W203 Lab 3 - Reducing Crime

Shane Andrade, Suhas Gupta, Vidhu Nath

December 11, 2018

Introduction

We are tasked with assisting a candidate's political campaign by using regression modeling and statistical inference to understand the factors affecting crimes rate in a county, that can help in formulating an effective political policy. The main dependent variable modeled in this paper is **crime rate** which is a part of the data set containing crime statistics from 1980s on a selection of North Carolina counties. Linear regression modeling using OLS method was used to examine and model crime rate and hypothesis testing was conducted to understand the statistical significance of model coefficients. This is followed by a discussion of practical significance of the coefficients on the given data. The objective of this analysis is to propose a linear model that can explain the dependence of crimes committed per capita in a county on various socio-economic factors and help the campaign with formulation of political policy.

Note: Throughout the report there are multiple instances where the figures and tables do not immediately follow the text that cross references them or the code chunk that generated them. This is due to the figure caption option being used that causes the figures/tables to float as Latex tries to find the most optimum space to render. Thus we would like the reader to be aware that a referenced figure/table can be on the following page in the report.

Initial EDA

We begin by loading in our data, saved in a file called *crime_v2.csv*. Libraries that will be used for various R functions in this report are initialized here.

```
packages <- c("car", "ggplot2", "stargazer", "scales", "gridExtra", "kableExtra",
              "float", "lmtest", "sandwich", "ggfortify")
lapply(packages, library, character.only = TRUE)

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

Lets take a quick look at the characteristics of the data set.

```
str(Crime, width = 60)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
```

```
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

```
stargazer(Crime, type = 'latex', title = "\\label{tab:crime_overview}Overview of
          variables in crime data set",
          header = FALSE)
```

Table 1: Overview of variables in crime data set

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
county	91	101.615	58.794	1.000	52.000	152.000	197.000
year	91	87.000	0.000	87.000	87.000	87.000	87.000
crmrte	91	0.033	0.019	0.006	0.021	0.040	0.099
prbarr	91	0.295	0.137	0.093	0.206	0.344	1.091
prbpris	91	0.411	0.080	0.150	0.365	0.457	0.600
avgsen	91	9.647	2.847	5.380	7.340	11.420	20.700
polpc	91	0.002	0.001	0.001	0.001	0.002	0.009
density	91	1.429	1.514	0.00002	0.547	1.568	8.828
taxpc	91	38.055	13.078	25.693	30.662	40.948	119.761
west	91	0.253	0.437	0.000	0.000	0.500	1.000
central	91	0.374	0.486	0.000	0.000	1.000	1.000
urban	91	0.088	0.285	0.000	0.000	0.000	1.000
pctmin80	91	25.495	17.017	1.284	9.845	38.142	64.348
wcon	91	285.358	47.487	193.643	250.782	314.795	436.767
wtuc	91	411.668	77.266	187.617	374.632	443.436	613.226
wtrd	91	211.553	34.216	154.209	190.864	225.126	354.676
wfir	91	322.098	53.890	170.940	286.527	345.354	509.466
wser	91	275.564	206.251	133.043	229.662	280.541	2,177.068
wmfg	91	335.589	87.841	157.410	288.875	359.580	646.850
wfed	91	442.901	59.678	326.100	400.240	478.030	597.950
wsta	91	357.522	43.103	258.330	329.325	382.590	499.590
wloc	91	312.681	28.235	239.170	297.265	329.250	388.090
mix	91	0.129	0.081	0.020	0.081	0.152	0.465
pctymle	91	0.084	0.023	0.062	0.074	0.083	0.249

We see from Table 1, that the data set contains 97 observations of 25 variables each. Most of the variables have numeric data type while there are a few categorical variables as well. The variable **prbconv** that represents probability of conviction is of **factor** data type. We believe that this is a coding error since the variable is defined as the ratio of convictions to arrests and looking at the data for other variables representing certainty of punishment **(prbarr, prbpris)**, there appears no reason for **prbconv** to have non-continuous values. We will thus convert this variable type to numeric and enable its use as a continuous random variable in regression modeling. It is also noteworthy that the variable **county** has a maximum value of 197 while there are only 97 observations in the data set. Thus, there must be skipped counties that prompts us to search for any missing values in the data set. This is important since linear model comparison and joint

Table 2: All rows with NA values in crime data set

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon	wtuc	wtrd	wfir	wser	wmfg	wfed	wsta	wloc	mix	pctymle
92	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
93	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
94	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
95	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
96	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
97	NA	NA	NA	NA	‘	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Note:
makecell[l]No values for any columns indicate coding error or missing survey results and these can be removed from analysis

hypothesis testing requires that same data is used for all unrestricted and restricted models. Even though R will do that automatically, we want to examine these data entries to avoid incorrect interpretation of results.

```
na_data = Crime[is.na(Crime$county),]
Crime_NA  = Crime[!is.na(Crime$county),]
knitr::kable(na_data, format = "latex",booktabs=T,
             caption="\\label{tab:na_data}All rows with NA values in crime data set") %>%
    kable_styling(latex_options = "scale_down" , "hold_position") %>%
    footnote(general="No values for any columns indicate coding error or missing survey
             results and these can be removed from analysis")
```

As we can see from Table 2, all of the rows between 92-97 contain `NA` values for all columns. The data are dependent on the county values as identifiers for the data set; as such, since there is no data entered in the data set for these rows, we will remove them from the main data set.

```
Crime_NA <- Crime[complete.cases(Crime),]
# set aside variable for ease of use
crmrte <- Crime_NA$crmrte
```

Now, lets examine the main variable of interest **crime rate per capita** and its distribution across counties.

```
# Different plots of crmrte
# Scatter plot
plot1 <- ggplot(Crime_NA, aes(x=county, y=crmrte)) +
        geom_point(col='steelblue4', shape=20, size=3) +
        labs(x = "County", y = "Crime Rate",
         title = "Crimes committed per capita in each county")
plot1 <- plot1 + theme(plot.title = element_text(hjust = 0.5, size=10))

# Histogram
plot2 <- qplot(crmrte, geom="histogram", bins=20,
         fill=I("steelblue4"), col=I("black"),
         xlab = "Crime Rate", ylab = "Frequency",
         main = "Distribution of crime rate across all counties",
         data=Crime_NA)
plot2 <- plot2 + theme(plot.title = element_text(hjust = 0.5, size=10))
```

```
grid.arrange(plot1, plot2, ncol=2)
```

From a visual inspection of the scatter plot in Figure 1 we see that most of the crime rates per county tend to be below the 0.050 mark, but there are several counties that have crime rate approximately double that value. The histogram in Figure 1 shows a positive skew indicating that most counties have a low crime rate. In order to understand the marginal effects of socio-economic factors on crime rate, we will perform a $log(x)$ transformation on this variable to improve the normality of the distribution. This is important to satisfy the normality of errors (MLR.6) assumption and make accurate inferences based on t and F statistics.

```
# Histogram with log(crmrte)
plot3 <- qplot(log(crmrte, 0.1), geom="histogram", bins=20,
           fill=I("steelblue4"), col=I("black"),
```

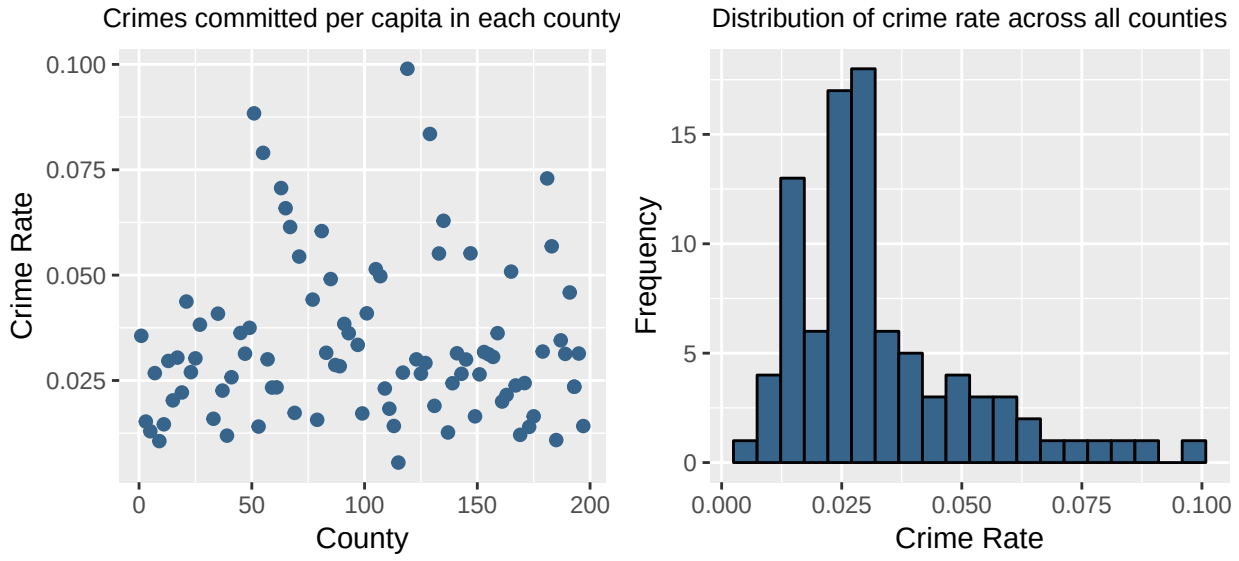

Figure 1: Scatter plot and distribution of outcome variable crime rate

```
            xlab = "Rate of Crime per County", ylab = "Frequency",
            main = "Distribution of log transformed crime rate data", data = Crime_NA)
plot3 + theme(plot.title = element_text(hjust = 0.5, size=11))
```

Following the transformation, the distribution in Figure 2 looks more normal.

Selection of Covariates

We will now examine the covariates in the data set with an aim to identify existing relationships with our outcome variable crime rate. In order to create an intuition about primary, secondary and tertiary covariates, we compute a correlation matrix of all the variables in the data set. The first row of the correlation matrix gives the co-variation of all independent variables with our outcome variable **crime rate**. We will exclude the non-numeric variables **(county, year, west, central, urban)** from the correlation matrix.

We will start the linear modeling process by looking at the variables and their individual correlations with the dependent variable **crime rate**. While there might be Multicolinearity between several of the variables, we want to focus initially on the direct relationship between the variables and **crime rate** then determine what causal effects might be absorbed by some of the variables.

```
Crime_NA$prbconv = as.numeric(as.character(Crime_NA$prbconv))

cat_vars = names(Crime_NA) %in% c("county", "year", "west","central","urban")
cor_data <- Crime_NA[!cat_vars]
cor_result <- round(cor(cor_data), 2)
first_cors <- sort(cor_result[1,], decreasing = T)

knitr::kable(first_cors,
             format = "latex",  col.names = c("Correlation"),booktabs=T,
             caption="\\label{tab:first_cors}Correlation of all
             covariates with crime rate") %>%
             kable_styling(latex_options = c("hold_position"))
```

The convenient aspect of this approach is that we are able to immediately see the different variable correlations

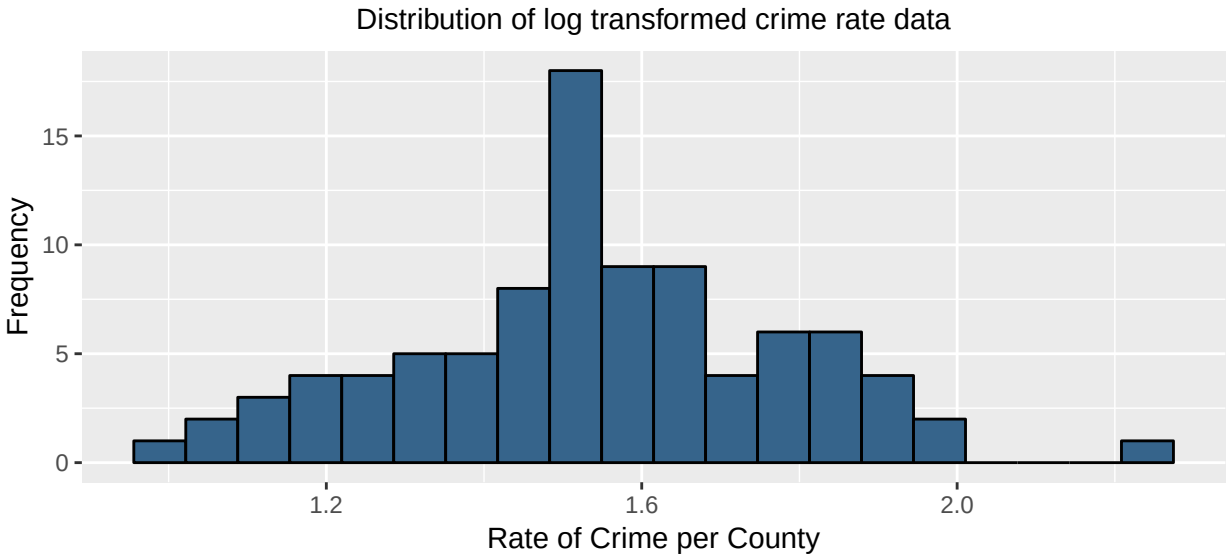

Figure 2: Histogram of log(Crime Rate)

Table 3: Correlation of all covariates with crime rate

	Correlation
crmrte	1.00
density	0.73
wfed	0.49
taxpc	0.45
wtrd	0.41
wcon	0.39
wmfg	0.35
wloc	0.35
wfir	0.33
pctymle	0.29
wtuc	0.23
wsta	0.20
pctmin80	0.19
polpc	0.17
prbpris	0.05
avgsen	0.03
wser	-0.05
mix	-0.13
prbarr	-0.39
prbconv	-0.39

directly with **crime rate**; and it helps us to select explanatory variables for regression modeling. In Table 3, one interesting aspect to note is that there are many different types of wages, ranging from strong positive correlations to weak negative correlations. We want to use **population density** and **tax revenue per capita** since they have high correlations with **crime rate**. Additionally, we will also choose **probability of arrest** and **probability of conviction** because they have high negative correlation with crime rate.

Primary Covariate EDA

To understand the nature of the selected variables, we'll look at each of them individually.

```
first_covs <- c("density", "taxpc", "prbarr","prbconv")
crime_first_covs <- names(Crime_NA) %in% first_covs
first_cov_data <- Crime_NA[crime_first_covs]
stargazer(first_cov_data, type='latex',
          title = "\\label{tab:first_covs_stats}Descriptive Statistics of Primary Covariates",
          header = FALSE)
```

Table 4: Descriptive Statistics of Primary Covariates

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
prbarr	91	0.295	0.137	0.093	0.206	0.344	1.091
prbconv	91	0.551	0.352	0.068	0.345	0.589	2.121
density	91	1.429	1.514	0.00002	0.547	1.568	8.828
taxpc	91	38.055	13.078	25.693	30.662	40.948	119.761

Of several interesting things to note from Table 4, the first is that the maximum values in variables **probability of arrest** and **probability of conviction** are greater than 1, which appears incorrect based on the general definition of probability. However these values arise due to the actual definition of these variables in the data set. These variables are proxies for certainty of punishment rather than true probabilities and are defined as ratios of crimes to arrests and arrests to convictions respectively. It is possible to get higher than 1 values for these ratios if false or incorrect arrests or erroneous convictions were made in a county.

Now we will perform EDA on our primary covariates (based on the highest correlation coefficients) to determine if any variable transformations are needed.

```
density <- Crime_NA$density
taxpc <- Crime_NA$taxpc
prbarr <- Crime_NA$prbarr
prbconv <- Crime_NA$prbconv
scatterplotMatrix(~ (density) + (taxpc) + (prbarr) + (prbconv) + log(crmrte),
                  data = Crime_NA,
                  regLine = list(col='darkred'), col=c("steelblue4"))
```

As can be seen by scatter plot matrix in Figure 3, and validated by the previous correlation chart, **population density** and **tax revenue per capita** have positive correlations with **crime rate** while **probability of arrest** and **probability of conviction** have the largest negative correlation with **crime rate**. All of the primary covariates appear to have a positive skew and we will attempt to make their distributions normal through transformations. We apply the $log(x)$ transformation for all variable except the population density. The data in population density can be better normalized using the square root transformation due to a large concentration near the min values. This transformation will allow us to achieve higher normality of errors in the linear model, and as a result, will lead to better interpretation of the p-values and other statistics.

```
scatterplotMatrix(~ sqrt(density) + log(taxpc) + log(prbarr) + log(prbconv) +
                      log(crmrte), data = Crime_NA,
                  regLine = list(col='darkred'), col=c("steelblue4"))
```

After applying the transformations, the distributions appear to be closer to normal than before (Figure 4) and we will now build our base model with these primary co-variates.

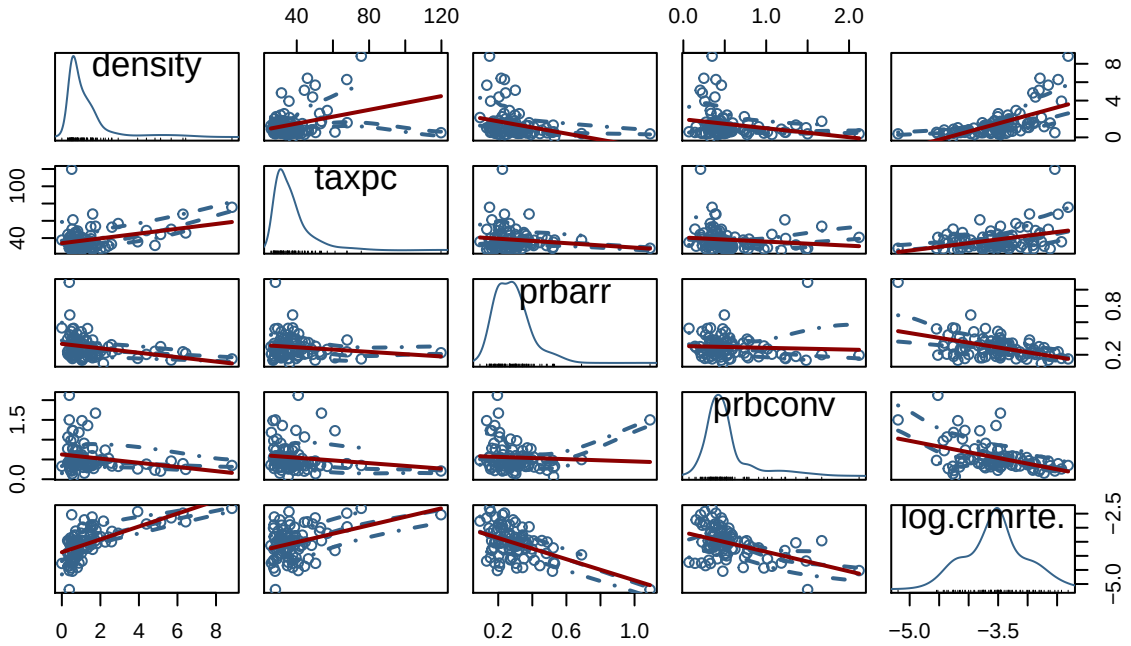

Figure 3: Scatterplot Matrix of Initial Covariates and Crime Rate

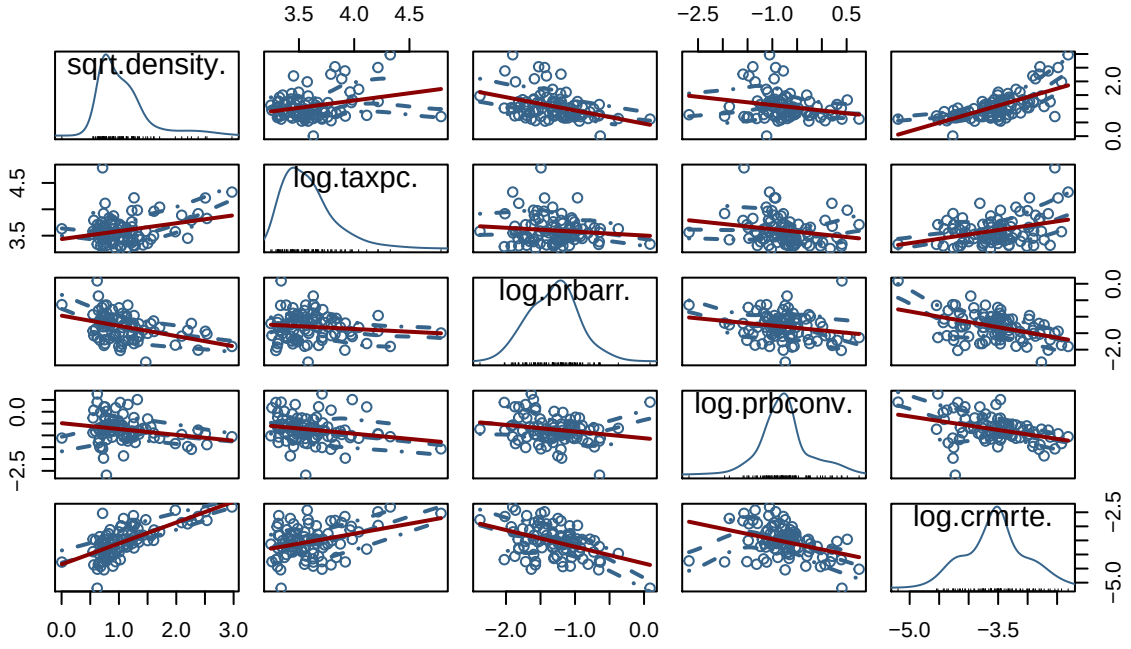

Figure 4: Scatterplot Matrix of Initial Covariates Transformed

Model 1 - Primary Covariates

The linear model for crime rate dependent on population density, tax revenue, probability of arrest and probability of conviction can be written as follows:

$$log(crmrte) = \beta_0 + \beta_1\sqrt{density} + \beta_2 log(taxpc) + \beta_3 log(prbarr) + \beta_4 log(prbconv) + u \quad (1)$$

We use the `lm` function in R for estimating the linear regression coefficients using the OLS method. Also, we will compute the robust standard errors using the `coef` function in R to safeguard against heteroskedasticity.

```
model1 = lm(log(crmrte) ~ sqrt(density) + log(taxpc) +
            log(prbarr) + log(prbconv),
            data = Crime_NA)

# Compute the robust standard errors for passing to stargazer call
se.model1 = sqrt(diag(vcovHC(model1)))
labels = c("sqrt(Population Density)", "log(Tax Revenue Per capita)",
           "log(Prob. of arrest)", "log(Prob. of conviction)")

stargazer(model1, type="latex", keep.stat=c("n",'adj.rsq'),
          title = "\\label{tab:model1_coefficients}OLS regression coefficients for Model1
          (Primary Covariates)",
          dep.var.labels = c("Log(Crime Rate)"),
          covariate.labels = labels,
          se = list(se.model1),
          star.cutoffs=c(0.05, 0.01, 0.001),
          header = FALSE)
```

Table 5: OLS regression coefficients for Model1 (Primary Covariates)

	Dependent variable:
	Log(Crime Rate)
sqrt(Population Density)	0.504^{***}
	(0.110)
log(Tax Revenue Per capita)	0.232
	(0.218)
log(Prob. of arrest)	-0.418^{**}
	(0.138)
log(Prob. of conviction)	-0.304^{*}
	(0.129)
Constant	-5.701^{***}
	(0.775)
Observations	91
Adjusted R^2	0.583
Note:	*p<0.05; **p<0.01; ***p<0.001

From Table 5 we can write the OLS regression line for predicting crime rate as:

$$log(crmrte) = -5.701 + 0.504\ \sqrt{density} + 0.232\ log(taxpc) - 0.418\ log(prbarr) - 0.304\ log(prbconv) + u \quad (2)$$

$$n = 91, R^2 = 0.583$$

Equation 2 says that as a group the four variables **density, taxpc, prbarr, prbconv** explain about 60% of the variation in **crime rate**. Each of the OLS slope coefficients has the anticipated sign. For an increase in density of 1 person per sq. mile, then holding the other factors fixed, the crime rate will increase by about 25% (note the square root in the equation). It appears that this is quite a high percentage change in the crime rate for a small change in population. However, since the equation models the percentage change in crime rate with respect to the explanatory variables, this relationship can be understood the following way. The relationship between crime rate and population density is exponential and hence the percentage change in crime rate takes on small values at the lower end of the range and attains increasingly larger values at the higher end of the range of crime rate. This is an interesting observation and says that densely populated areas in the county observe exponentially higher crime rates compared to sparsely populated areas and it is something that the political campaign can consider in policy making. A positive change in crime rate is predicted by an increase in tax revenue per capita but the interpretation of the slope coefficient is different. $\beta_2 = 0.232$ means that a 1% increase in tax revenue per capita leads to a 0.232% increase in crime rate, when all other factors are kept constant. The variables depicting the certainty of punishment have a similar interpretation as tax revenue, but the signs for their OLS slope coefficients are negative. $\beta_3 = -0.418$ and $\beta_4 = -0.304$ predict that for 1% increase in probability of arrest and probability of conviction (individually for each, keeping other factors constant), the crime rate falls by 0.418% and 0.304% respectively.

Classical Linear Model Assumptions for Model 1

```
autoplot(model1, colour="steelblue4", smooth.colour = 'darkred')
```

Since this is our first model, we will not do a deep dive here of the verification for CLM assumptions and instead save that for our final model. For our exploratory models, we will only call out the assumptions that we feel need discussion.

Looking at the ***Residuals vs Fitted*** plot in Figure 5, we can see the parabolic shape in smoothing curve which indicates there may be a violation of the zero conditional mean. This means that the estimators for our OLS slope coefficients will be biased. We can argue that as $n \to \infty$, the bias of the estimator converges to 0, giving us $cov(x_i, u) = 0$, where x_i is each of our independent variables. We can see in Table 6 that each of the co-variances are in fact 0.

```
# Create a table for the covariances
variables <- c("$\\sqrt{\\text{density}}$",
               "$log(\\text{tax per captia})$",
               "$log(\\text{probability of arrest})$",
               "$log(\\text{probability of conviction})$")

covariances <- c(
  cov(sqrt(Crime_NA$density), model1$residuals),
  cov(log(Crime_NA$taxpc), model1$residuals),
  cov(log(Crime_NA$prbarr), model1$residuals),
  cov(log(Crime_NA$prbconv), model1$residuals))

df <- data.frame(variables, covariances)
names(df)[1] = "Variable"
names(df)[2] = "Covariance"

knitr::kable(df,
```

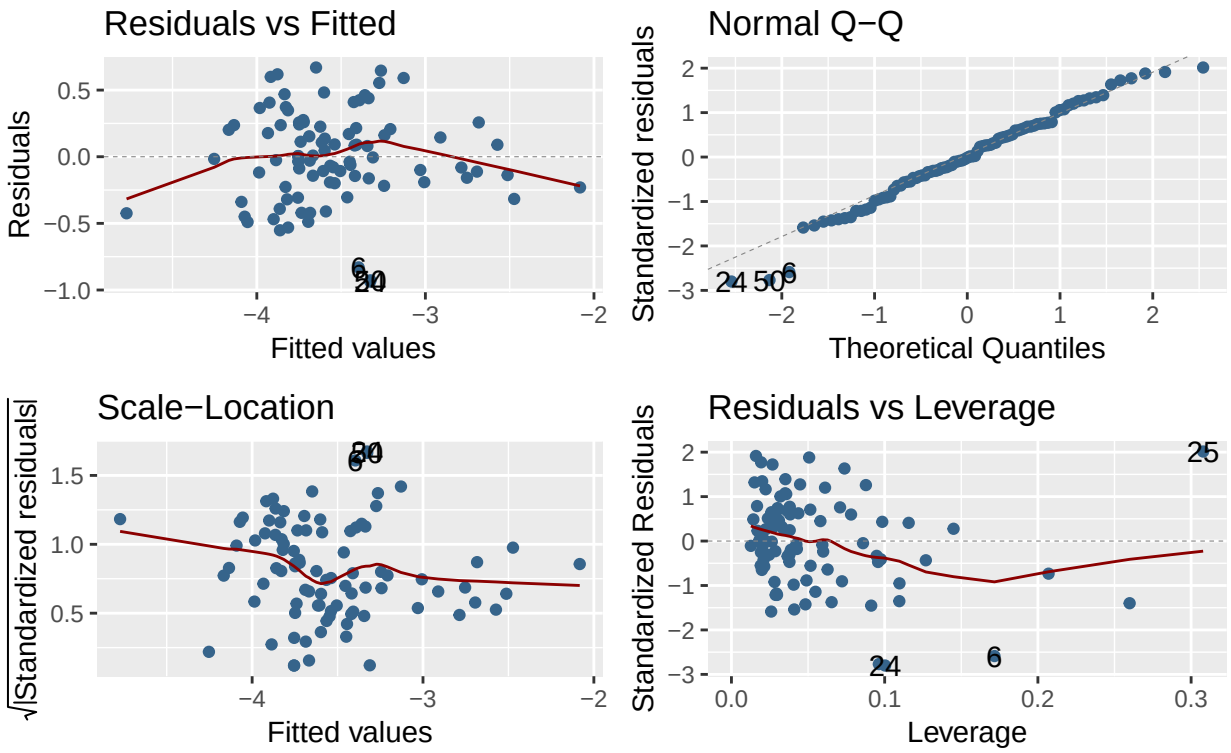

Figure 5: Diagnostic Plots for Model 1

```
            format = "latex", escape=FALSE,booktabs=T,
            caption="\\label{tab:model1_cors} Covariance of Model1 Residuals
            with Explanatory Variables") %>%
            kable_styling(latex_options = c("hold_position"))
```

Table 6: Covariance of Model1 Residuals with Explanatory Variables

Variable	Covariance
$\sqrt{\text{density}}$	0
log(tax per captia)	0
log(probability of arrest)	0
log(probability of conviction)	0

The ***Residuals vs Fitted*** plot in Figure 5 also shows appears to show some variation in the variance of the errors. We will perform a Breusch-Pagan test to verify homoskedasticity using the NULL hypothesis H_0 : there is homoskedasticity and alternate hypothesis H_A : there is heteroskedasticity

```
lmtest::bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 20.656, df = 4, p-value = 0.0003705
```

Our p-value is less than our $\alpha(= 0.05)$, so we can reject the H_0 of there being homoskedasticity. This is expected as this is a first attempt at modeling the data and shows that we are missing covariates in the

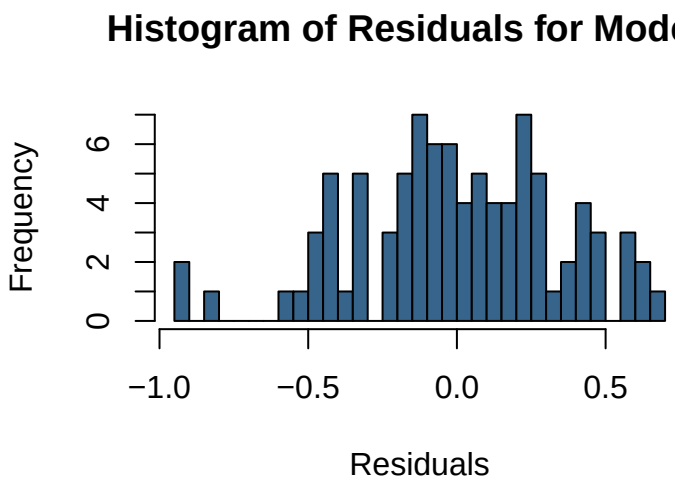

Figure 6: Histogram of Residuals for Model 1

model. We will rectify this by adding explanatory variables in model2 and model3. The Q-Q plot in Figure 5 shows that errors are approximately normally distributed. The histogram of the residuals in Figure 6 does show a slight negative skew, but we have a sufficient number of observations for the CLT to apply.

```
hist(model1$residuals, breaks=25, main='Histogram of Residuals for Model 1',
     xlab='Residuals', col="steelblue4")
```

Secondary Covariate EDA

To improve the accuracy of the model and meet the violating CLM assumptions described in the preceding sections, we will add more explanatory covariates to the equation from the data set. We examine the **percent of young males** and the **percent of minority** variables based on the magnitude of correlation from the correlation table (Table 3). Note that **police per capita** is positively correlated with crime rate. The positive correlation of 0.190 between crime rate and police presence makes intuitive sense since police are the direct responders to crime. However, it doesn't make sense to use police percentage as an explanatory variable for explaining causal effects on crime rate. Higher police presence in a county is unlikely to cause more crime. However, we do expect higher police presence in counties with higher crime rate that explain the high correlation value.

There are many wage variables in the data set that have a high correlation with **crime rate**. Since we are primarily concerned with formulating actionable policy insights for a political campaign, we propose to look at the combined effect of all wages on the crime rate. Since the range of data for each wage type is quite different from others, we will re-scale the wages between 0 and 1 before combining them.

```
wage_var_names = names(Crime_NA) %in% c("wcon", "wtuc", "wtrd", "wfir",
                                        "wser", "wmfg", "wfed", "wsta", "wloc")
wage_data <- Crime_NA[wage_var_names]
stargazer(wage_data, type = 'latex', header = FALSE,
          title = "\\label{tab:wage2_stats}Overview of data in all
          wage variables in the data set")
```

With this information, we can see that all of the wages fall roughly into the same range: [211.553 and 442.902]. Since there seems to be a high max value but normal mean value for the **weekly service wage**, we plot those values to visualize the data in this variable.

```
plot <- ggplot(Crime_NA, aes(x=county, y=wser)) +
    geom_point(col='steelblue4', shape=20, size=3) +
```

Table 7: Overview of data in all wage variables in the data set

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
wcon	91	285.358	47.487	193.643	250.782	314.795	436.767
wtuc	91	411.668	77.266	187.617	374.632	443.436	613.226
wtrd	91	211.553	34.216	154.209	190.864	225.126	354.676
wfir	91	322.098	53.890	170.940	286.527	345.354	509.466
wser	91	275.564	206.251	133.043	229.662	280.541	2,177.068
wmfg	91	335.589	87.841	157.410	288.875	359.580	646.850
wfed	91	442.901	59.678	326.100	400.240	478.030	597.950
wsta	91	357.522	43.103	258.330	329.325	382.590	499.590
wloc	91	312.681	28.235	239.170	297.265	329.250	388.090

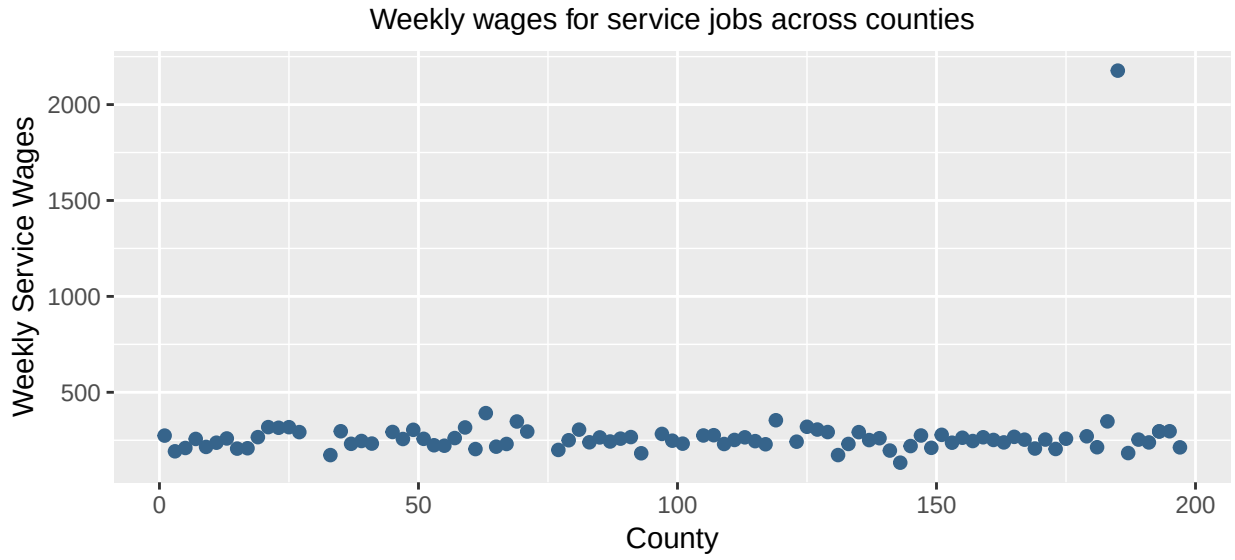

Figure 7: Weekly Service Wages by County

```
    xlab("County") + ylab("Weekly Service Wages") +
    ggtitle("Weekly wages for service jobs across counties")

plot + theme(plot.title = element_text(hjust = 0.5, size=11))
```

It's evident from Figure 7 that the max value of the **weekly service wage** is an outlier and scaling this data on the same scale as the other wages could affect the final slope coefficient estimates for **weekly service wage**. We will therefore not combine this variable with the other wages and exclude it from **model2** regression. We will use this variable separately in the third model to understand its impact on crime rate predictions.

```
# Rescale
wages <- rescale(Crime_NA$wfed,c(0,1)) + rescale(Crime_NA$wtrd,c(0,1)) +
    rescale(Crime_NA$wcon,c(0,1)) + rescale(Crime_NA$wmfg,c(0,1)) +
    rescale(Crime_NA$wloc,c(0,1)) + rescale(Crime_NA$wfir,c(0,1)) +
    rescale(Crime_NA$wtuc,c(0,1)) + rescale(Crime_NA$wsta,c(0,1))
```

With the covariates decided, we create a scatter plot matrix of these covariates and our outcome variable **crime rate**.

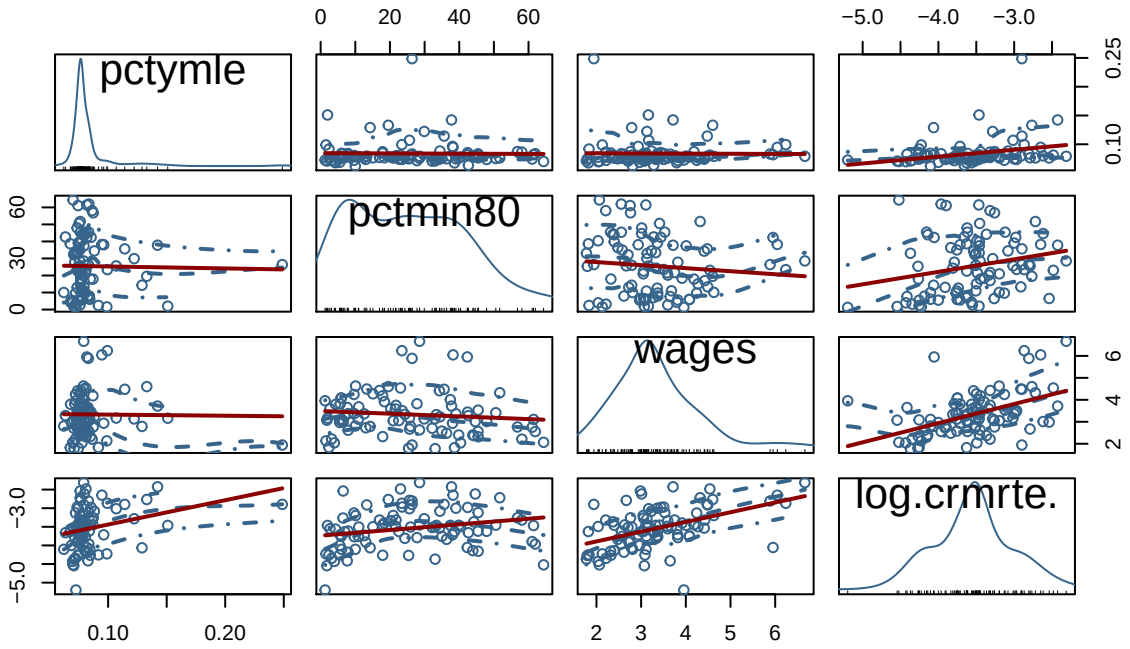

Figure 8: Scatterplot Matrix of Secondary Covariates

```
scatterplotMatrix(~ pctymle + pctmin80 + wages + log(crmrte), data = Crime_NA,
                  regLine = list(col='darkred'), col=c("steelblue4"))
```

Visualizing the scatter plots, we notice positive skew for the variable **percent young male**. We will apply the $log(x)$ transformation to it, in an effort to make it more normal.

```
scatterplotMatrix(~ log(pctymle) + pctmin80 + wages + log(crmrte), data = Crime_NA,
                  regLine = list(col='darkred'), col=c("steelblue4"))
```

Following the transformation, the data for **percent young male** looks more normal than before.

Model 2 Estimation - Secondary Covariates

Adding the additional covariates the second linear model has the form:

$$\begin{aligned} log(crmrte) = \beta_0 &+ \beta_1\sqrt{density} + \beta_2 log(taxpc) + \beta_3 log(prbarr) \\ &+ \beta_4 log(prbconv) + \beta_5 log(pctymle) + \beta_6 pctmin80 \\ &+ \beta_7 wages + u \end{aligned} \tag{3}$$

We then use the `lm` function in R to fit OLS regression line for these covariates:

```
model2 = lm(log(crmrte) ~ sqrt(density) + log(taxpc) +
              log(prbarr) + log(prbconv) + log(pctymle) +
              (pctmin80) + wages,
            data = Crime_NA)

# Compute the robust standard errors
se.model2 = sqrt(diag(vcovHC(model2)))

# Pass the robust errors to stargazer
```

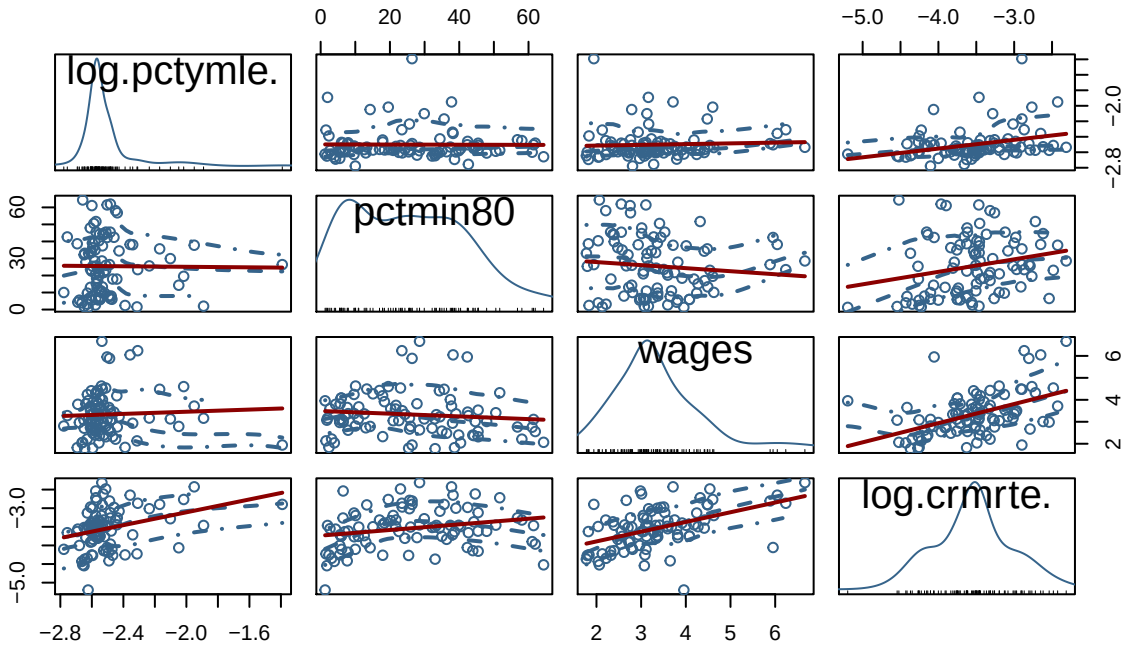

Figure 9: Scatterplot Matrix of Secondary Covariates Transformed

```
stargazer(model1, model2, type="latex",
          title = "OLS regression coefficients (Model1 and Model2) comparison",
          keep.stat=c('adj.rsq','n'), header = FALSE,
          se = list(se.model1, se.model2),
          star.cutoffs=c(0.05, 0.01, 0.001))
```

The adjusted R-squared value shows an increase from 0.583 in **model1** to 0.710 in **model2** which indicates that the additional variables explain more variation in crime rate than our first model. The fitted OLS regression line can be written as:

$$\begin{aligned} log(crmrte) = -5.530 + 0.438\sqrt{density} + 0.167log(taxpc) - 0.449log(prbarr) \\ - 0.343log(prbconv) + 0.173log(pctymle) + 0.011pctmin80 \\ + 0.062wages + u \end{aligned} \tag{4}$$

$$n = 91, R^2 = 0.710$$

The OLS slope coefficients in **model2** have the expected signs. We see that the magnitude of the slope coefficients for population density and tax revenue percentage have decreased while those for probability of arrest and conviction have increased. The explanation for the OLS slope coefficient estimators obtained for the newly added variables **percentage of young male**, **percentage of minnonrities** and **wages**, goes as follows: A 1% increase in young male percentage is predicted to increase the **crime rate** by 0.17% when other factors are kept constant. Similarly, a unit change in minority percentage is estimated to increase crime rate by 1.1%. Lastly, the combined and scaled wages variables is estimated to increase the crime rate by 6.2%. We refrain from making inferences about these effects until we have fitted our third model with all the remaining covariates.

Classical Linear Model Assumptions for Model 2

```
autoplot(model2, colour = 'steelblue4', smooth.colour = 'darkred')
```

Table 8: OLS regression coefficients (Model1 and Model2) comparison

	Dependent variable:	
	log(crmrte)	
	(1)	(2)
sqrt(density)	0.504^{***}	0.438^{**}
	(0.110)	(0.149)
log(taxpc)	0.232	0.167
	(0.218)	(0.294)
log(prbarr)	-0.418^{**}	-0.449^{***}
	(0.138)	(0.104)
log(prbconv)	-0.304^{*}	-0.343^{***}
	(0.129)	(0.104)
log(pctymle)		0.173
		(0.209)
pctmin80		0.011^{***}
		(0.002)
wages		0.062
		(0.068)
Constant	-5.701^{***}	-5.530^{***}
	(0.775)	(1.161)
Observations	91	91
Adjusted R^2	0.583	0.710
Note:	*p<0.05; **p<0.01; ***p<0.001	

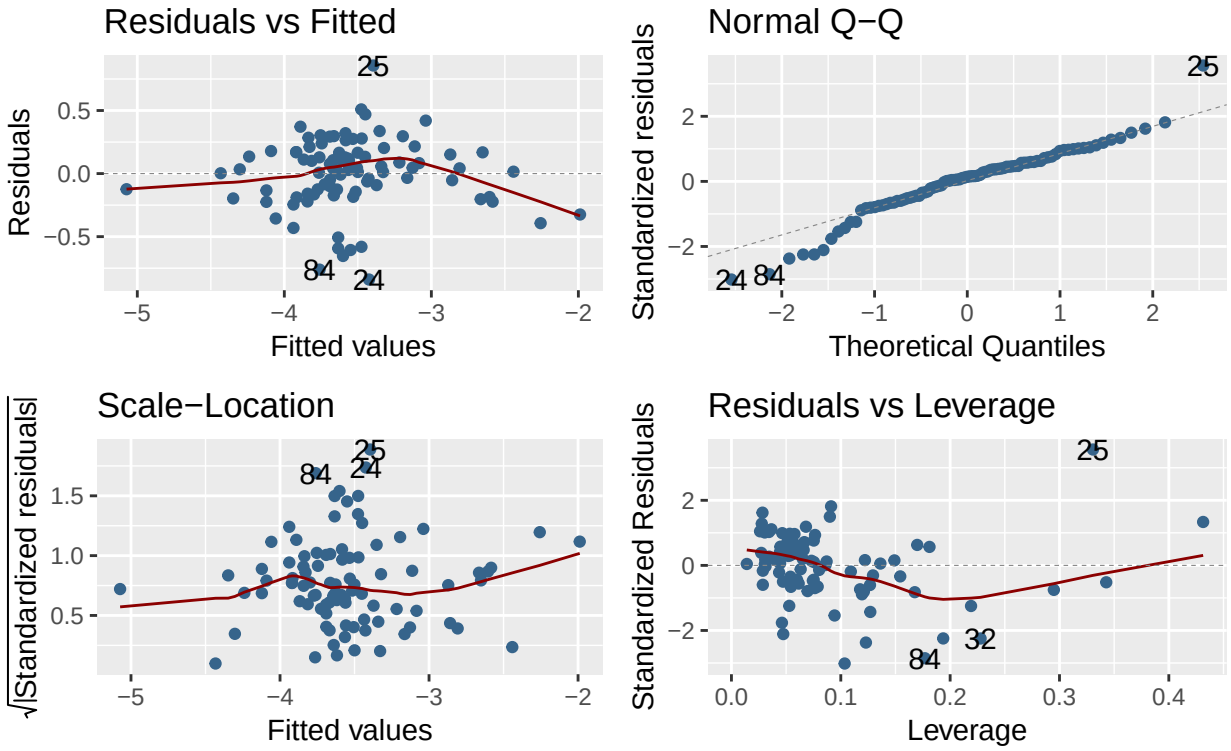

Figure 10: Diagnostic Plots for Model 2

As noted before, we will not do a deep dive here of the verification for CLM assumptions and instead save that for our final model. For our exploratory models, we will only call out the assumptions that we feel need discussion. We can perform a Breusch-Pagan test to verify homoskedasticity using the NULL and alternate hypotheses H_0 : there is homoskedasticity and H_A : there is heterskedasticity.

```
lmtest::bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 25.834, df = 7, p-value = 0.0005393
```

Our p-value is less than than our $\alpha(= 0.05)$, so we must reject the H_0 of there being homoskedasticity. We will continue to refine the model to achieve homoskedasticity. The Q-Q plot in Figure 10 shows some deviation from normality on the extremes. The histogram of the residuals in Figure 11 does show a slight negative skew, but we do have a sufficient number of observations for the CLT to apply.

```
hist(model2$residuals, breaks=25, main='Histogram of Residuals for Model 2',
     xlab='Residuals', col="steelblue4")
```

Teritiary Covariate EDA

For our final model, we want to determine if our model can be further improved. To that end, we will incorporate majority of the remaining variables in our data set, including the categorical variables. First, we will examine the remaining numerical variables and examine the relationship they have with **crime rate**.

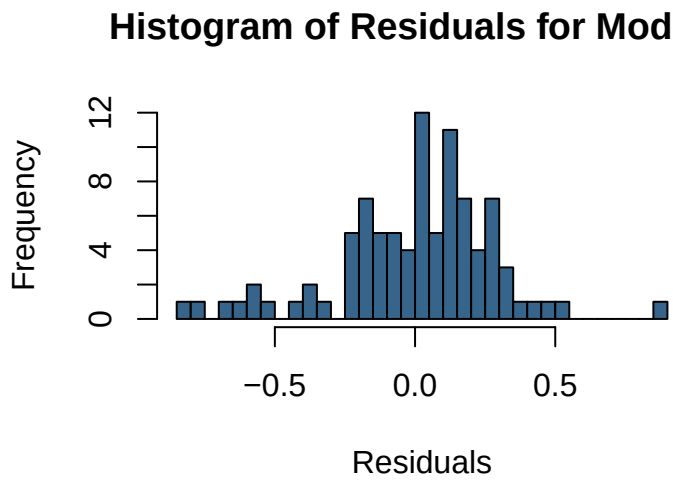

Figure 11: Histogram of Residuals for Model 2

```
rem_var_names = names(Crime_NA) %in% c("prbpris", "avgsen", "mix")
rem_data <- Crime_NA[rem_var_names]

stargazer(rem_data, type = 'latex',
          title = "Descriptive statistics of additional covariates in model2", header = FALSE)
```

Table 9: Descriptive statistics of additional covariates in model2

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
prbpris	91	0.411	0.080	0.150	0.365	0.457	0.600
avgsen	91	9.647	2.847	5.380	7.340	11.420	20.700
mix	91	0.129	0.081	0.020	0.081	0.152	0.465

```
ggplot(Crime_NA, aes(x=county, y=avgsen)) +
    geom_point(col='steelblue4', shape=20, size=3) +
    xlab("County") + ylab("Average Sentence")
```

We can see that for **average sentence** in Figure 12, the max seems to be quite high compared to the mean and standard deviations. The scatter plot does show some values are seem to be outliers, but we don't believe these are extreme.

```
bp1 <- ggplot(Crime_NA, aes(x=as.factor(urban), y=crmrte)) +
    geom_point(col='steelblue4', shape=20, size=3) +
    xlab("Urban") + ylab("Crime Rate") + geom_boxplot() +
    scale_x_discrete(labels=c("0" = "not urban", "1" = "urban"))

bp2 <- ggplot(Crime_NA, aes(x=as.factor(west), y=crmrte)) +
    geom_point(col='steelblue4', shape=20, size=3) +
    xlab("West") + ylab("Crime Rate") + geom_boxplot() +
    scale_x_discrete(labels=c("0" = "not west", "1" = "west"))

bp3 <- ggplot(Crime_NA, aes(x=as.factor(central), y=crmrte)) +
    geom_point(col='steelblue4', shape=20, size=3) +
    xlab("Central") + ylab("Crime Rate") + geom_boxplot() +
```

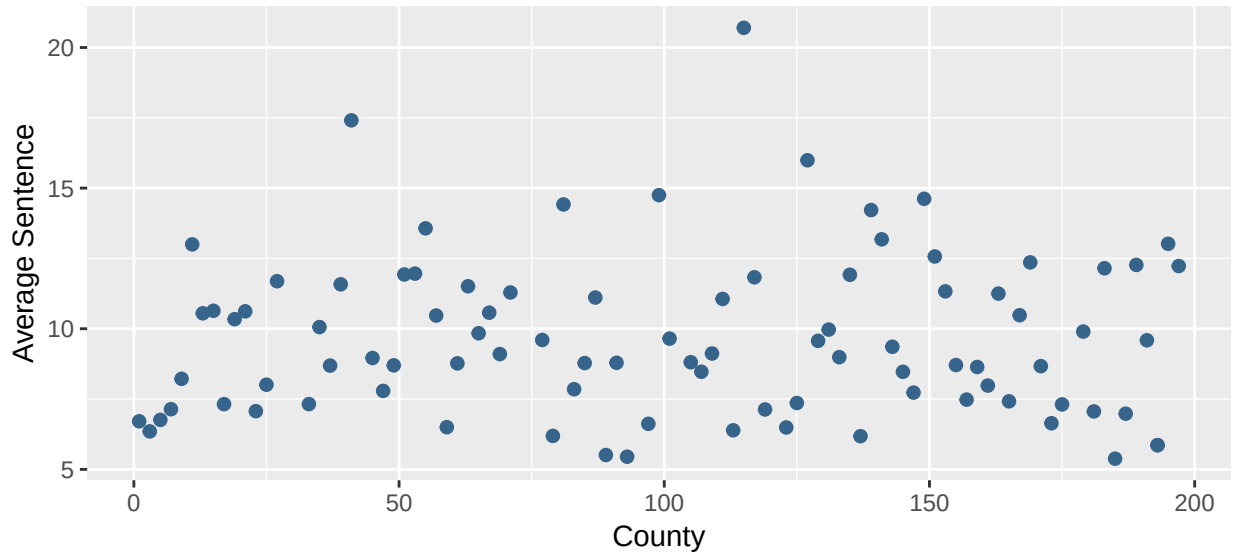

Figure 12: Plot of average Sentence versus county

```
    scale_x_discrete(labels=c("0" = "not central", "1" = "central"))

grid.arrange(bp1, bp2, bp3, ncol=3)
```

The box plots of the regional data **urban**, **west**, and **central** in Figure 13 indicate that the largest discrepancy of **crime rate** appears between urban and rural areas. As such, the **urban** descriptor would be important to consider for our model.

```
scatterplotMatrix(~ prbpris + avgsen + mix + wser + log(crmrte), data = Crime_NA,
                  regLine = list(col='darkred'), col=c("steelblue4"))
```

Figure 14 shows scatter plots of the tertiary Covariates and the outcome variable **crime rate**. Note that we will also reintroduce **weekly service wages** back into the model to help incorporate all the wage variables in our examination. We can see that **average sentence**, **mix**, and **weekly service wages** seem to be skewed positively in Figure 14, so we will apply the $log(x)$ transform to them. Since **urban** is categorical, it is excluded from this scatter plot matrix.

Model 3 Estimation - Tertiary Covariate

The third model can be written as :

$$\begin{aligned} log(crmrte) = \beta_0 &+ \beta_1\sqrt{density} + \beta_2 log(taxpc) + \beta_3 log(prbarr) \\ &+ \beta_4 log(prbconv) + \beta_5 log(pctymle) + \beta_6 pctmin80 \\ &+ \beta_7 wages + \beta_8 prbpris \\ &+ \beta_9 log(avgsen) + \beta_{10} log(mix) \\ &+ \beta_{11} log(wser) + \beta_{12}(urban) + u \end{aligned} \tag{5}$$

Putting this into R's `lm` command, we can fit an OLS regression line for the third model:

```
model3 = lm(log(crmrte) ~ sqrt(density) + log(taxpc) +
              log(prbarr) + log(prbconv) + log(pctymle) +
              (pctmin80) + (wages) +
```

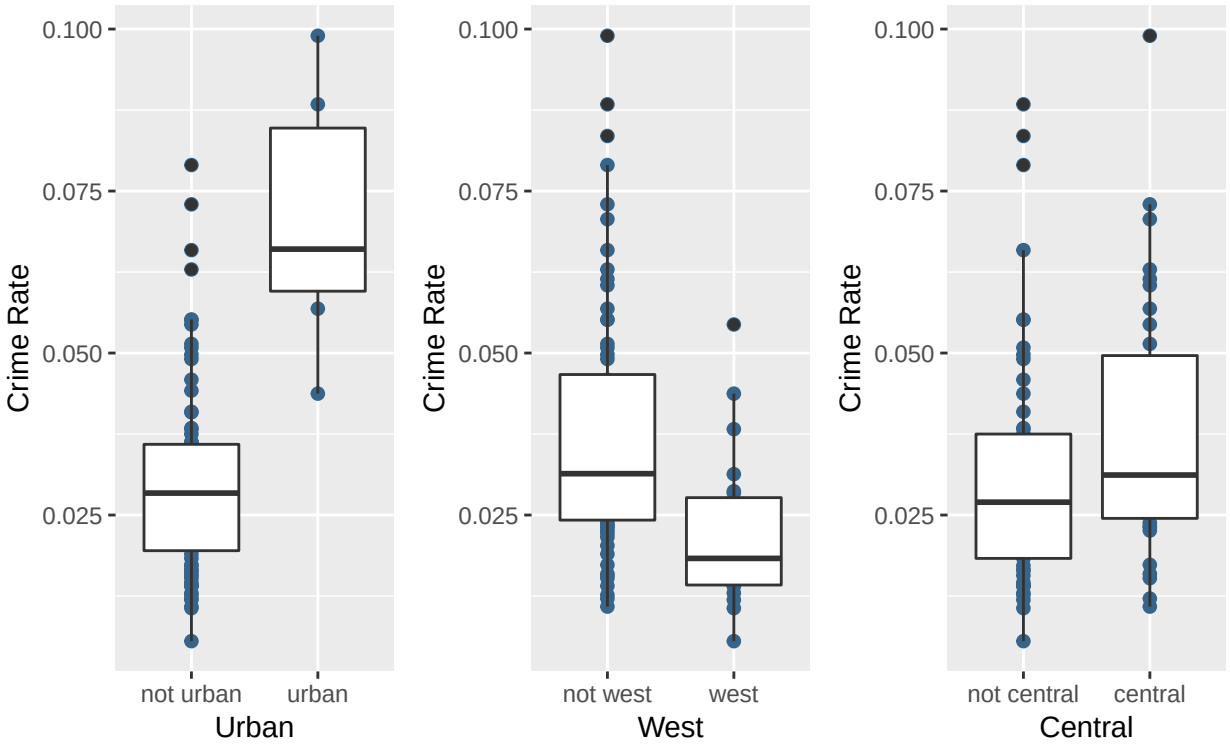

Figure 13: Boxplots of crimes rate distribution versus region categories

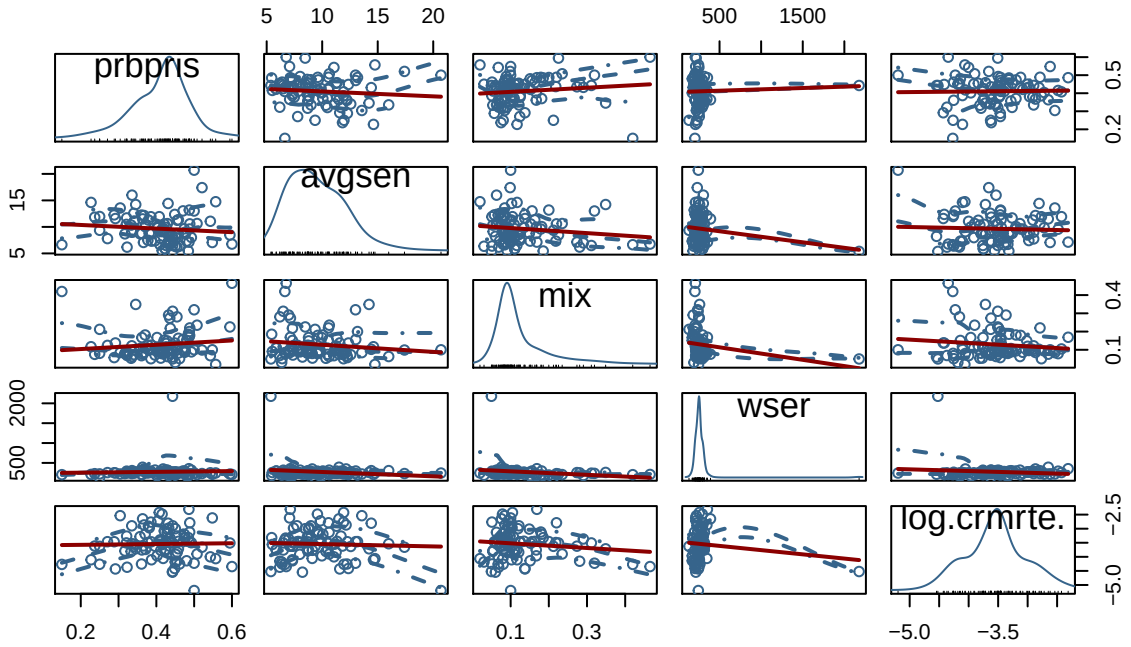

Figure 14: Scatterplot Matrix of Tertiary Covariates

```
                (prbpris) + log(avgsen) + log(mix) +
                log(wser) + factor(urban),
              data = Crime_NA)

# Compute the robust standard errors
se.model3 = sqrt(diag(vcovHC(model3)))
# Pass the robust errors to stargazer
stargazer(model1, model2, model3, type="latex",
          title="OLS regression coefficient comparison (Model1, Model2, Model3)",
          keep.stat=c('adj.rsq','n'), header = FALSE,
          se = list(se.model1, se.model2,se.model3),
          star.cutoffs=c(0.05, 0.01, 0.001))
```

The adjusted R-squared value increased to 0.743 in **model3** and it is noteworthy that this increase is not as large as the one observed between **model1** and **model2**. It should also be noted that the variables that have the greatest negative impact are related to probability of arrests, probability of convictions and average sentence. The definition of probability of arrest is directly tied to crime rate, as an example. This means the coefficients of those variables might confound the explanation for the variance of **crime rate**, i.e. might affect the R^2 values in ways that aren't immediately obvious. Additionally, the probabilities of the events further along the justice system are incumbent on the step before, meaning that the probability of conviction is almost wholly based on an arrest, and prison sentencing based off of conviction, and so on.

We can now write the OLS regression line for **model3** as:

$$\begin{aligned} log(crmrte) = -3.848 &+ 0.565\sqrt{density} + 0.309log(taxpc) - 0.479log(prbarr) \\ &- 0.282log(prbconv) + 0.250log(pctymle) + 0.012pctmin80 \\ &+ 0.430wages - 0.496prbpris \\ &- 0.097log(avgsen) + 0.070log(mix) \\ &- 0.331log(wser) - 0.340(urban) + u \\ &n = 91, R^2 = 0.710 \end{aligned} \tag{6}$$

The OLS slope coefficients for service wages doesn't seem to have the expected sign. We expect the crime rate to increase in areas with higher service wages. However, we do see that the coefficient is not statistically significant and we will test whether it can be removed from the final model without affect the accuracy of our causal predictions.

Classical Linear Model Assumptions for Model 3

```
autoplot(model3, colour = 'steelblue4', smooth.colour = 'darkred')
```

As noted before, we will not do a deep dive here of the verification for CLM assumptions and instead save that for our final model. For our exploratory models, we will only call out the assumptions that we feel need discussion.

We can perform a Breusch-Pagan test to verify the homoskedasticity using the NULL and alternate hypothesis as H_0 : there is homoskedasticity and H_A : there is heteroskedasticity.

```
lmtest::bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
```

Table 10: OLS regression coefficient comparison (Model1, Model2, Model3)

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
sqrt(density)	0.504^{***}	0.438^{**}	0.644^{***}
	(0.110)	(0.149)	(0.146)
log(taxpc)	0.232	0.167	0.315
	(0.218)	(0.294)	(0.287)
log(prbarr)	-0.418^{**}	-0.449^{***}	-0.465^{***}
	(0.138)	(0.104)	(0.087)
log(prbconv)	-0.304^{*}	-0.343^{***}	-0.273^{*}
	(0.129)	(0.104)	(0.107)
log(pctymle)		0.173	0.221
		(0.209)	(0.174)
pctmin80		0.011^{***}	0.012^{***}
		(0.002)	(0.002)
wages		0.062	0.082
		(0.068)	(0.064)
prbpris			−0.436
			(0.362)
log(avgsen)			−0.068
			(0.127)
log(mix)			0.053
			(0.088)
log(wser)			−0.329
			(0.215)
factor(urban)1			-0.400^{*}
			(0.199)
Constant	-5.701^{***}	-5.530^{***}	-3.907^{*}
	(0.775)	(1.161)	(1.595)
Observations	91	91	91
Adjusted R^2	0.583	0.710	0.743
Note:		*p<0.05; **p<0.01; ***p<0.001	

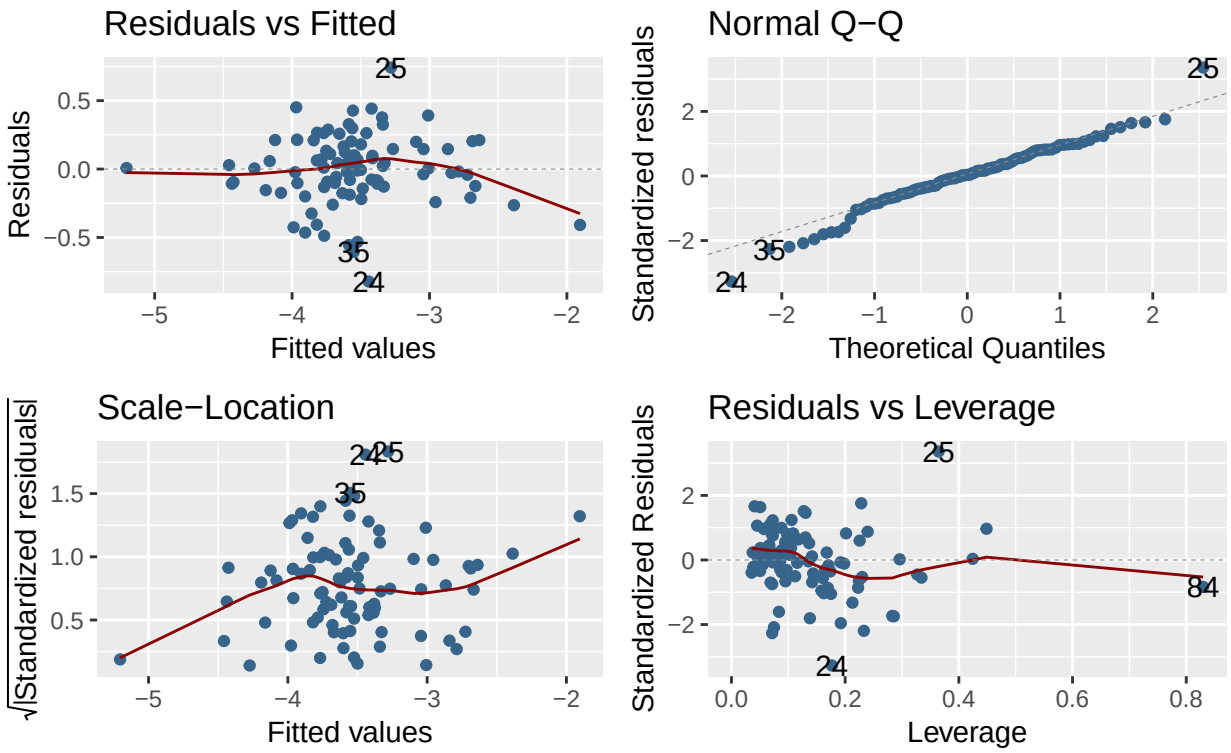

Figure 15: Diagnostic Plots for Model 3

```
## BP = 26.168, df = 12, p-value = 0.01016
```

Our p-value is less than our $\alpha(= 0.05)$, so we can reject the H_0 of there being homoskedasticity. The Q-Q plot in Figure 15 shows some deviation from normality on the extremes. The histogram of the residuals in Figure 16 does show a slight negative skew, but we do have a sufficient number of observations for the CLT to apply.

```
hist(model3$residuals, breaks=25, main='Histogram of Residuals for Model 3',
     xlab='Residuals', col="steelblue4")
```

Omitted and Excluded Variables

1. The **county identifier** and **year** variables were excluded from the linear models. These county identifier appears to be simply included for keeping track of the records while the year data is constant across all rows.
2. Coefficients in **model2** show a positive linear relationship with percent of young male (**pctymle**) and percent minority(**pctmin80**):
    - Positive coefficients for these covariates indicate a positive relationship with crime rate. However, it is not apparent as why the increase in young male or minority population in a county would lead to higher crime rate. We suspect that there are omitted variables in the data set that bias the coefficients for these two variable. One such variable could be the income/wages of young male and minorities in each of the counties. Intuitively it might help to understand the economic status of the these segments of the county population and may shed some light on the causal effects of these two variables on the outcome variable **crime rate**. The omitted wage variable for young male or minorities can be have a positive or negative linear coefficient.
    - If it turns out to have a positive coefficient then it would mean that existing linear coefficients for

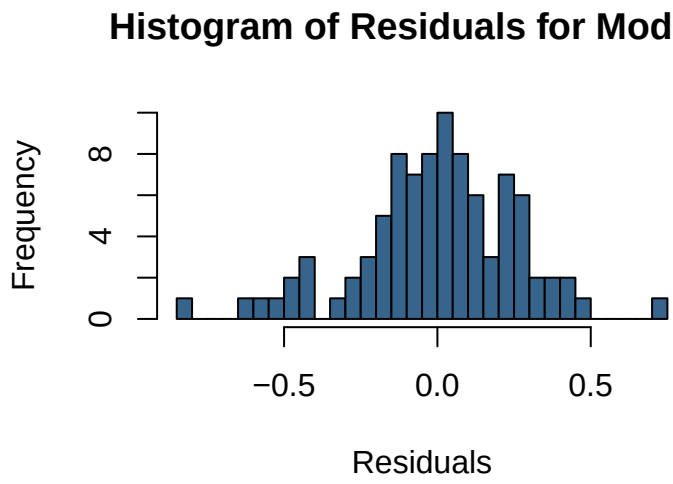

Figure 16: Histogram of Residuals for Model 3

these covariates in **model2** have a positive bias and we are overestimating their **negative** effect on crime rate. If the omitted variable of young male or minority wages has a negative bias then we are underestimating the effect of young male and minority population on county's crime rate.

3. Positive linear coefficient for police per capita (**polpc**):
   - Linear **model3** for crime rate shows a positive coefficient for the police per capita variable. This doesn't make intuitive sense and we suspect that there is an omitted variable in the data, with a negative correlation to crime rate. One such variable could be the percent of population that is unemployed in the county. We expect this variable to have a negative linear coefficient for the crime rate model and reduce the coefficient value related to police per capita values.

Final Model Selection and Hypothesis Testing

We will use **model3** as the starting point for converging on our final model. The function `coeftest` will be used to compute robust standard errors and p-values for all the OLS coefficients.

```
model3_coefs = coeftest(model3, vcov=vcovHC)[,]
knitr::kable(model3_coefs, format="latex",booktabs=T,
              escape=FALSE, caption="\\label{tab:model3_pvals}Robust standard
             errors and p-values for model3 estimated OLS coefficients") %>%
             kable_styling(latex_options = c("hold_position"))
```

We see from Table 11 that coefficients for **density, probability of arrest**, and **percentage of minorities** are highly statistically significant (even at lower than 1% significance level). The coefficients for **conviction probability** and **urban neighborhoods** are only significant at the 10% level while the other variable coefficients are not statistically significant at any level. Thus, we need to make adjustments to our model, and we would like to remove these variables from the final model to improve the parsimony. However, we first need to test whether these variables are jointly significant for predicting crime rate. We formulate the following NULL and alternate hypothesis for this test:

$$H_0 : \beta_2 = 0, \beta_5 = 0, \beta_7 = 0, \beta_8 = 0, \beta_9 = 0, \beta_{10} = 0, \beta_{11} = 0$$

$$H_a : \text{At least one of the OLS coefficients is not zero}$$

Table 11: Robust standard errors and p-values for model3 estimated OLS coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.9073773	1.5945980	-2.4503840	0.0165097
sqrt(density)	0.6435018	0.1456309	4.4187175	0.0000317
log(taxpc)	0.3154566	0.2869536	1.0993298	0.2750058
log(prbarr)	-0.4645758	0.0866566	-5.3611099	0.0000008
log(prbconv)	-0.2734213	0.1065918	-2.5651255	0.0122320
log(pctymle)	0.2207948	0.1735855	1.2719652	0.2071647
pctmin80	0.0119809	0.0020229	5.9227633	0.0000001
wages	0.0815964	0.0637233	1.2804790	0.2041714
prbpris	-0.4362946	0.3618727	-1.2056577	0.2315942
log(avgsen)	-0.0683887	0.1271675	-0.5377841	0.5922572
log(mix)	0.0527449	0.0883508	0.5969938	0.5522405
log(wser)	-0.3291145	0.2153779	-1.5280789	0.1305382
factor(urban)1	-0.4002701	0.1990774	-2.0106249	0.0478212

```
joint_hypothesis1 = linearHypothesis(model3, c("log(taxpc)=0", "log(pctymle)=0","wages=0",
                    "prbpris=0","log(avgsen)=0","log(mix)=0","log(wser)=0"), vcov=vcovHC)

knitr::kable(joint_hypothesis1,
  format="latex",  escape=FALSE,booktabs=T,
  caption="\\label{tab:joint_hypothesis1}Joint hypothesis
  testing (F-test) for irrelevant variables added to model3") %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 12: Joint hypothesis testing (F-test) for irrelevant variables added to model3

Res.Df	Df	F	Pr(>F)
85	NA	NA	NA
78	7	1.002511	0.4360509

Results from table 12 show that we cannot reject the NULL hypothesis that all the variables are jointly zero. Thus in an attempt to improve the inference we will remove these variables from our model and run OLS regression again to estimate the new coefficients for equation 7.

$$\begin{aligned} log(crmrte) = \beta_0 + \beta_1\sqrt{density} + \beta_2 log(prbarr) \\ + \beta_4 log(prbconv) + \beta_6 pctmin80 \\ + \beta_{12}(urban) + u \end{aligned} \tag{7}$$

```
# Fit an OLS regression line for the model
model4 = lm(log(crmrte) ~ sqrt(density) + log(prbarr) +
           + log(prbconv) + (pctmin80) + factor(urban),
           data = Crime_NA)

# Compute the robust standard errors
se.model4 = sqrt(diag(vcovHC(model4)))

# Pass the robust errors to stargazer
stargazer(model3, model4, type = 'latex', header = FALSE,
```

```
          title="\\label{tab:model4_reg_output}OLS regression coefficient table
          comparing restricted and unrestricted versions of model3",
          keep.stat=c('adj.rsq','n', 'ser','f'),
          se = list(se.model3,se.model4),
          star.cutoffs=c(0.1, 0.05, 0.01))
```

From the output regression table 13, we note that the slope coefficients maintain their sign between the restricted and unrestricted models. The largest magnitude change in slope coefficients is observed for **probability of conviction**. The prediction of the negative impact of this variables on **crime rate** has increased by approximately 10% after removing the statistically insignificant variables from the model. We also notice that the statistical significance of **urban** has reduced in the restricted model and its slope coefficient is only significant at 10% level now (compared to 5% level in unrestricted model). The adjusted R^2 value has decreased in the restricted model by approximately 3% which is a good trade off with the reduction in number of explanatory variables. Since the significance level is subject to researcher's interpretation bias, we report the p-values for the slope coefficient's t-statistics below:

```
final_model_pvalues = coeftest(model4, vcov=vcovHC)[,c(1,4)]
slope_values = format(final_model_pvalues[,1], digits=2)
p_values = format(final_model_pvalues[,2], digits=3)

result_frame = data.frame(slope_values,p_values)
row.names(result_frame) = c('Intercept','$\\sqrt{Density}$','log(Probability of Arrest)',
                            'log(Probability of Conviction)',
                            'Percentage of Minority in 1980s','Urban')

knitr::kable(result_frame, format="latex",booktabs=T,
              escape=FALSE, col.names=c("Coefficient Estimate","P-Value"),
             align = rep('c'), caption="\\label{tab:final_model_pvalues}
             P-values for OLS slope coefficients in final linear model") %>%
             kable_styling(latex_options = c("hold_position"))
```

The final model equation for predicting *crime rate* can be written as follows:

$$\begin{aligned} log(crmrte) = -5.460 + 0.682\sqrt{density} - 0.472log(prbarr) - 0.365log(prbconv) \\ + 0.012(pctmin80) - 0.294(urban) + u \end{aligned} \tag{8}$$

$$n = 91, R^2 = 0.712$$

The final linear model for predicting crime rate is represented by equation 8. The r-squared value from OLS regression has a value of 0.712 which means that our explanatory variables explain 71.2% of the variation in the crime rate sample data. The population density seems to be the biggest factor influencing crime rate. Keeping other factors constant, a 1 unit increase in population density causes about approximately 34% increase in crime rate. A 1% increase in probability of arrest or probability of conviction (ceteris paribus) causes approximately 0.47% and 0.36% decrease in crime rate respectively. Similarly, the effect of minority percentage is also very small since a unit increase in minority percentage will only cause a 1.2% increase in crime rate. Thus, even though the OLS coefficients for certainty of punishment and percentage of minority have statistical significance, they have very little practical significance. The variable `urban` is a factor with only two levels: 0 and 1. Hence, its OLS slope coefficient for only adds to the intercept of our model when urban = 1. It is worth mentioning that the intercept of the final regression line is negative. Since the outcome variable is the *log* of **crime rate**, this means that the when values of all explanatory variables are zero, the **crime rate** is $e^{-5.46} = 0.00425$. This value is 1.5 standard deviations above the minimum crime rate in the sample data set. Thus, a small amount of crime is predicted by our model to exist in all counties before the affect of any explanatory variables is considered.

Table 13: OLS regression coefficient table comparing restricted and unrestricted versions of model3

	Dependent variable:	
	log(crmrte)	
	(1)	(2)
sqrt(density)	0.644***	0.682***
	(0.146)	(0.129)
log(taxpc)	0.315	
	(0.287)	
log(prbarr)	−0.465***	−0.472***
	(0.087)	(0.094)
log(prbconv)	−0.273**	−0.365***
	(0.107)	(0.105)
log(pctymle)	0.221	
	(0.174)	
pctmin80	0.012***	0.012***
	(0.002)	(0.002)
wages	0.082	
	(0.064)	
prbpris	−0.436	
	(0.362)	
log(avgsen)	−0.068	
	(0.127)	
log(mix)	0.053	
	(0.088)	
log(wser)	−0.329	
	(0.215)	
factor(urban)1	−0.400**	−0.294*
	(0.199)	(0.160)
Constant	−3.907**	−5.460***
	(1.595)	(0.106)
Observations	91	91
Adjusted R^2	0.743	0.712
Residual Std. Error	0.277 (df = 78)	0.293 (df = 85)
F Statistic	22.645*** (df = 12; 78)	45.587*** (df = 5; 85)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 14: P-values for OLS slope coefficients in final linear model

	Coefficient Estimate	P-Value
Intercept	-5.460	6.54e-66
$\sqrt{Density}$	0.682	1.03e-06
log(Probability of Arrest)	-0.472	2.64e-06
log(Probability of Conviction)	-0.365	8.12e-04
Percentage of Minority in 1980s	0.012	2.77e-07
Urban	-0.294	6.96e-02

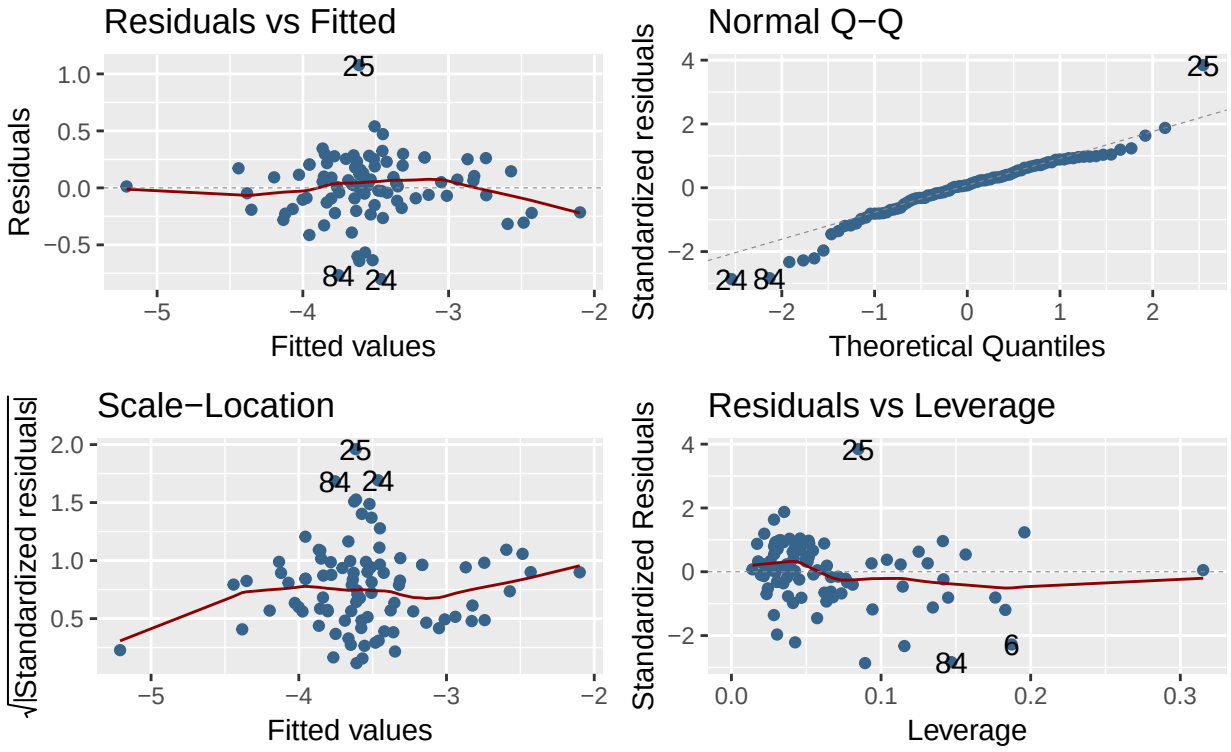

Figure 17: Diagnostic Plots for final model

MLR Assumption Validation

We will now examine all the MLR assumptions in detail for the final regression model.

```
autoplot(model4, colour='steelblue4', smooth.colour = 'darkred')
```

The diagnostic plots for the final regression model (equation 8) are shown in Figure 17.

A. MLR Assumption 1: Nonlinearity in Parameters

We have not constrained the error term and the equation 8 is linear in parameters that were estimated using the OLS method. Thus this assumption is satisfied.

B. MLR Assumption 2: Random Sampling

As discussed earlier, we have 91 observations but the numbering of the counties goes up to 197, so there may

Table 15: Final Model Residuals Covariance with Explanatory Variables

Variable	Covariance
$\sqrt{\text{density}}$	0.0000000
$log(\text{probability of arrest})$	0.0000000
$log(\text{probability of conviction})$	0.0000000
$log(\text{percentage of minority})$	0.0137139
$(urban)$	0.0000000

be some counties missing. For our purpose, we have assumed the counties were selected randomly and are i.i.d., supporting our claim is that there is no indication for any systematic order to the selection of counties highlighted in the description for the data collection.

C. MLR Assumption 3: No Perfect Multicollinearity

We computed the correlation coefficients for all the variables in the data set and did not find perfect multicolinearity among any of them. Moreover, R will tell us if there is any perfect colinearity.

D. MLR Assumption 4: Zero Conditional Mean

Looking at the Residuals vs Fitted plot in Figure 5, we can see a parabolic shape for the residual means. This indicates that there may be a violation of the zero conditional mean, which can lead to our coefficient estimators being biased for predicting the changes in population crime rates. However, we argue that as $n \to \infty$, the bias of the estimator converge to 0, giving us $cov(x_i, u) = 0$, where x_i is each of our independent variables.

```
# Create a table for the covariances
variables <- c("$\\sqrt{\\text{density}}$",
               "$log(\\text{probability of arrest})$",
               "$log(\\text{probability of conviction})$",
               "$log(\\text{percentage of minority})$",
               "$\\text(urban)$")

covariances <- c(
  cov(sqrt(Crime_NA$density), model4$residuals),
  cov(log(Crime_NA$prbarr), model4$residuals),
  cov(log(Crime_NA$prbconv), model4$residuals),
  cov(log(Crime_NA$taxpc), model4$residuals),
  cov(Crime_NA$urban, model4$residuals))

df <- data.frame(variables, covariances)
names(df)[1] = "Variable"
names(df)[2] = "Covariance"

knitr::kable(df, format = "latex", booktabs=T, escape=F,
             caption="\\label{tab:model4_cors}Final Model Residuals
             Covariance with Explanatory Variables")
```

We can see in Table 15 that each of the co-variances are in fact 0.

E. MLR Assumption 5: Homoskedasticity

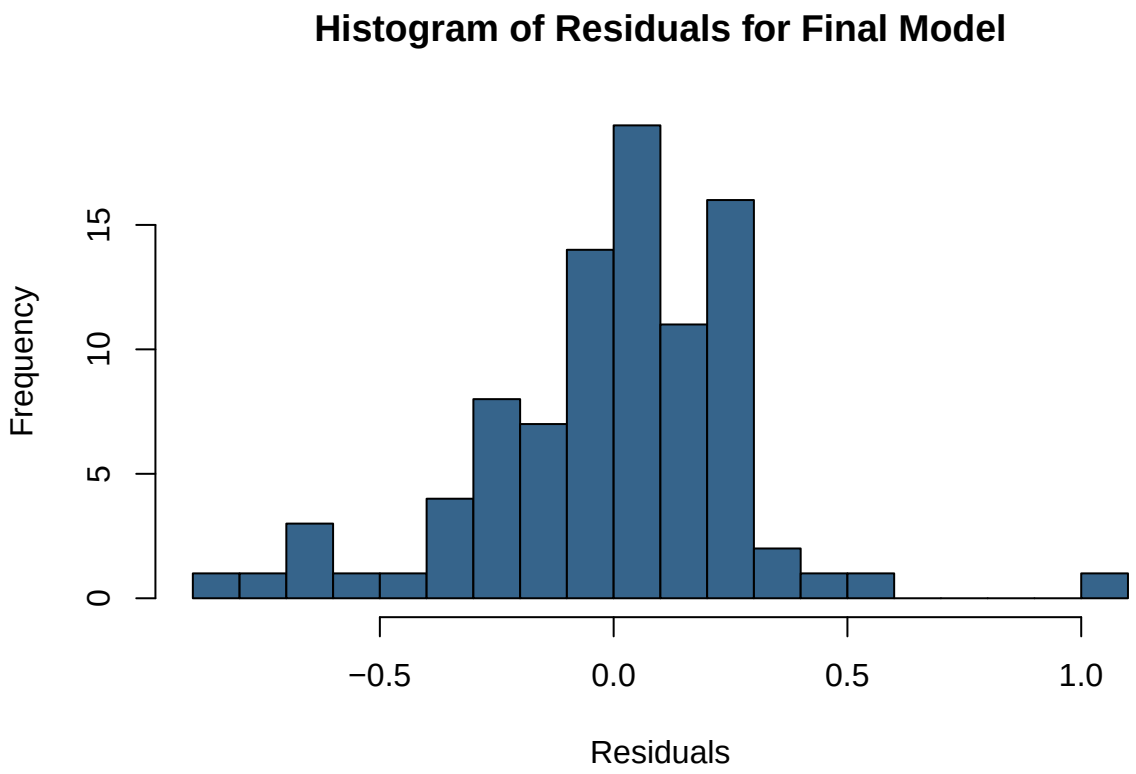

Figure 18: Histogram of Residuals for Final Model

We perform a Breusch-Pagan test to verify the homoskedasticity, using the NULL and alternate hypothesis H_0 : there exists homoskedasticity and H_A : there exists heteroskedasticity.

```
lmtest::bptest(model4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model4
## BP = 13.541, df = 5, p-value = 0.01881
```

Our p-value is less than our $\alpha(= 0.05)$, so we can reject the H_0 of there being no heteroskedasticity. However, looking at the diagnostic plots for the model we can see that the mean deviates from zero only for the extreme ends of the variable range. Also, we have used heteroskedasticity robust errors in all our analysis to account for violation of MLR.5.

F. MLR Assumption 6: Normality of Errors

The Q-Q plot in Figure 17 appears to have a fairly normal distribution of the errors, with some deviation at the lower values of errors. The histogram of the residuals in Figure 18 also shows a near-normal distribution.

```
hist(model4$residuals, breaks=25, main='Histogram of Residuals for Final Model',
     xlab='Residuals', col = 'steelblue4')
```

Conclusion and Policy Proposal

From our analysis, we can come to some conclusions as to what suggestions to provide to the political campaign. The practical significance of the OLS coefficients of variables will have precedence over their statistical significance in our interpretations. policy. Since our outcome variable is log(crime rate), our model predicts percentage changes to **crime rate** with changes to the explanatory variables.

1. A unit change in **population density** is predicted to increase the **crime rate** by approximately 34%. New zoning laws could be implemented to reduce population densities in areas of high crimes rates. This can be achieved by expanding habitable areas by providing more housing permits to builders that allows expansion of cities and reduces the population density.
2. Each percent increase in the **probability of arrest** leads to a 0.472% decrease in **crime rate** while a percent increase in **probability of conviction** decreases **crime rate** by 0.343%. Compared to the effects of population density, the impact of these variables has low practical significance. Stricter penalties for crimes and reduction of loopholes in the criminal code, can lead to reduction in crime rate but the effect may not be a good trade off for the government resources that need to be spend on this effort.
3. Similarly the coefficient estimate for **percentage minority** has high statistical significance but the magnitude of the value shows low practical significance. A unit reduction in minority percentage is estimated to reduce crime rate by 1.2%. We believe that the high statistical significance of the OLS coefficient for this variable is thus confounded by the effects of population density. Highly populous areas will tend to have a higher percentage of minorities and thus show a high correlation with **crime rate**. The effects of urban neighborhoods predicted by our model is important to understand.
4. Since **urban** is a categorical variable with levels: 0 and 1, its coefficient adds to the intercept of the regression line when all other variables are zero. Thus we predict that urban neighborhoods tend to have higher base crime rate without the effects of explanatory variables. Omitted variables like wealth in the neighborhoods can be confounding the effect of **urban** variable on **crime rate**.

Based on the above discussion, we suggest as a policy to modify zoning laws and expand densely populated cities and neighborhoods by approving more housing permits. Reducing the population density is predicted to bring down crime rate significantly in North Carolina counties.