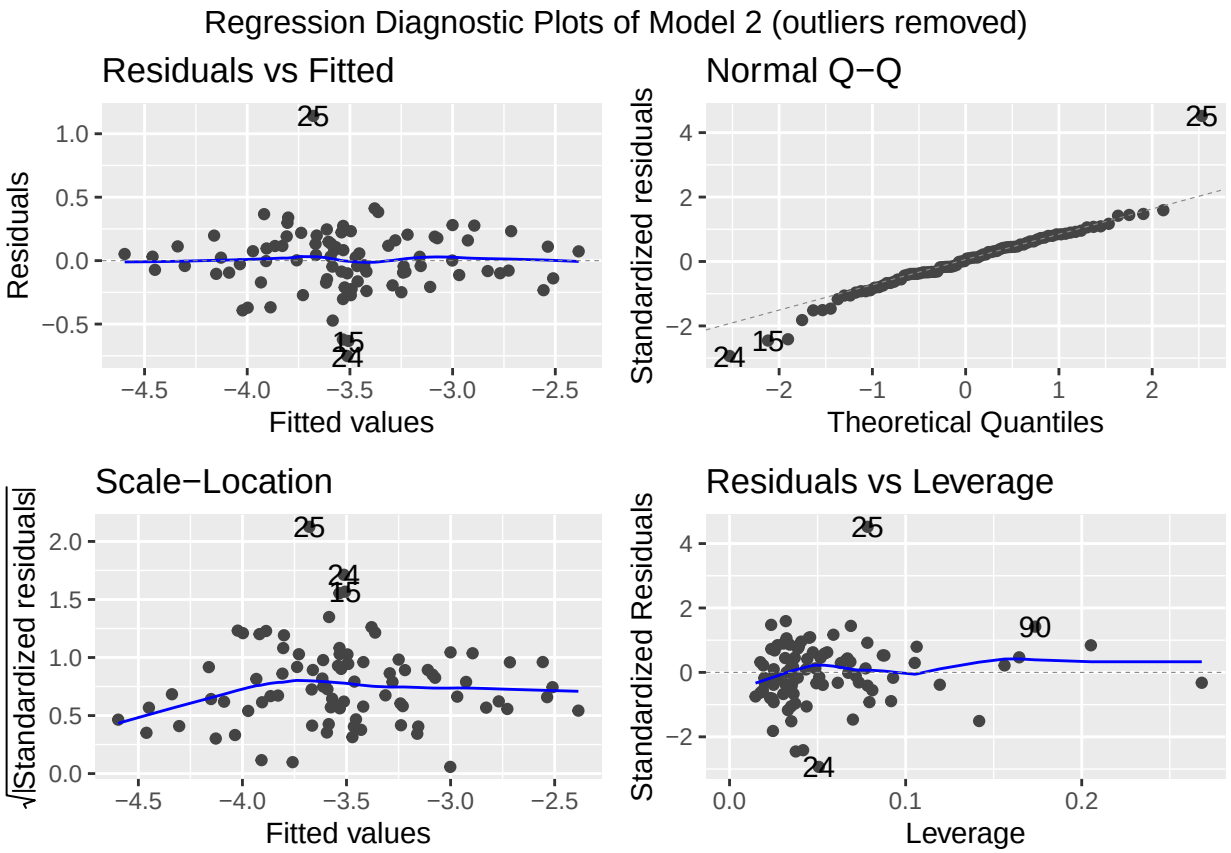
Point 51 & 79 appears to have high leverage and high influence. We use a Cook's distance of 1 to exclude any outliers from the dataset.

```
# Removing outlier points with high leverage/influence using cook's d
cooksd <- cooks.distance(model_second_pre)
influential <- as.numeric(names(cooksd)[(cooksd > 1)])
C_screen2 <- C[-influential, ]

# Rerunning model minus influential points
model_second <- lm(model_second_base, data = C_screen2)

# Residual plots
p<-autoplot(model_second, top="Regression Diagnostic Plots of Model 2 (outliers removed)")

gridExtra::grid.arrange(grobs = p@plots,
                        top="Regression Diagnostic Plots of Model 2 (outliers removed)")
```



We can see that our residuals vs leverage plot appears much more even now that we have removed point 51 & 79 from our dataset. Now that we have addressed the outliers for this model, let us examine the CLM assumptions for this model.

CLM Assumptions

CLM Assumption 1: The second model is linear.

The linear regression model stated above does not violate the linearity.

CLM Assumption 2: The data is a random sample of data which is independently and identically distributed.

Without prior knowledge of how the data was generated and sourced, and considering that the data provided had errors and was modified from the original Cornwell and Trumball study, we had trouble satisfying this assumption at face value.

For the purposes of our OLS model, we will assume that the data provided by the campaign is independently and identically distributed. However, it is possible that this may not be the case. For instance, it is plausible that a county may under-report or over-report crime statistics in order to appear safer compared to other areas of North Carolina.

Additionally, we will assume that it is a random sample. A more extensive study on the determinants of crime, using individual level data rather than county aggregated statistics may yield different results.

We have considered subsampling our data to generate a random sample - however, given the relatively small amount of data (>100 observations), we did not feel comfortable doing so. For the purposes of our model, we will assume that this is true, however, it is possible that it may be false.

CLM Assumption 3: The second model does not have multicollinearity.

```
stargazer(vif(model_second), header=F, title = "VIF for Model 2")
```

Table 5: VIF for Model 2

prbarr	prbconv	log(density)	log(pctmin80)
1.399	1.297	1.334	1.016

Our Variance Inflation Factor test results were less than 4 for both key variables in the second model. We do not need to worry about large standard errors. We will assume that multicollinearity is not an issue for this model.

CLM Assumption 4: The second model satisfies the zero conditional mean of errors and exogeneity.

The plot of points on the Residuals vs Fitted plot appears to be an even band across the x axis. By looking at the superimposed line in this graph, we believe that this model satisfies the Zero Conditional Mean criteria for the Classical Linear Model. While we debated logging our dependent variable, we discovered that logging crime rate ended up yielding a better model fit, and better satisfaction of the Zero Conditional Mean criteria.

CLM Assumption 5: Errors in the model are heteroscedastic.

We will use a Breush Pagan test to examine if the errors in the model are heteroscedastic or homoscedastic.

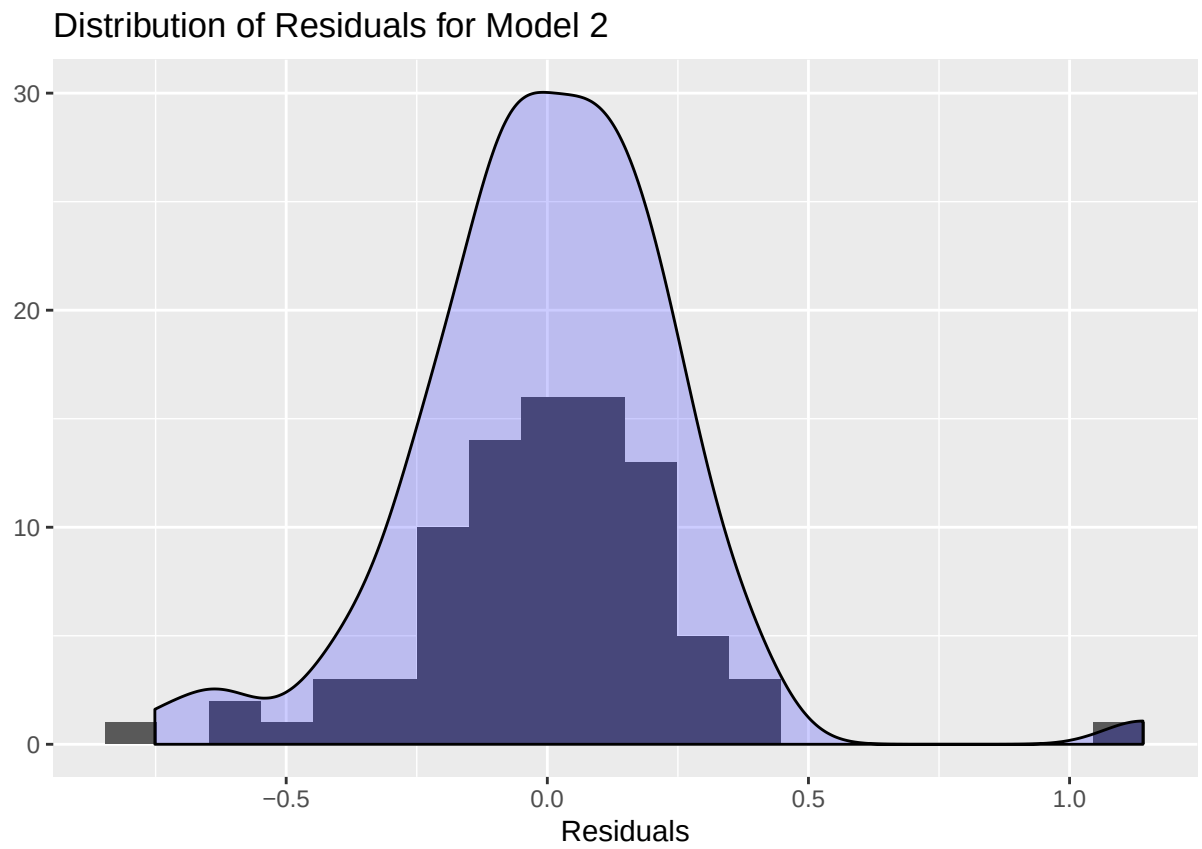
```
bptest(model_second)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_second
## BP = 9.6876, df = 4, p-value = 0.04603
```

The p value is less than 0.05, which means we can reject the null hypothesis that there is homoscedasticity. The line is not perfectly flat for this model's scale-location plot, which indicates some level of heteroscedasticity, albeit fairly small. To err on the side of caution, we will use heteroscedastic robust errors as a form of mitigation.

CLM Assumption 6: Errors in the model are normally distributed.

```
ggplot(model_second, aes(model_second$residuals)) + geom_histogram(bins=20) +xlab("Residuals") +
ylab("")+ geom_density(alpha=.2, fill="blue",aes(y=.2 * ..count..))+
  ggtitle("Distribution of Residuals for Model 2")
```



According to the histogram, the residuals of our initial model appears to be normally distributed. To be sure, let's take a closer look by applying the Shapiro-Wilk test

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

```
##
##  Shapiro-Wilk normality test
##
## data:  model_second$residuals
## W = 0.93605, p-value = 0.0003002
```

The null hypothesis is that these residuals are drawn from a population with a normal distribution. Since the p-value of Shapiro-Wilk test is less than 0.05, we reject the null hypothesis. The SW test does not allow us to assume normal distribution of errors. However, we realize that we're working with a dataset that has a sample size greater than 30. We will rely on OLS asymptotic properties.

Model interpretation

```
stargazer(model_second, header = F, type = "latex", omit.table.layout= "n",
          title="Regression Results for Model 2")
```

We believe that second model (model_second in our code) is our strongest model that best fits the needs of the political campaign and has the strongest explanatory power when it comes to crime rate. Despite only having 4 factors prbconv, prbarr, log(pctmin80), and log(density), we have a relatively high adjusted R2 value of .744 and a relatively high F statistic (64.138).

The p-values on all of our selected explanatory variables are all extremely low, which is good, since this tests for whether or not our coefficient is equal to zero (no effect). Since all of our predictors have low p-values,

Table 6: Regression Results for Model 2

	Dependent variable:
	log(crmrte)
prbarr	-1.597^{***}
	(0.313)
prbconv	-0.656^{***}
	(0.094)
log(density)	0.324^{***}
	(0.043)
log(pctmin80)	0.229^{***}
	(0.031)
Constant	-3.395^{***}
	(0.145)
Observations	88
R^2	0.756
Adjusted R^2	0.744
Residual Std. Error	0.263 (df = 83)
F Statistic	64.138^{***} (df = 4; 83)

they are meaningful additions to our model. Our coefficients appear to be practically significant as well. Below is the complete picture of the model with coefficients:

$$log(crmrte) = -3.395 - 1.597(prbarr) - 0.656(prbconv) + 0.324(log(density)) + 0.229(log(pctmin80)) + u$$

Positively Correlated Factors

If the density (number of people per square mile) goes up by 1 percent, we expect the crime rate to go up by approximately .324 percent. This appears to indicate that higher density tend to be associated with higher rates of crime.

If the percentage of minorities goes up by 1 percent relative to the current percentage (e.g: 3 to 3.3 percent would be a 10 percent increase), we expect the crime rate to go up by approximately .229 percent.

Negatively Correlated Factors

If the probability of arrest in a county changes by 1 percent, we expect the crime rate to drop by 1.597 percent.

If the probability of conviction in a county changes by 1 percent, we expect the crime rate to drop by 0.656 percent.

Policy recommendation

Assuming that our client does not want to enact any policy to affect the density of cities, the population distribution of our counties, or the distribution of minorities in a county but still wants to enact policies that would have a downward effect on the crime rate, our suggestion would be to improve the behavior of police

and government officials (specifically, by increasing arrests and convictions), in order to positively impact the crime rate.

Model 3

Overview

We have already identified many covariates as part of Model 2. We will create a new model that contains all the covariates and see if additional covariates help improve the model at all. The key purpose of this model is to demonstrate the robustness of the results to model specification.

We will not consider the issues with multicollinearity in this model. We are also going to consider the covariates even if they are not statistically significant to see the impact of those decisions.

Our model will be as follows:

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

We can find the coefficients and standard errors for this model using the functions in R.

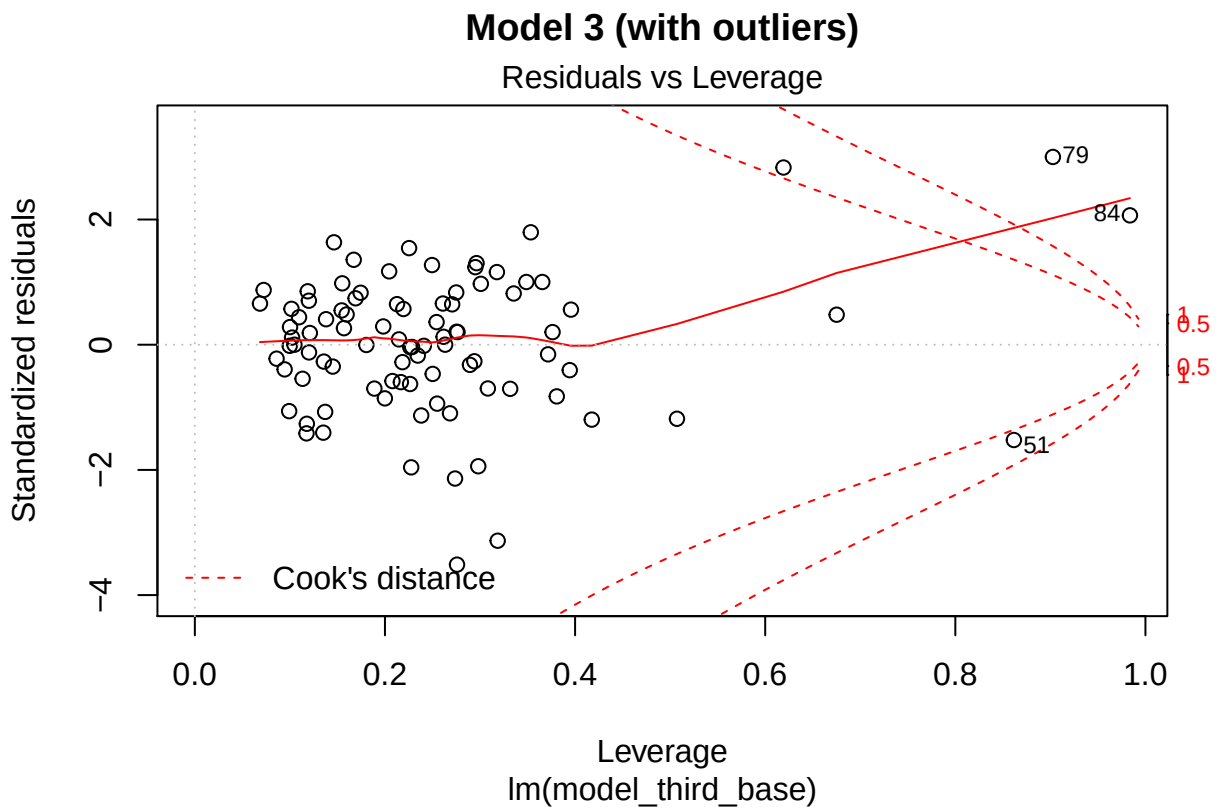
```
# Code to generate third model
model_third_base_pt1 = "log(crmrte) ~ prbarr + prbconv + log(density) + polpc"
model_third_base_pt2 = "+ prbpris + avgsen + taxpc + west + central"
model_third_base_pt3 = "+ urban + log(pctmin80) + wcon + wtuc + wtrd + "
model_third_base_pt4 = "wfir + wser + wmfg + wfed + wsta"
model_third_base_pt5 = "+ wloc + log(mix) + pctymle"

# Model using concatenated strings
model_third_base = paste(model_third_base_pt1, model_third_base_pt2, model_third_base_pt3,
model_third_base_pt4, model_third_base_pt5)
model_third_pre <- lm(model_third_base,data=C)
```

Outliers

We want to address any outliers that may be influencing our model line. We will use a Cook's distance of 1 in order to properly remove lines that are highly influential and have high leverage.

```
# Adding plot with outliers
plot(model_third_pre, which=5, main = "Model 3 (with outliers)")
```



Points 79 and 84 appear to have high leverage and high influence. We use a Cook's distance of 1 to exclude any outliers from the dataset.

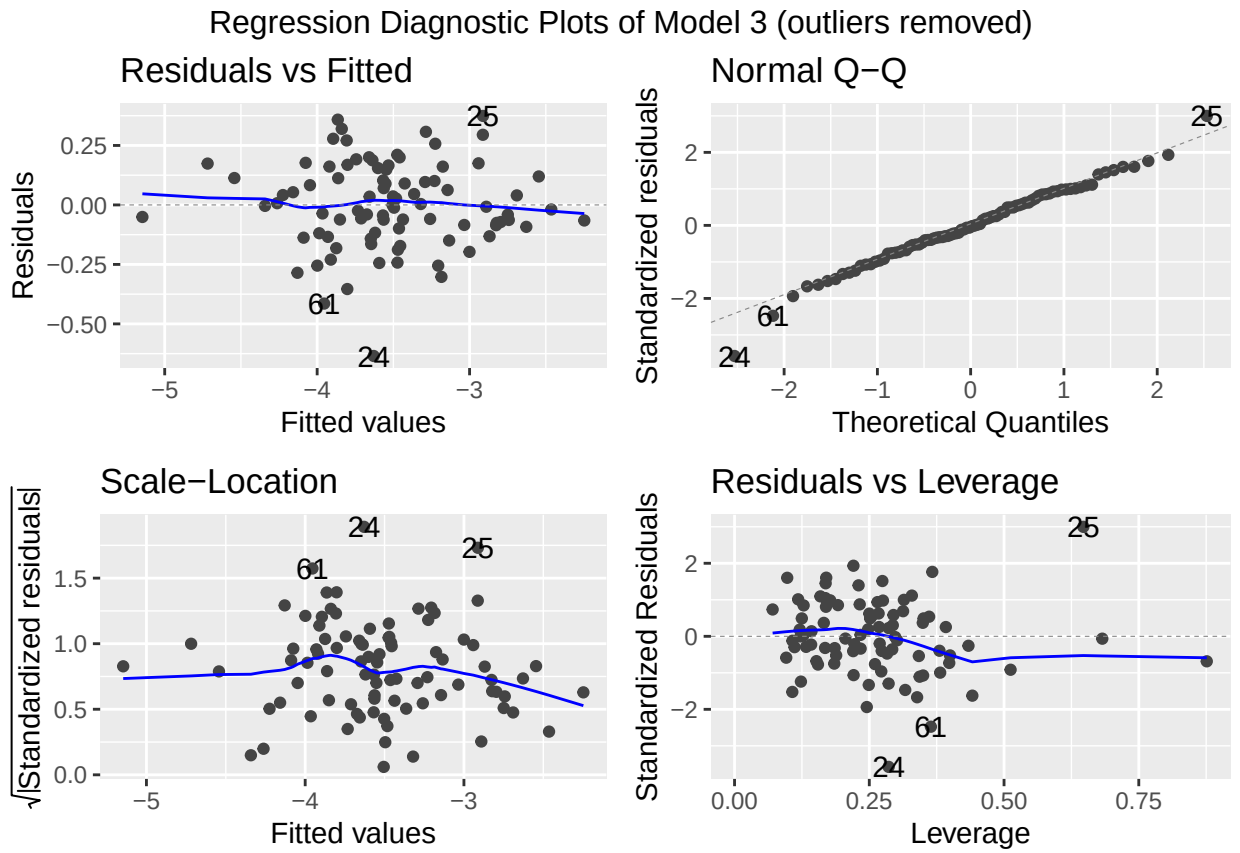
```
# Removing outlier points with high leverage/influence using cook's d
cooksd <- cooks.distance(model_third_pre)

# Finding the influential points
influential <- as.numeric(names(cooksd)[(cooksd > 1)])
C_screen3 <- C[-influential, ]

# Rerunning model minus influential points
model_third <- lm(model_third_base, data = C_screen3)

# Residual plots
p<-autoplot(model_third, top="Regression Diagnostic Plots of Model 3 (outliers removed)")

gridExtra::grid.arrange(grobs = p@plots,
                        top="Regression Diagnostic Plots of Model 3 (outliers removed)")
```



We can see that our residuals vs leverage plot appears much cleaner now that we have removed point 79 & 84 from our dataset.

CLM Assumptions

CLM Assumption 1: Linearity in parameters

The regression model is linear as described by the linear equation. Only coefficients are assumed to be linear.

CLM Assumption 2: The data is from a random sample which is independently and identically distributed

Without prior knowledge of how the data was generated and sourced, and considering that the data provided had errors and was modified from the original Cornwell and Trumball study, we had trouble satisfying this assumption at face value.

For the purposes of our OLS model, we will assume that the data provided by the campaign is independently and identically distributed. However, it is possible that this may not be the case. For instance, it is plausible that a county may under-report or over-report crime statistics in order to appear safer compared to other areas of North Carolina.

Additionally, we will assume that it is a random sample. A more extensive study on the determinants of crime, using individual level data rather than county aggregated statistics may yield different results.

We have considered subsampling our data to generate a random sample - however, given the relatively small amount of data (>100 observations), we did not feel comfortable doing so. For the purposes of our model, we will assume that this is true, however, it is possible that it may be false.

CLM Assumption 3. Model 3 has multicollinearity

VIF values is more than 4 for the density, west and pctmin80 variables and it indicates that there is collinearity and standard error might affect the ability to model the data. It is better to remove some collinear variables to avoid that issue.

```
# Checking the VIF to identify multicollinearity in the model
stargazer(vif(model_third), header=F, type="latex", flip=TRUE, title = "VIF for Model 3")
```

Table 7: VIF for Model 3

prbarr	2.319
prbconv	2.113
log(density)	4.417
polpc	3.384
prbpris	1.323
avgsen	1.815
taxpc	2.088
west	4.485
central	2.110
urban	2.533
log(pctmin80)	4.038
wcon	2.252
wtuc	1.657
wtrd	3.100
wfir	2.850
wser	2.633
wmfg	1.964
wfed	3.382
wsta	1.794
wloc	2.535
log(mix)	2.004
pctymle	1.583

CLM Assumption 4: Model 3 satisfies zero conditional mean of errors and exogeneity

We can call a model as **unbiased** if it satisfies the consistency assumptions(First 3 assumptions) and 4th assumption of CLM.

Zero conditional mean for errors can be identified by looking at the blue line in the Residuals vs Fitted chart above for this model. Since it is approximately flat, we can assume that it satisfies the zero conditional mean assumption. This is a stronger assumption and so we don't need to test for exogeneity

CLM Assumption 5: Errors in the model are heteroscedastic

We can call a model as **BLUE** if it is unbiased and it is homoscedastic. Based on the Residuals vs Fitted chart above, Model 3 doesn't look homoscedastic.

We can check if the model has homoscedasticity with a Breusch-Pagan test.

```
# Running the Breusch-Pagen test to identify homoscedasticity
bptest(model_third)
```

```
##
##  studentized Breusch-Pagan test
##
```

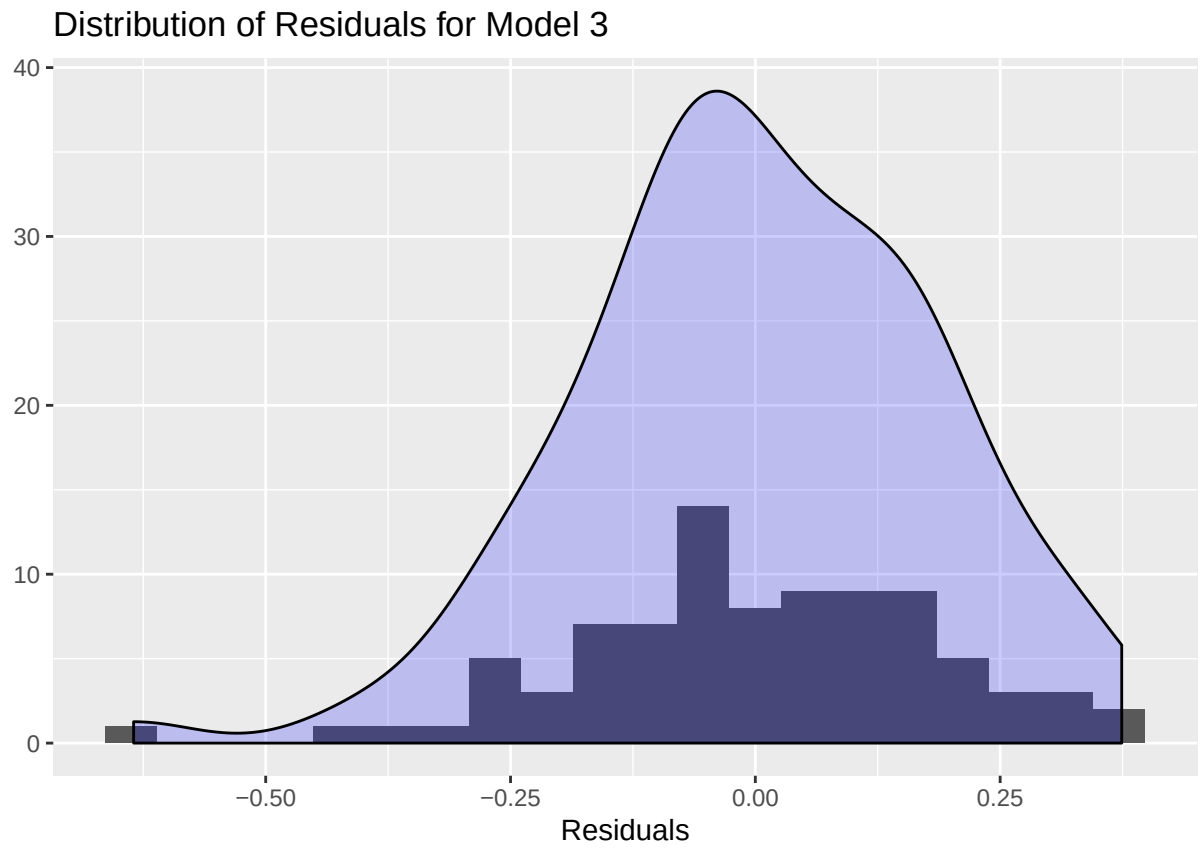
```
## data:  model_third
## BP = 35.583, df = 22, p-value = 0.03364
```

The null hypothesis for this test is that the errors are homoscedastic. Since p-value=0.03364(<0.05) and it is less than 0.05, we reject the null hypothesis. Therefore, the errors are heteroscedastic.

CLM Assumption 6: Errors in the model are normally distributed

A view of the histogram below shows that the errors look almost normally distributed.

```
# Plot to show normality
ggplot(model_third, aes(model_third$residuals)) + geom_histogram(bins=20) +xlab("Residuals") +
ylab("")+ geom_density(alpha=.2, fill="blue",aes(y=.2 * ..count..)) +
  ggtitle("Distribution of Residuals for Model 3")
```



We can also consider the formal Shapiro-Wilk test of normality.

```
# Running Shapiro-Wilk test to test normality of residuals
shapiro.test(model_third$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_third$residuals
## W = 0.98333, p-value = 0.3191
```

The null hypothesis for the Shaprio-Wilk test is that the errors are normally distributed. Since the p-value=0.3191 (> 0.05), we fail to reject the null hypothesis and conclude that the errors are normally distributed.

Responding to CLM Assumptions Violations

We seem to have a violation of homoscedasticity and we also have multicollinearity. Therefore, we conclude that exclusion of these variables from model 2 is justified.

We can use robust standard errors to address heteroscedasticity.

Model Interpretation

This model includes all the covariates from the dataset. We are adding all these variates in spite of seeing multi-collinearity between them. This model also includes covariates that have a very high p-value and are not statistically significant.

```
# Interpreting the coefficients through stargazer
stargazer( model_third, type = "latex",
           title = "Regression Output for Model 3",
           keep.stat = c("adj.rsq", "n", "f", "ser"),
           star.cutoffs = c(0.05, 0.01, 0.001),
           header = F, single.row = TRUE)
```

Table 8: Regression Output for Model 3

	Dependent variable:
	log(crmrte)
prbarr	-1.620^{***} (0.251)
prbconv	-0.584^{***} (0.104)
log(density)	0.309^{***} (0.062)
polpc	182.443^{***} (41.901)
prbpris	-0.235 (0.338)
avgsen	-0.016 (0.011)
taxpc	0.005^{*} (0.002)
west	0.044 (0.111)
central	-0.146^{*} (0.067)
urban	-0.128 (0.124)
log(pctmin80)	0.229^{***} (0.047)
wcon	0.001 (0.001)
wtuc	0.0003 (0.0004)
wtrd	0.001 (0.001)
wfir	-0.001 (0.001)
wser	-0.002^{**} (0.001)
wmfg	0.00000 (0.0004)
wfed	0.002^{*} (0.001)
wsta	-0.0003 (0.001)
wloc	0.0001 (0.001)
log(mix)	0.034 (0.060)
pctymle	2.124 (1.198)
Constant	-3.992^{***} (0.500)
Observations	88
Adjusted R^2	0.848
Residual Std. Error	0.210 (df = 65)
F Statistic	23.089^{***} (df = 22; 65)
Note:	*p<0.05; **p<0.01; ***p<0.001

This model seems to work as it produces a statistically significant F statistic and a very high adjusted R2 value of 0.848. It also shows the general direction of all the covariates and their impact in deciding the crime rate.

Below is the complete picture of the model with coefficients:

$$log(crmrte) = -3.992 - 1.62(prbarr) - 0.584(prbconv) + 0.309(log(density)) + 182.443(polpc) - 0.235(probpris)$$
$$-0.016(avgsen) + 0.005(taxpc) + 0.044(west) - 0.146(central) - 0.128(urban) + 0.229(log(pctmin80)) + 0.001(wcon)$$
$$+0.0003(wtuc) + 0.001(wtrd) - 0.001(wfir) - 0.002(wser) + 0.0000(wmfg) + 0.002(wfed) - 0.0003(wsta) + 0.0001(wloc)$$
$$+0.034(log(mix)) + 2.124(pctymle) + u$$

There are multiple coefficients in this model that are close to zero and have minimal impact on the crime rate irrespective of their direction. Additionally, there are multiple variables in this model that are not statistically significant. The direction and the impact of the variables are reduced or increased due to the presence of multicollinearity in this model.

Positively correlated factors

Density(density), police per capita(polpc), percentage of minorities(pctmin80), offence mix(mix), western N.C(west) and percentage young male(pctymle) have a positive correlation to the crime rate. However, pctymle, mix and west variables are not statistically significant and do not need to be considered.

If the density (number of people per square mile) goes up by 1 percent, we expect the crime rate to go up by approximately .309 percent.

If the percentage of minorities goes up by 1 percent relative to the current percentage (e.g: 3 to 3.3 percent would be a 10 percent increase), we expect the crime rate to go up by approximately .229 percent.

If the police per capita increases by 1 unit, then the crime rate increases by 182.44 percent. We noticed that police per capita and crime rate appears to be positively correlated. This seems counter intuitive at first glance, since one would think that the crime rate would go down with more police. However, it may be the case that an increase in police per capita is a response to high levels of crime in a county.

Negatively correlated factors

Deterrent variables such as probability or arrest(prbarr), probability of conviction(prbconv), probability of prison sentence(prbpris),average sentence(avgsen) and other area variables like Urban and central have negative correlation to the crime rate. However, avgsen, urban and central variables are not statistically significant and need not be considered.

If the probability of arrest goes up by 1 percent, we can expect the crime rate to go down by 1.62 percent.

If the probability of prison sentence goes up by 1 percent, we can expect the crime rate to go down by .235 percent.

If the probability of conviction goes up by 1 percent, we can expect the crime rate to go down by .584 percent.

Factors with minimal impact

Wage variables like wfed, wser, wcon, wtuc, wtrd, wfir, wmfg, wsta,wloc and other variables like taxpc are not significant and may not have a high impact since they have coefficients closer to zero and any change in those variables don't affect the crime rate much. Furthermore, all these variables except wfed and wser are not statistically significant and need not be considered.

Even though wfed and wser are statistically significant, their impact is going to be minimal since their coefficients are very small.

Comparison of the models

We will be comparing the final versions of our 3 models below using the package stargazer. Note that we will be using to get robust standard errors to ensure that the standard errors are robust to heteroscedasticity.

```
# Getting the Robust standard error to pass to Stargazer for comparison for all 3 models
invisible((se.model_first = coeftest(model_first, vcov = vcovHC)[ , "Std. Error"]))
invisible((se.model_second = coeftest(model_second, vcov = vcovHC)[ , "Std. Error"]))
invisible((se.model_third = coeftest(model_third, vcov = vcovHC)[ , "Std. Error"]))

# We pass the standard errors into stargazer through the se argument.
stargazer( model_first, model_second, model_third, type = "latex",
           title = "Linear Models predicting crime rate",
           keep.stat = c("adj.rsq", "n", "f", "ser", "aic", "wald"),
           se = list(se.model_first, se.model_second, se.model_third),
           star.cutoffs = c(0.05, 0.01, 0.001),
           header = F,
           single.row = TRUE) # Omit more output related to errors
```

Table 9: Linear Models predicting crime rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
log(density)	0.425*** (0.055)	0.324*** (0.056)	0.309*** (0.067)
polpc			182.443** (68.122)
prbpris			−0.235 (0.430)
avgsen			−0.016 (0.014)
taxpc			0.005 (0.006)
west			0.044 (0.116)
central			−0.146* (0.071)
urban			−0.128 (0.164)
log(pctmin80)		0.229*** (0.030)	0.229*** (0.056)
wcon			0.001 (0.001)
wtuc			0.0003 (0.001)
wtrd			0.001 (0.002)
wfir			−0.001 (0.001)
wser			−0.002* (0.001)
wmfg			0.00000 (0.0005)
wfed			0.002* (0.001)
wsta			−0.0003 (0.001)
wloc			0.0001 (0.002)
log(mix)			0.034 (0.088)
pctymle			2.124* (0.998)
prbarr	−0.982* (0.492)	−1.597*** (0.369)	−1.620*** (0.309)
prbconv		−0.656*** (0.123)	−0.584*** (0.171)
Constant	−3.261*** (0.152)	−3.395*** (0.227)	−3.992*** (0.556)
Observations	89	88	88
Adjusted R^2	0.509	0.744	0.848
Residual Std. Error	0.383 (df = 86)	0.263 (df = 83)	0.210 (df = 65)
F Statistic	46.662*** (df = 2; 86)	64.138*** (df = 4; 83)	23.089*** (df = 22; 65)
Note:			*p<0.05; **p<0.01; ***p<0.001

Adjusted R^2

We utilize adjusted R-squared values for our model rather than the R-squared to evaluate how our regression models fit the data. This is because the R-squared statistic doesn't account for overfitting, which leads to inflated R-squared values that will not decrease. The Adjusted R-squared is more appropriate, since it penalizes for the number of additional terms in the model.

According to the stargazer table above, model 3 had the highest adjusted R-squared value (0.848), followed by model 2's adjusted R-squared value (0.744), and model 1's adjusted R-squared value (0.509). While it's natural for us to focus on the highest adjusted R-squared value, it's crucial to pay attention to the differences in adjusted R-squared among the models by number of variables added. To be specific, the increase of adjusted R-squared value from model 1 to model 2 was greater than the increase of adjusted R-squared value from model 2 to model 3, despite the fact that only 2 variables were added between model 1 and model compared to the 18 additional variables between model 2 and model 3.

F Statistic

The F statistic is a test where the test distribution has an F-distribution under the null hypothesis. It's used when comparing statistical models which have been fitted to a dataset, in order to identify the model which best fits the population from the data that it was sampled.

Among the 3 models, model had the highest F-statistic value with 64.138, followed by model 1's F-statistic value of 46.662, and model 3's F-statistic value of 23.089. We can conclude that model 2 best fits the data at hand.

AIC

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

The fact that our third model had the lowest AIC should not be too alarming, since our third model was force fitted by inclusion of all variables. From the review of AIC values among all 3 models, we can tell that model 2 improved significantly from model 1.

Along with adjusted R-squared values, F-Statistic values, and AIC values, it's important to pay close attention to the movement of coefficients. To be specific, the coefficients of the prbarr variable changes drastically from model 1 (-0.982) to model 2 (-1.597), while the change from model 2 (-1.597) to model 3 (-1.620) was less. Similar pattern exists for the density variable as the change from model 1 (0.425) to model 2 (0.324) is far greater than the change from model 2 (0.324) to model 3 (0.309). This observation tells us that our model 2 was close to best fit, since model 3 with all variables didn't necessarily improve the results.

Robust Standard Errors

We included heteroscedastic robust errors (also known as White Standard Errors) in order to account the fact that all 3 of our models showed evidence of heteroscedasticity. Heteroscedasticity-consistent standard errors allow us to fit our models that contain heteroscedastic residuals.

Conclusion

A comparison of all the 3 models is shown in the regression output above. Although all the 3 models discussed above attempt to model crime rate, we believe that the second model is the best model for this data as it has a nice balance between parsimony and accuracy. On one hand, model 1 is incomplete and misses some critical covariates. On the other hand, model 3 has many covariates that are unnecessary and may not have a huge impact on the outcome.

Given the data provided by the campaign, and from our analysis, we believe that from a policy level, improved training and directing the behavior of the police force, as well as improving the behavior of district attorneys, will likely have the best chance of improving crime rate. It is important to note that our models do not necessarily imply a causal effect - however, from a practical standpoint, it does stand to reason that some criminals may be deterred from committing a crime if a county is perceived to be "tough on crime".

Areas of improvement

Omitted Variables

We believe that there are several potential factors that may explain have an impact on the crime rate that were not included in the provided dataset. Given more time and funding, we would be interested in investigating areas such as:

- Education: Highly educated areas may be more correlated with lower rates of crime.
- Rate of gun ownership: Certain counties with high rates of crime may have high levels of gun ownership.
- Rates of alcohol & substance consumption: Areas with high levels of alcohol & substance abuse may be more correlated with high levels of crime.

These are omitted variables that we believe may have an impact on the crime rate in a county. We believe that attributes like these may also have a strong relationship with the rate of crime in a county, and would potentially be uncorrelated with our existing explanatory variables.

Research Design

Furthermore, the original study by Cornwell and Trumball used panel data to infer causal effects. In fact, the outdated panel data may not be relevant with the current crime rate since crime evolved with the introduction of technology and the internet. We also do not recommend using data from 1980 to prescribe policy for 2020.

Given additional time and resources, we would request for a more comprehensive dataset from the campaign to investigate this issue further. We believe that there are many potential factors that drive an individual to commit a crime and studying the phenomenon on an individual level may yield interesting insights as well.

Appendix 1: Exploratory Data Analysis

Data Preparation

The summary table of our dataset indicates the presence of NA values across all columns. Furthermore, it's important to pay special attention to 'prbconv' variable for the following reason. While other variables have 6 NA values, 'prbconv' variable has 5 null values and 1 tilde (special character). A simple na.omit() function will not work for our dataset.

```
# Simple tail() check on the 'prbconv' variable allowed us to identify the suspect events
knitr::kable(
  C[1:6,1:10 ], caption = 'Top 6 Rows of Our Dataset')
```

Table 10: Top 6 Rows of Our Dataset

crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central
0.0356036	0.298270	0.5275960	0.436170	6.71	0.0018279	2.4226327	30.99368	0	1
0.0152532	0.132029	1.4814800	0.450000	6.35	0.0007459	1.0463320	26.89208	0	1
0.0129603	0.444444	0.2678570	0.600000	6.76	0.0012343	0.4127660	34.81605	1	0
0.0267532	0.364760	0.5254240	0.435484	7.14	0.0015299	0.4915572	42.94759	0	1
0.0106232	0.518219	0.4765630	0.442623	8.22	0.0008602	0.5469484	28.05474	1	0
0.0146067	0.524664	0.0683761	0.500000	13.00	0.0028820	0.6113361	35.22974	1	0

```
knitr::kable(
  C[1:6,11:20 ], caption = 'Top 6 Rows of Our Dataset (continued)')
```

Table 11: Top 6 Rows of Our Dataset (continued)

urban	pctmin80	wcon	wtuc	wtrd	wfir	wser	wmfg	wfed	wsta
0	20.21870	281.4259	408.7245	221.2701	453.1722	274.1775	334.54	477.58	292.09
0	7.91632	255.1020	376.2542	196.0101	258.5650	192.3077	300.38	409.83	362.96
0	3.16053	226.9470	372.2084	229.3209	305.9441	209.6972	237.65	358.98	331.53
0	47.91610	375.2345	397.6901	191.1720	281.0651	256.7214	281.80	412.15	328.27
0	1.79619	292.3077	377.3126	206.8215	289.3125	215.1933	290.89	377.35	367.23
0	1.54070	250.4006	401.3378	187.8255	258.5650	237.1507	258.60	391.48	325.71

```
knitr::kable(
  C[1:6,21:23 ], caption = 'Top 6 Rows of Our Dataset (continued)')
```

Table 12: Top 6 Rows of Our Dataset (continued)

wloc	mix	pctymle
311.91	0.0801688	0.0778710
301.47	0.0302267	0.0826069
281.37	0.4651163	0.0721154
299.03	0.2736220	0.0735373
342.82	0.0600858	0.0706976
275.22	0.3195266	0.0989192

The entries above provide a sample look into what the data looks like for the report.

As a result of our evaluation of the dataset, we found several issues. We found empty rows in the provided dataset that interfered with how the variables were being read in R. We also believe the duplicate entry, for one county in the year 1987 is a data entry error. Since we do not want a single county to be double counted, we will remove this row and chalk it up to data entry error.

```
# Looking at the bottom 10 rows of our dataset
knitr::kable(
  A[92:97,1:10 ], caption = 'Last 6 Rows of Our Dataset')
```

Table 13: Last 6 Rows of Our Dataset

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc
92	NA	NA	NA	NA		NA	NA	NA	NA	NA
93	NA	NA	NA	NA		NA	NA	NA	NA	NA
94	NA	NA	NA	NA		NA	NA	NA	NA	NA
95	NA	NA	NA	NA		NA	NA	NA	NA	NA
96	NA	NA	NA	NA		NA	NA	NA	NA	NA
97	NA	NA	NA	NA	‘	NA	NA	NA	NA	NA

```
knitr::kable(
  A[88:89,1:10 ], caption = 'Duplicate Row Found in our EDA')
```

Table 14: Duplicate Row Found in our EDA

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc
88	193	87	0.0235277	0.266055	0.588859022	0.423423	5.86	0.0011789	0.8138298	28.51783
89	193	87	0.0235277	0.266055	0.588859022	0.423423	5.86	0.0011789	0.8138298	28.51783

```
# Creating new dataframe C with unique values only
C <- unique(B)
print("Number of samples in dataframe C:")
```

[1] "Number of samples in dataframe C:"

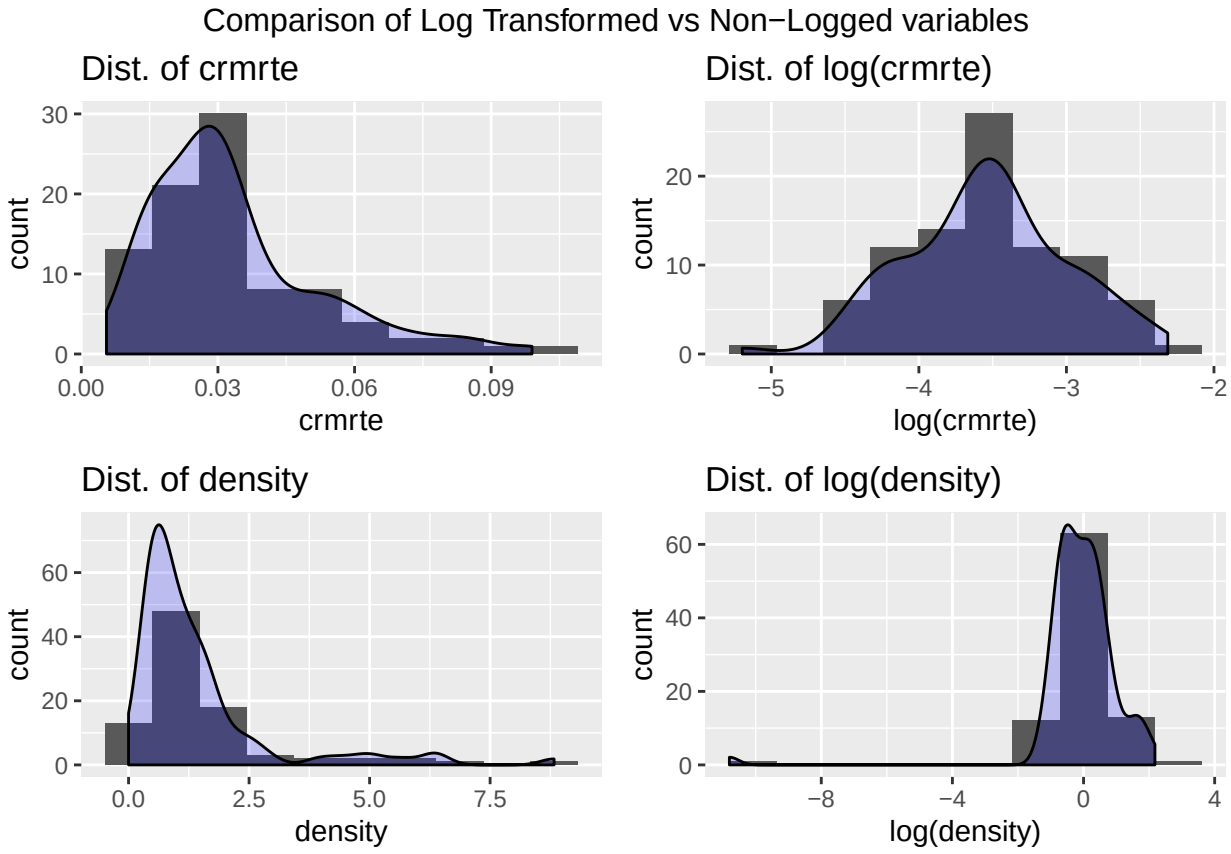
```
nrow(C)
```

[1] 90

Appendix 2: Log Transformation

Based on our examination of the Q-Q Plot, model_first did not have the best fit. We will go ahead and take a log of the variable for another look at the quality of our model.

```
# Comparison of Log Transformed vs Non-Logged variables
p1 <- ggplot(C, aes(C$crmrte))+geom_histogram(bins=10)+xlab("crmrte")+
  geom_density(alpha=.2, fill="blue") + ggtitle("Dist. of crmrte")
p2 <- ggplot(C, aes(log(C$crmrte)))+geom_histogram(bins=10)+xlab("log(crmrte)") +
  geom_density(alpha=.2, fill="blue",aes(y=.3 * ..count..))+ ggtitle("Dist. of log(crmrte)")
p3 <- ggplot(C, aes(C$density))+geom_histogram(bins=10)+xlab("density")+
  geom_density(alpha=.2, fill="blue",aes(y=1.3 * ..count..))+ ggtitle("Dist. of density")
p4 <- ggplot(C, aes(log(C$density)))+geom_histogram(bins=10)+xlab("log(density)")+
  geom_density(alpha=.2, fill="blue",aes(y=1.5 * ..count..))+ ggtitle("Dist. of log(density)")
grid.arrange(p1,p2,p3,p4, ncol=2, top="Comparison of Log Transformed vs Non-Logged variables")
```



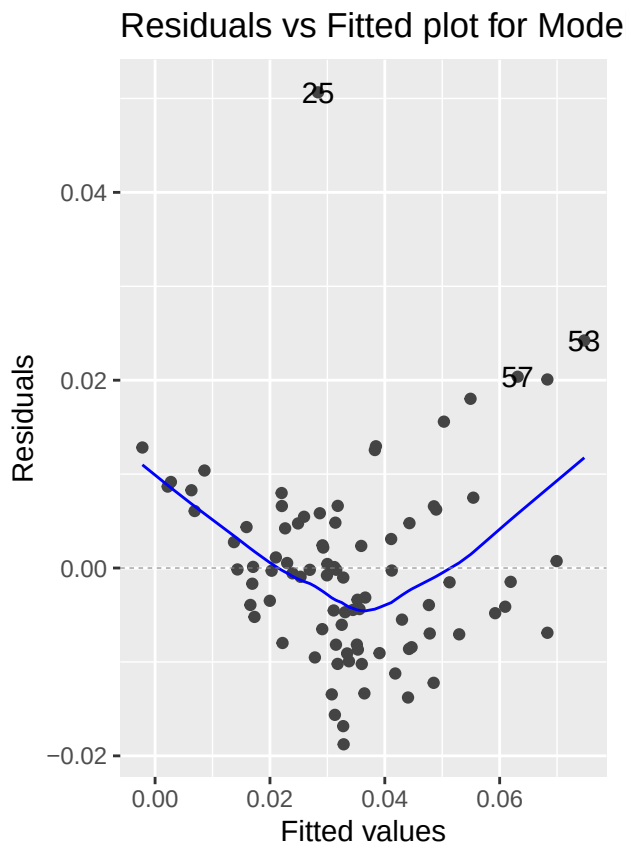
According to the chart matrix above, histograms of the log transformed variables are present on the right side and histograms of the regular variables are present on the left side. For crime rate, the log transformation shapes the distribution of the variable into a better normal shape. It's important to note that values of the crime rate will change after the log transformation, which is why we're seeing negative x values on the right side. For density, the log transformation brings the mean/median closer to 0. Applying log transformations will help with overall fit of the models. However, we must note that certain degree of parsimony will be lost in a trade off for better fitting models.

Zero Conditional Mean

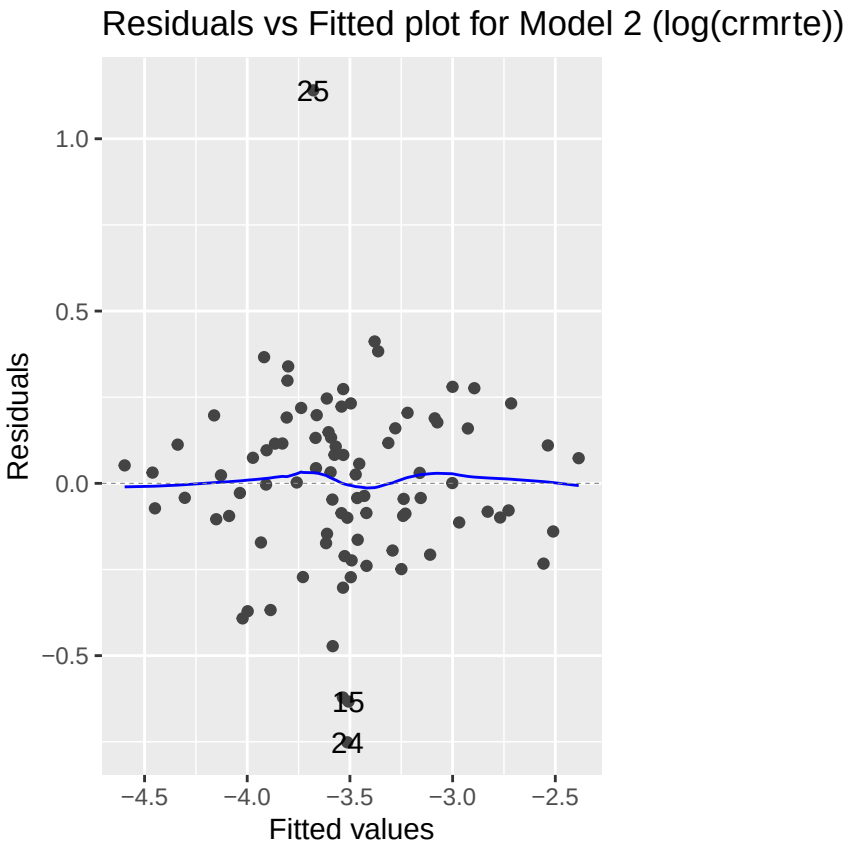
```
# First model with log() on crmrte
model_second_log <- lm(log(crmrte)~log(density)+prbarr+prbconv+log(pctmin80), data=C_screen2)

# First model with no log
model_second_nolog <- lm(crmrte~log(density)+prbarr+prbconv+log(pctmin80), data=C_screen2)

#residual plots for the model with log on crmrte
autoplot(model_second_nolog, which=1) +
  ggtitle("Residuals vs Fitted plot for Model 2 (crmrte)")
```




```
autoplot(model_second_log, which=1) +
  ggtitle("Residuals vs Fitted plot for Model 2 (log(crmrte))")
```



For Models 1-3, we also decided to log transform our dependent variable. For the purposes of discussion, we have shown a comparison of the Residuals vs Fitted plot for Model 2 - the key distinction being one has the dependent variable crmrte, and the other having the dependent variable log(crmrte). The independent variables remain the same. As you can see from the non log transformed version of Model 2, even with removing outliers from the dataset, we do not have homoscedasticity, and there is a clear U shape bend in the Residuals vs Fitted plot. We cannot assume Zero Conditional Mean for this model.

On the other hand, the Residuals vs Fitted plot for Model 2 is a straight line. While homoscedasticity would be a stretch for this model (due to the middle band of points), the justification for Zero Conditional Mean is much stronger. A similar logic applied to models 1 and 3. While we may be losing some parsimony with a log transformation of our dependent variable, we are able to satisfy one of the key assumptions for OLS Regressions - Zero Conditional Mean.