Lab3 Final, w203: Statistics for Data Science

Avinash Chandrasekaran, Deepak Nagaraj, Saurav Datta

April 13, 2018

0. Introduction

Our team has been hired to provide research for a political campaign. The campaign has obtained a dataset of crime statistics for a selection of counties in North Carolina. Our task is to examine the data to help the campaign understand the determinants of crime and to generate policy suggestions that are applicable to local government.

The data provided consists of 25 variables and 97 different observations collected in a given year. Moreover the dataset obtained is a single cross-section of data collected from variety of different sources. For the analysis made in this research, we will assume that the data collected from different counties in NC were randomly sampled.

Our primary analysis of data will include ordinary least squares regressions to make causal estimates and we will clearly explain how omitted variables may affect our conclusions. We begin our research by conducting exploratory analysis of the dataset to gain a better understanding of the variables.

Agenda

1. Data input and cleanup
2. Exploratory Data Analysis
3. Correlation Analysis
4. Model development and analysis
5. Conclusion

---

1. Data input and cleanup

Our first step is to detect anomalies such as missing and duplicate values and to clean up the dataset before deeper dive into regression analysis.

```
# Read the csv file
crime_data_raw = read.csv("crime_v2.csv")
```

```
summary(crime_data_raw)
tail(crime_data_raw, n=8)
```

There appears to be 6 rows of NA's across all variables. We can simply use na.omit(), because the number of all-NA rows matches the count on all the variables.

We noticed that 'prbconv' is a factor while the rest of the variables are numeric.

County and Year variables just represent the different counties and the year the data was collected. Year is always 87. Hence, we can safely remove these from the dataset for further analysis.

We also noticed a duplicate record (record #89) in the dataset. As this could potentially affect our regression analysis, we will remove the duplicate record.

```r
# remove NA rows
crime_data = na.omit(crime_data_raw)
# convert factor to numeric for variable prbconv
crime_data$prbconv = as.numeric(levels(crime_data$prbconv)[crime_data$prbconv])
crime_data = crime_data %>% dplyr::select(-c(year, county))
# convert percentages into (0, 100) range
crime_data$pctmin80 = crime_data$pctmin80 * 100
crime_data$pctymle = crime_data$pctymle * 100
# convert probabilities into (0, 100) range
crime_data$prbarr = crime_data$prbarr * 100
crime_data$prbconv = crime_data$prbconv * 100
# remove duplicate record
duplicated(crime_data)[duplicated(crime_data)==TRUE]
```

```
## [1] TRUE
```

```r
crime_data = distinct(crime_data)
```

Influential outliers

We now present some influential outliers we found during our analysis that affect our regression fit. We measure influence via Cook's distance (>1).

Observation #84: extremely high service wages. This observation is so skewed that it pulls the regression line of fit. The county corresponds to Warren; there is nothing in Wikipedia to suggest that this county can command such high service wages.

We think this is a measurement error and we will remove it.

```r
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))

# Notice the outlier
boxplot(crime_data$wser, main="Service wages")

m <- lm(log(crmrte) ~ wser, data = crime_data)

# Notice the leverage of the data point
plot(m, which = 5, main = "Leverage", caption = NA)

# Note how the line of fit changes after removal
plot(crime_data$wser, log(crime_data$crmrte), main = "Before removal")
abline(m)

crime_data_tmp <- crime_data %>% dplyr::slice(-84)
plot(crime_data_tmp$wser, log(crime_data_tmp$crmrte), main = "After removal")
m <- lm(log(crmrte) ~ wser, data = crime_data_tmp)
abline(m)

# Remove
crime_data <- crime_data_tmp
```
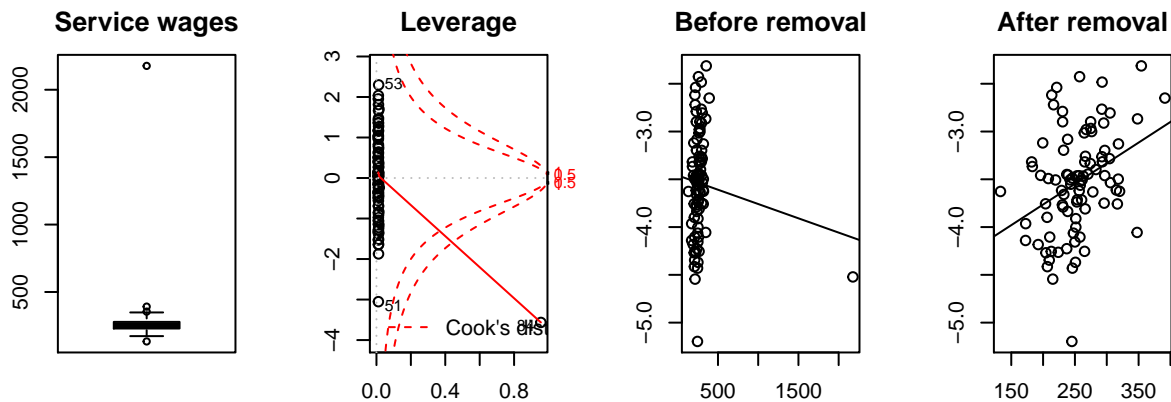
We saw some outliers with a distinct pattern: they come from counties that are heavy on tourism, but with low resident population. This causes high crime rates and police presence even in the presence of low density, because density probably counts only residents. We were enabled by Zach's report, where he states that the county numbers correspond to FIPS codes. We looked up county information on NCPedia. We remove these outliers to improve our model fit.

Observation #51: highly influential observation, which has very low crime rate, yet very high police per capita. This observation corresponds to Madison County in the Smoky mountains of western NC.

```r
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))

# Notice the outlier
boxplot(crime_data$polpc, main = "Police p.c.")
m <- lm(log(crmrte) ~ polpc, data = crime_data)

# Notice the leverage of the data point
plot(m, which = 5, main = "Leverage", caption = NA)

# Note how the line of fit changes after removal
plot(crime_data$polpc, log(crime_data$crmrte), main = "Before removal")
abline(m)

crime_data_tmp <- crime_data %>% dplyr::slice(-51)
plot(crime_data_tmp$polpc, log(crime_data_tmp$crmrte), main = "After removal")
m <- lm(log(crmrte) ~ polpc, data = crime_data_tmp)
abline(m)

# Remove
crime_data <- crime_data_tmp
```
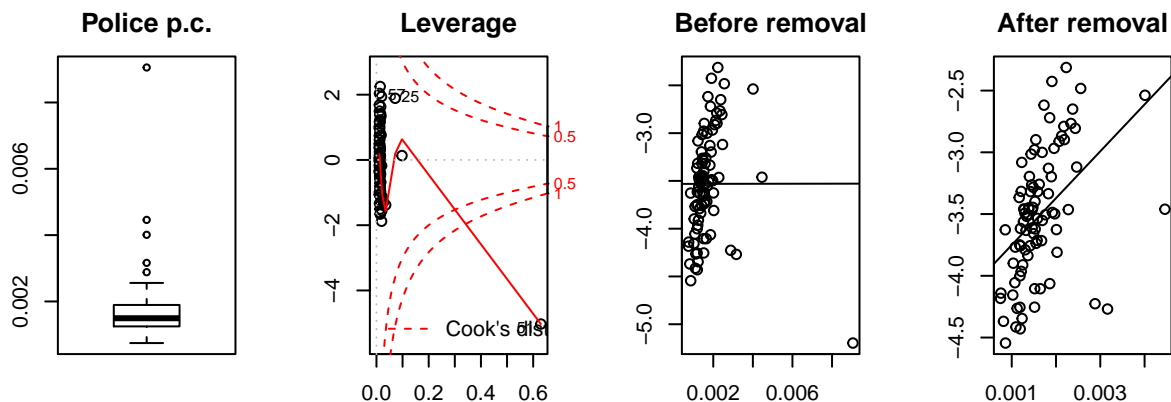
Observation #78: extremely low density. This corresponds to Swain County, western North Carolina. This county has almost all its land area in a national park and is an influential outlier. It also derives revenue from tourism.

```r
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))

# Notice the outlier
boxplot(log(crime_data$density), main = "Density")
m <- lm(log(crmrte) ~ log(density), data = crime_data)

# Notice the leverage
plot(m, which = 5, caption = NA, main = "Leverage")

# Print "before-after" graphs
plot(log(crime_data$density), log(crime_data$crmrte), main = "Before removal")
abline(m)

crime_data_tmp <- crime_data %>% dplyr::slice(-78)
plot(log(crime_data_tmp$density), log(crime_data_tmp$crmrte), main = "After removal")
m <- lm(log(crmrte) ~ log(density), data = crime_data_tmp)
abline(m)

# Remove
crime_data <- crime_data_tmp
# See below
crime_data <- crime_data %>% slice(-25)
```
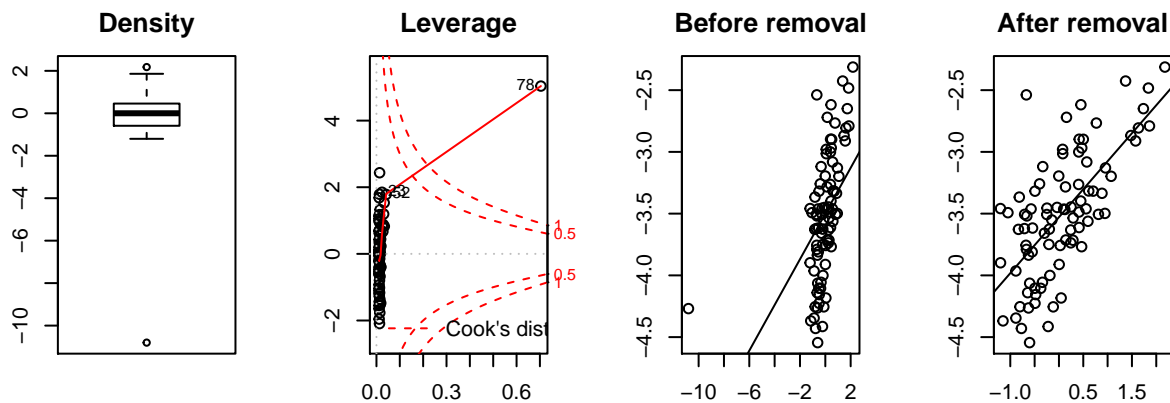

Observation #25 causes outliers in the final model fit due to a lot of influence. It shows very high crime rate and very high police per capita, while also showing very low density, very low minority and highest tax per capita.

The observation corresponds to Dare County on the eastern seaboard, consisting mostly of a sliver of an island. NCPedia lists seasonal tourism as the primary industry of the county.

---

2. Exploratory Data Analysis

We will now try to get a sense of each variable in the dataset. We define a utility function to describe a variable, visible in the original R markdown, but omitted here to leave more space for analysis.

Single variable analysis

Crime rate

Crime rate is the key dependent variable of interest.

```r
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
hist(crime_data$crmrte, main = "Crime rate",
     xlab="Crime Rate")
boxplot(crime_data$crmrte, main = "Boxplot")
hist(log(crime_data$crmrte), main = "Crime rate",
     xlab="Log of Crime Rate")
crime_data$log_crmrte = log(crime_data$crmrte)
```
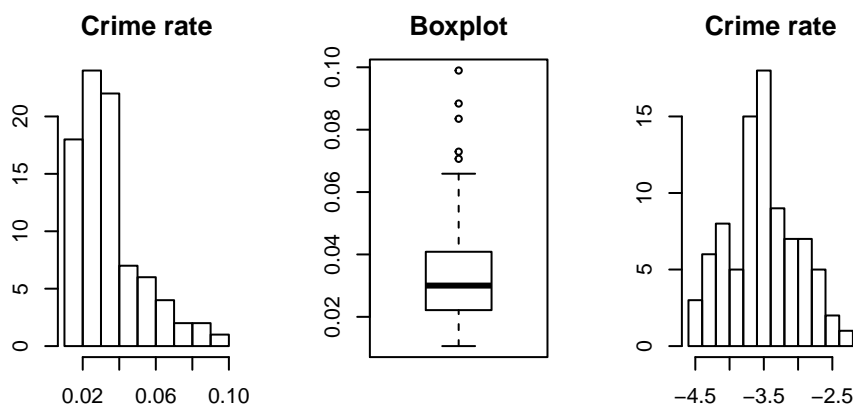

Looking at the histogram, the distribution is positively skewed to the left. We can take the log transformation which makes the variable appear more normally distributed.

Crime rate is mostly low, but there are some observations that show high crime rate (positive outliers). This causes skew.

Probability of arrest

```r
f_describe_col(crime_data$prbarr)
```

```
## [1] "Correlation:  -0.37"
```
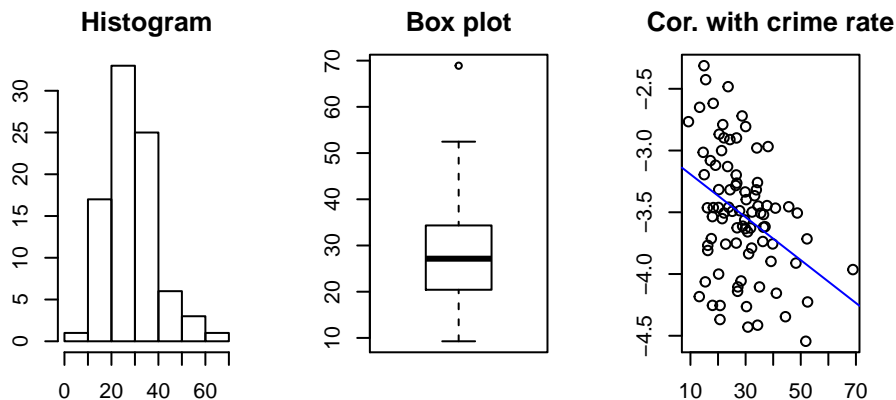
The plot looks fairly normal; there is only one outlier that corresponds to the high arrest probability of 0.7. It is not influential, so we will keep it.

There is fairly negative correlation of -0.37: as probability of arrests increases, crime rate goes down. It may be that arrests are a deterrent, indicating causality.
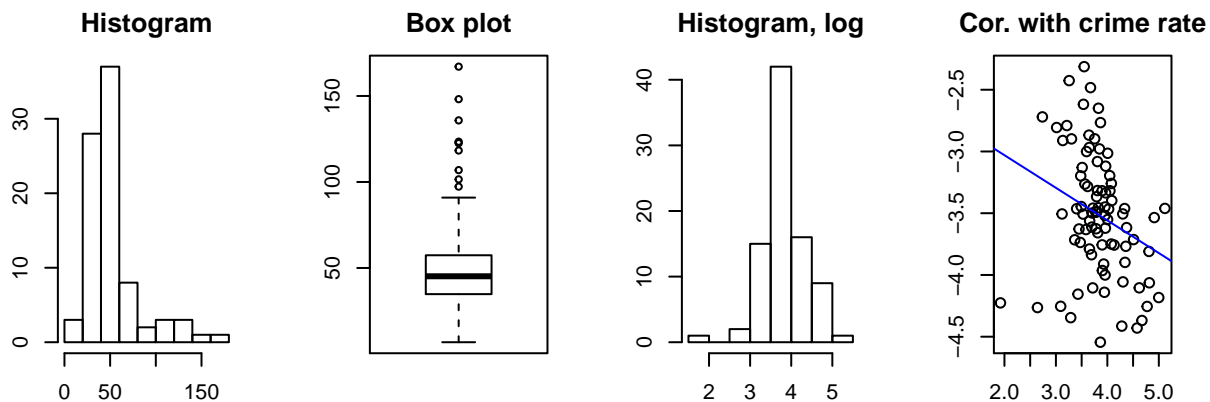
We will include *prbarr* as an independent variable in our model.

Probability of conviction

```
f_describe_col(crime_data$prbconv, do_log=TRUE)
```

```
## [1] "Correlation:  -0.275"
```

```
crime_data$log_prbconv = log(crime_data$prbconv)
```


This variable has quite a bit of left skew. It also has many outliers after the 3rd quartile. There are a few beyond 1 as well. Again, this is because we are not looking at a real probability but a ratio of convictions to arrests. It is possible, although perhaps uncommon, that a suspect is arrested once but convicted on multiple charges.

Taking a log transform improves the skew, although the spread is still quite a bit. There are no outliers with large influence as measured by Cook's distance (not shown).

There is moderate negative correlation with crime rate of -0.3. As convictions go up, crime rate goes down. Since we have already considered *prbarr*, let us check if *prbconv* has high correlation with *prbarr*:

```
print(cor(crime_data$prbarr, crime_data$prbconv))
```

```
## [1] -0.3058084
```

```
print(cor(crime_data$prbarr, crime_data$log_prbconv))
```

```
## [1] -0.299908
```

We don't see much correlation and therefore, will include *log_prbconv* in our model.
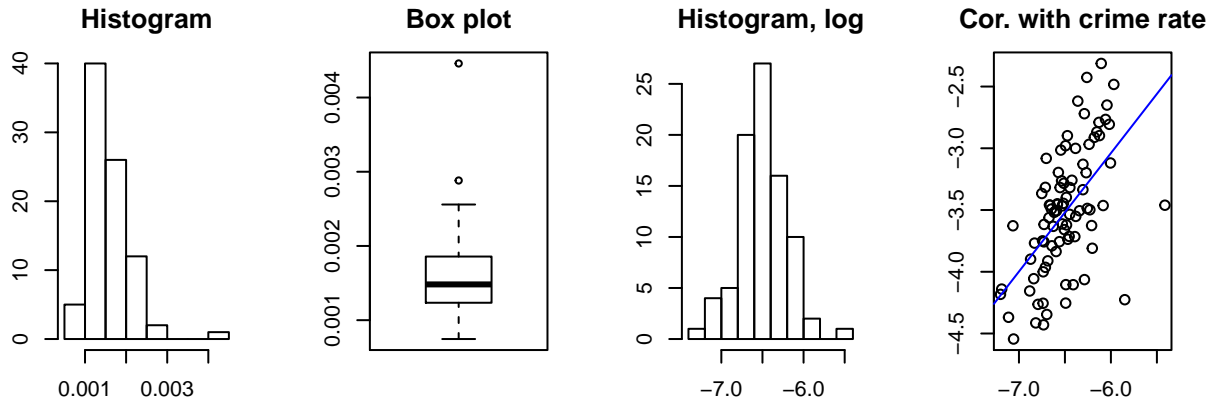

Police per capita

Police per capita has positive skew. We observed that taking the log tranformation made the distribution more normal.

```
f_describe_col(crime_data$polpc, do_log=TRUE)
```

[1] "Correlation: 0.579"

```
crime_data$log_polpc = log(crime_data$polpc)
```

Histogram	Box plot	Histogram, log	Cor. with crime rate

We see fairly strong positive correlation of 0.6 with crime rate: high number of police per capita is associated with high crime rate. But do police *cause* crime? It is probably an effect: more police may have been deployed to deal with higher amount of crime. If that is the case, it is worth questioning further why the additional police has not lowered the crime rate: are they ineffective?
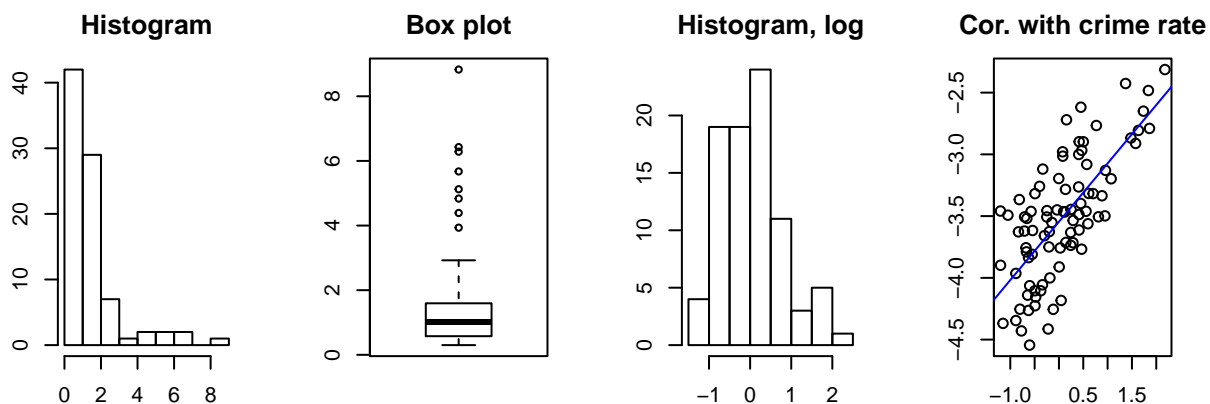
For our first model, we will *not* include this variable, but we will include it in a second model.

Population density

```
f_describe_col(crime_data$density, do_log=TRUE)
```

[1] "Correlation: 0.715"

```
crime_data$log_density = log(crime_data$density)
```

Histogram	Box plot	Histogram, log	Cor. with crime rate

The histogram of density shows quite a bit positive skew: they probably correspond to cities and the countryside. The log transformation shows a more promising normal distribution. There are no outliers with large leverage as measured by Cook's distance.

We see remarkably high positive correlation with crime rate. It may be that high population density indicates greater scope for hiding or cooperation in order to commit crime, indicating causality. We will surely consider this variable in our model.

Tax revenue per capita

```
f_describe_col(crime_data$taxpc, do_log=TRUE)
```

```
## [1] "Correlation:  0.29"
```

```
crime_data$log_taxpc = log(crime_data$taxpc)
```
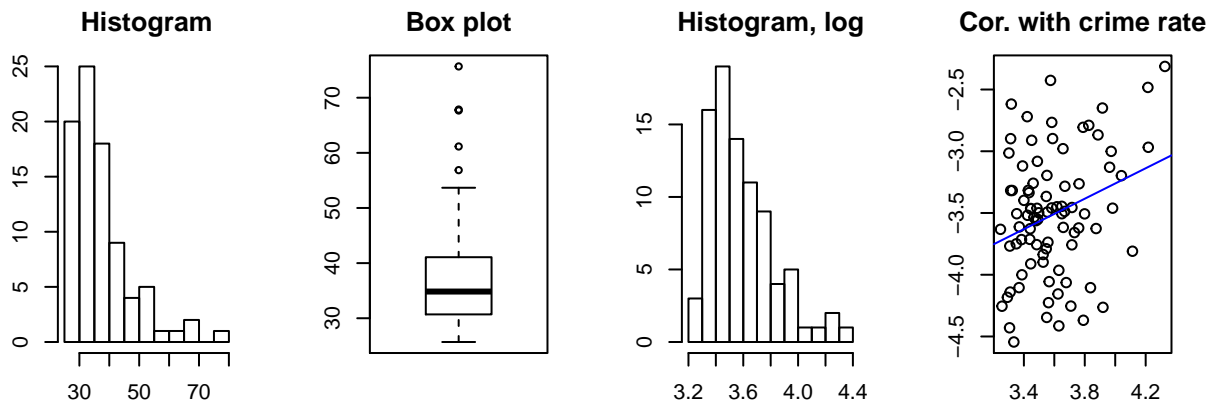

Tax revenue also shows positive skew, probably in line with density: urban areas may generate more taxes than the countryside.

We also see considerable positive correlation with crime rate. It may be that tax revenue is a proxy for wealth, and high amount of wealth attracts crime. On the other hand, it is worth checking if we are spending tax dollars wisely in combating crime: if that were the case, counties with higher tax revenue would probably see lower crime.

We will *not* include this variable in a first model.

Percent minority

```
f_describe_col(crime_data$pctmin80)
```

```
## [1] "Correlation:  0.297"
```
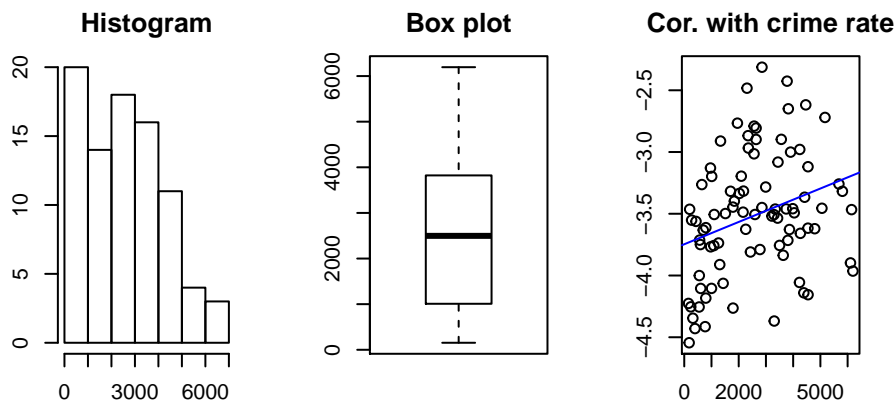
Minority percentage has a bit positive skew, but no outliers. The range is quite limited. We will not try to transform this variable.

There is a fair amount of positive correlation with crime rate (0.27). It may be that as minorities increase, there is loss of social homogeneity and/or hate crime.
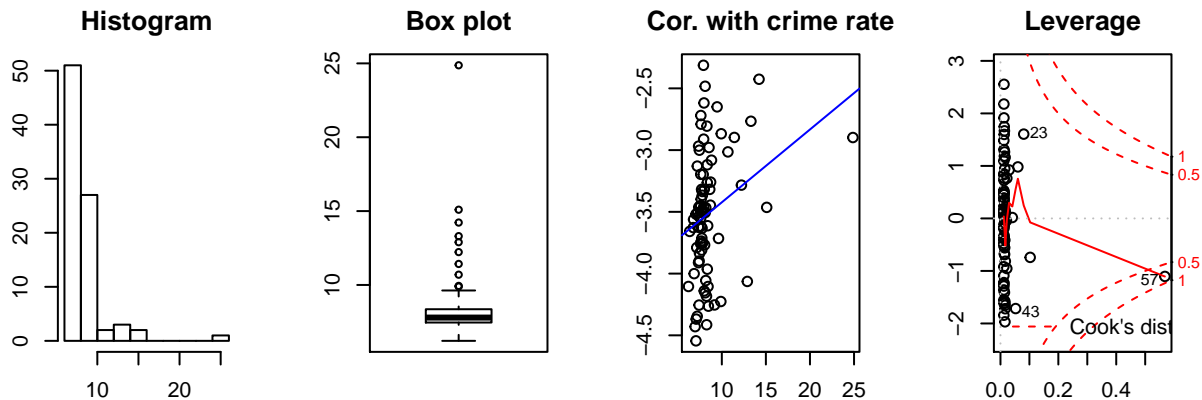
We will include this variable in our model to check for significance.

Percent of young males

```
f_describe_col(crime_data$pctymle, plot_model=TRUE)
```

```
## [1] "Correlation:  0.28"
```


We see moderate positive correlation with higher percentage of young males. Boxplot shows outliers, but none has outsized influence (Cook's distance > 1).

A high percentage of young males can indicate higher aggressiveness and risk, causing higher rate of crime. We may also see the effect of omitted variables like youth unemployment or low education levels.

We will include this variable in our model.

Categorical variables: West/Central, Urban

Let us check if crime rate has patterns in the categorical variables. We see two categories:

- Urban or rural
- Geographic: west, central, other

```
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
crime_data_urban <- crime_data %>% filter(urban == 1)
crime_data_rural <- crime_data %>% filter(urban == 0)
boxplot(crime_data_urban$crmrte, ylim=c(0.0, 0.1), main = "Urban")
boxplot(crime_data_rural$crmrte, ylim=c(0.0, 0.1), main = "Rural")
```
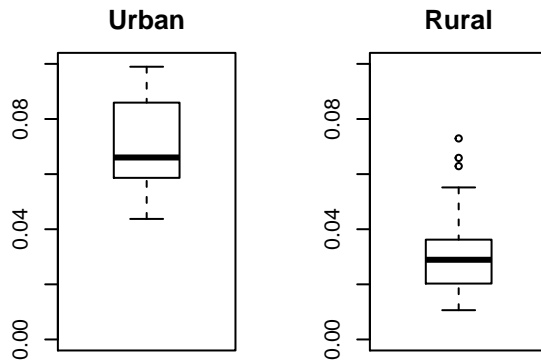
Urban   **Rural**

We see that crime rate is much higher in urban areas than rural areas, confirming our previous observation of positive correlation with density as well. Median crime rate is double the value in rural areas.

We have only 8 counties termed urban, so it is a small sample. Therefore, and since we have already selected density, we will *not* include this variable in our model.

We also compared crime rate across urban vs. rural, *ceteris paribus*.

```r
m <- lm(log(crmrte) ~ factor(urban), data = crime_data)
summary(m)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ factor(urban), data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94517 -0.23010  0.04276  0.26244  0.98154
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.59955    0.04849 -74.235  < 2e-16 ***
## factor(urban)1  0.91643    0.15898   5.764 1.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4282 on 84 degrees of freedom
## Multiple R-squared:  0.2835, Adjusted R-squared:  0.2749
## F-statistic: 33.23 on 1 and 84 DF,  p-value: 1.32e-07
```

We see that being an urban county is highly statistically significant for crime rate. Also, there is a difference of 0.92 (or 92%) between crime rate in urban and rural areas. This is a lot and shows that crime is strongly correlated with urban areas.

We then performed an analysis on the geographic distribution. From this , we see that crime is higher in central region. This is because the big cities of North Carolina are close to its center geographically. The west is mostly mountains with very few people. There are 2 cities in the "Other" region, and they show as outliers with high crime rate in the corresponding boxplot.

```r
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
crime_data_west <- crime_data %>% filter(west == 1)
crime_data_central <- crime_data %>% filter(central == 1)
crime_data_other <- crime_data %>% filter(west == 0 & central == 0)
boxplot(crime_data_west$crmrte, ylim=c(0.0, 0.1), main = "West")
boxplot(crime_data_central$crmrte, ylim=c(0.0, 0.1), main = "Central")
```

10

```r
boxplot(crime_data_other$crmrte, ylim=c(0.0, 0.1), main = "Other")
```
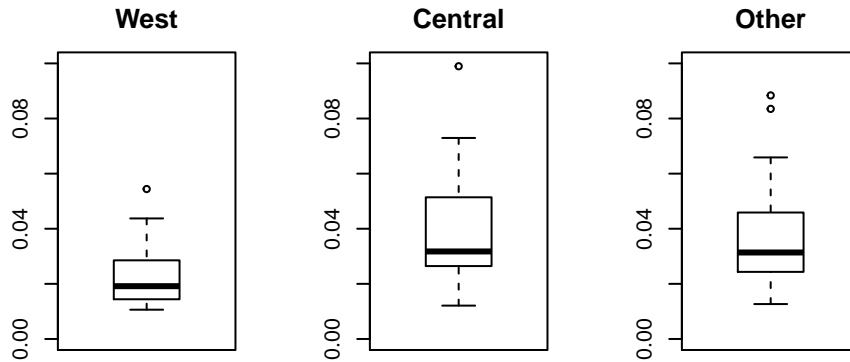

From our analysis of the geographic distribution, we observed the following:

- 27 observations have been marked as central with no overlap on west or urban

- 18 observations have been marked as west with no overlap on central or urban

- 2 observations have been marked as urban with no overlap on central or west

- No observations is marked as urban and central and west

- 1 observation has been marked as both central and west.

- 1 observation has been marked as both west and urban

- 5 observations have been marked as both central and urban

- 32 observations have not been marked as either west nor central nor urban

We interpret this to mean that an attempt was made to mark the observations geographically as well as by urban (or rural). However, we feel this assignment isn't complete as some regions haven't been given any label.

In general the west region has more occurences of lower crime rate than the central region. There is a significant overlap between the observations marked as west/central and those marked as urban. Also, there are no regions marked as rural.

This analysis provides another way of looking at crime as a function of urban, dense areas. It also suggests a different regression model based on the regions, which we attempt later.

Wage distribution

```r
par(mfrow=c(3,4), mai=c(0.35,0.35,0.35,0.35))
hist(crime_data$wcon, main="wcon")
hist(crime_data$wloc, main="wloc")
hist(crime_data$wtrd, main="wtrd")
hist(crime_data$wtuc, main="wtuc")
hist(crime_data$wfir, main="wfir")
hist(crime_data$wser, main="wser")
hist(crime_data$wmfg, main="wmfg")
hist(crime_data$wfed, main="wfed")
hist(crime_data$wsta, main="wsta")
```
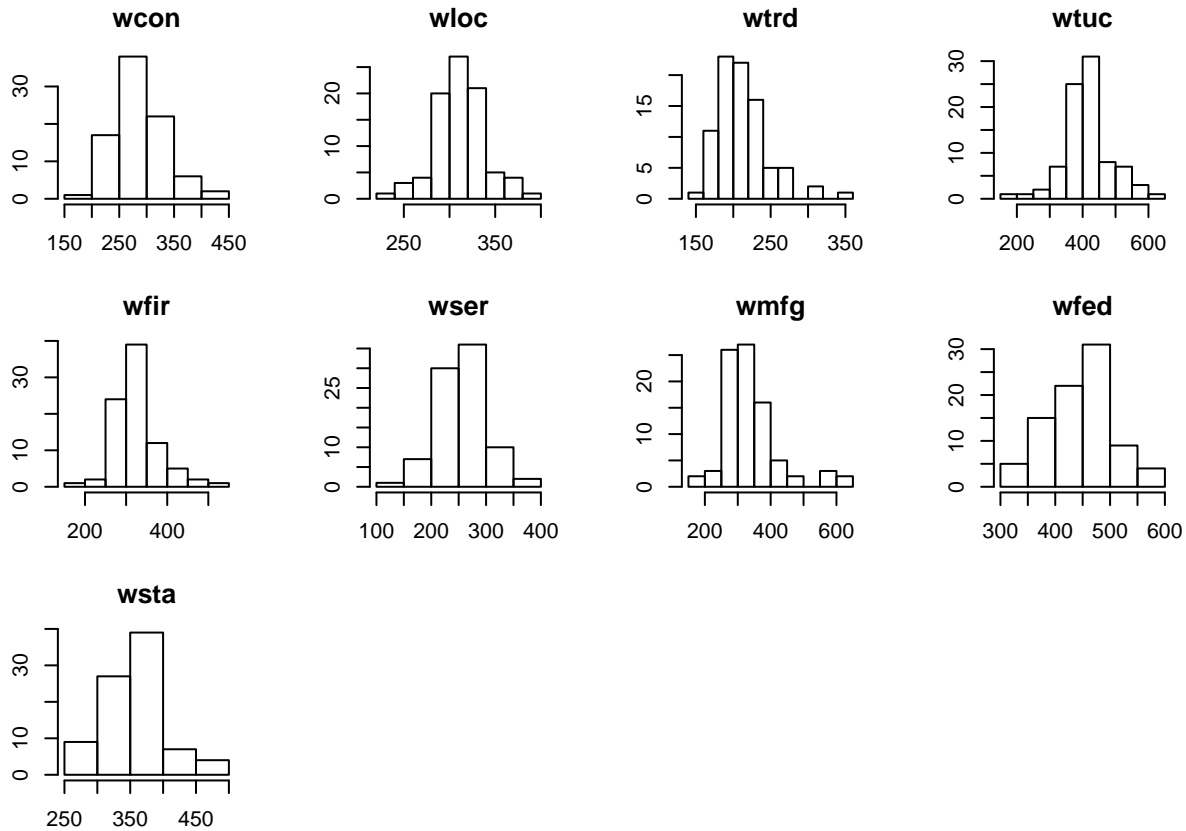
Most of the wage variables conform to normal distributions. We do not have to worry about transformations.

Let us look at which of them have high correlation with crime rate, considering all those with $R > 0.25$ (arbitrarily). If they have a high degree of correlation, we will pick one of them to represent all.

```
wage_cols = c("log_crmrte", "wcon", "wloc", "wtrd", "wtuc", "wfir",
              "wser", "wmfg", "wfed", "wsta")
corrplot(cor(crime_data[, wage_cols]), type="upper", diag=TRUE, addCoef.col="white",
         addCoefasPercent = TRUE, order="hclust", method="ellipse")
```
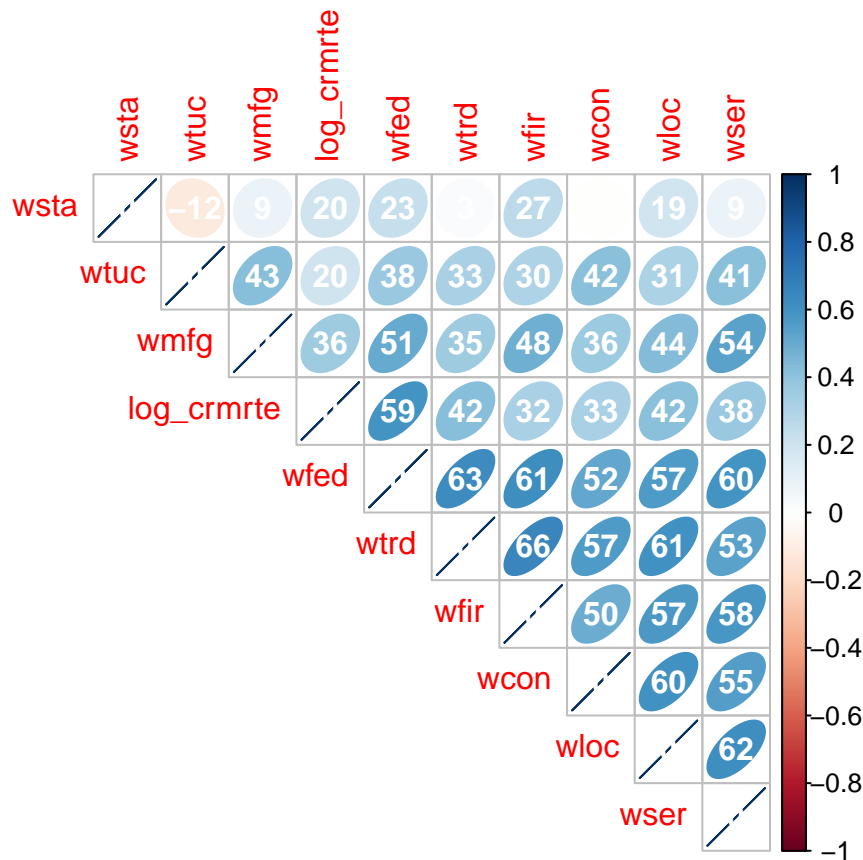
Indeed, a lot of the wage categories above have a high degree of correlation among them, although all are less than 70.

Higher wages are correlated positively with crime rate, but may indicate higher wealth or a different omitted variable. They cannot be causal in and of themselves.

We will *not* include wages in a first model. We will use $wfed$ as a representative variable for all wages, for a second model we will propose.

Other variables

The following variables show low to nonexistent correlation with crime rate. We will *not* consider them in our regression model.

- Probability of prison sentence $(R = 0.0537)$
- Offense mix $(R = 0.0136)$
- Average sentence duration $(R = 0.0438)$

3. Correlation Analysis

Let us check for correlations across predictor pairs.

The correlation plot between the different predictors is as follows:

```
corrplot(
  cor(crime_data[,
                 c("prbarr", "log_prbconv", "prbpris", "avgsen",
                   "log_polpc", "log_density", "log_taxpc", "pctmin80", "mix",
                   "pctymle", "wfir", "wfed")
                 ]),
        type = "upper",
  diag=TRUE, addCoef.col="white", addCoefasPercent = TRUE,
  order="hclust", method="ellipse")
```
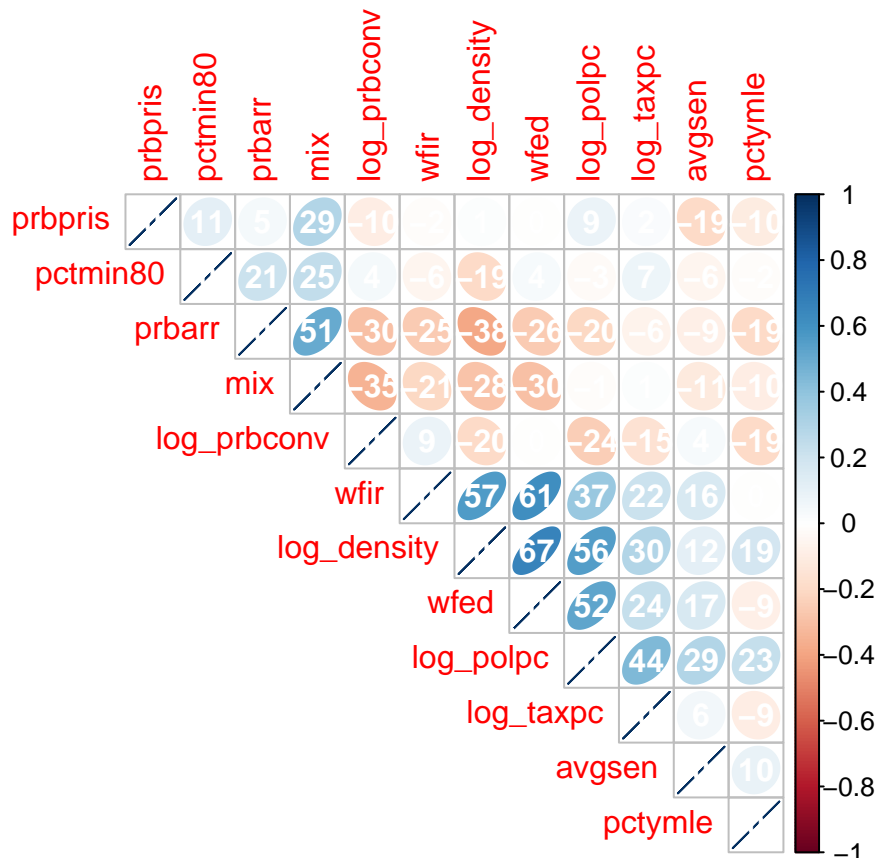

The following correlations are worth a remark:

- We do not see high correlations (>0.7), both positive or negative. This helps us avoid multicollinearity issues.

- Wages correlate highly (0.67) with population density, probably because it is tuned to cost of living.

- Police per capita correlates highly with density (0.56) and tax revenue (0.44), perhaps indicating federal laws on police counts.

- Probability of arrest correlates negatively (-0.38) with density, indicating that criminals get away with crime in populous areas. It also correlates negatively (-0.30) with probability of conviction, indicating higher chance of false arrests when a large number of arrests are made.

4. Model development and analysis

Summary of variables

We will now fit three models based on our investigation. A quick overview:

- Model 1 will have variables we think are causal in nature
- Model 2 will have a few additional variables that show high correlation
- Model 3 will have almost all variables: a useful benchmark

Here is a summary table of variables we will use in our models.

Variable	Transform?	Model1?	Model2?	Model3?	Remarks
county	N/A				Unused
year	N/A				Unused
prbarr		Y	Y	Y	Causal
prbconv	log	Y	Y	Y	Causal
prbpris				Y	No corr. found
avgsen				Y	No corr. found
polpc	log		Y	Y	Effect, not cause
density	log	Y	Y	Y	Causal
taxpc	log		Y	Y	Omit var: wealth
west	N/A				Categ, sep. model
central	N/A				Categ, sep. model
urban					Too few observations
pctmin80		Y	Y	Y	Causal
wcon					Ignored
wtuc					-"-
wtrd					
wfir				Y	
wser					
wmfg					
wfed			Y	Y	Representative for wages
wsta					Ignored
wloc					-"-
mix	log			Y	No corr. found
pctymle		Y	Y	Y	Causal, weak cor

Model 1

For the first model, here are the variables we will consider:

- We believe that the following can directly cause higher crime: high density, higher percentage of minorities, higher percentage of young men.

- We also think the following cause lower crime: high probability of arrest and conviction.

$$\log\left(crmrte\right) = \beta_0 + \beta_1 \; prbarr + \beta_2 \; \log\left(prbconv\right) + \beta_3 \; \log\left(density\right) + \beta_4 \; pctmin80 + \beta_6 \; pctymle$$

Let us fit the model:

```
model1 = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
            pctmin80 + pctymle, data=crime_data)
summary(model1)$adj.r.squared
```

```
## [1] 0.7836206
```

The model shows a fit of about 0.784 as measured by adjusted R^2.

Statistical significance:

coeftest() shows that all our variables are statistically significant. We need this over regular *summary()* because our model is not homoskedastic (explained later).

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

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -2.7440e+00  3.8150e-01 -7.1928 3.017e-10 ***
## prbarr      -1.3844e-02  3.1794e-03 -4.3543 3.918e-05 ***
## log_prbconv -2.3649e-01  7.5035e-02 -3.1517  0.002285 **
## log_density  4.2012e-01  4.6307e-02  9.0723 6.332e-14 ***
## pctmin80     1.5017e-04  1.5296e-05  9.8177 2.192e-15 ***
## pctymle      1.3753e-02  6.8297e-03  2.0136  0.047409 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore the model becomes (using 2 significant digits for coefficients):

$$\log\left(crmrte\right) = -2.7 - 0.014 \cdot prbarr - 0.24 \cdot \log\left(prbconv\right) + 0.42 \cdot \log\left(density\right) + 0.00015 \cdot pctmin80 + 0.014 \cdot pctymle$$

Practical significance:

From the above regression equation, we can infer the following. All statements have *ceteris paribus* assumptions:

- If we hold all other variables constant, we have a "base crime rate" of $e^{-2.7} = 0.067$.

- For a unit increase in "probability" of arrest, crime rate goes down by approximately $|e^{-0.014} - 1| = 0.014$, or 1.4%. This is a large effect. It shows that law enforcement is a very good deterrent to crime.

- For a 1% increase in "probability" of conviction, crime rate decreases by about 0.24%. This is not a lot. It may be because most of the effect is already absorbed by *prbarr*.

- For a 1% increase in density, crime rate increases by about 0.42%. This is a large effect. As we put more people closer, crime is easier to commit and harder to spot.

- For a 1% increase in minorities, crime rate increases by about $|e^{0.00015} - 1| = 0.015\%$. This is a minuscule effect. Compared to the animosity the majority may have over minorities, the effect is very little.

- For a 1% increase in young males, crime rate increases by about $e^{0.014} - 1 = 0.015$, or 1.5%. This is a large effect as well.

Verification of CLM Assumptions

Let us verify if the classical linear model assumptions hold on our model.

```
par(mfrow=c(2,2), mai=c(0.35,0.35,0.35,0.35))
plot(model1, which=c(1,2,3))
hist(model1$residuals, main="Residuals")
```
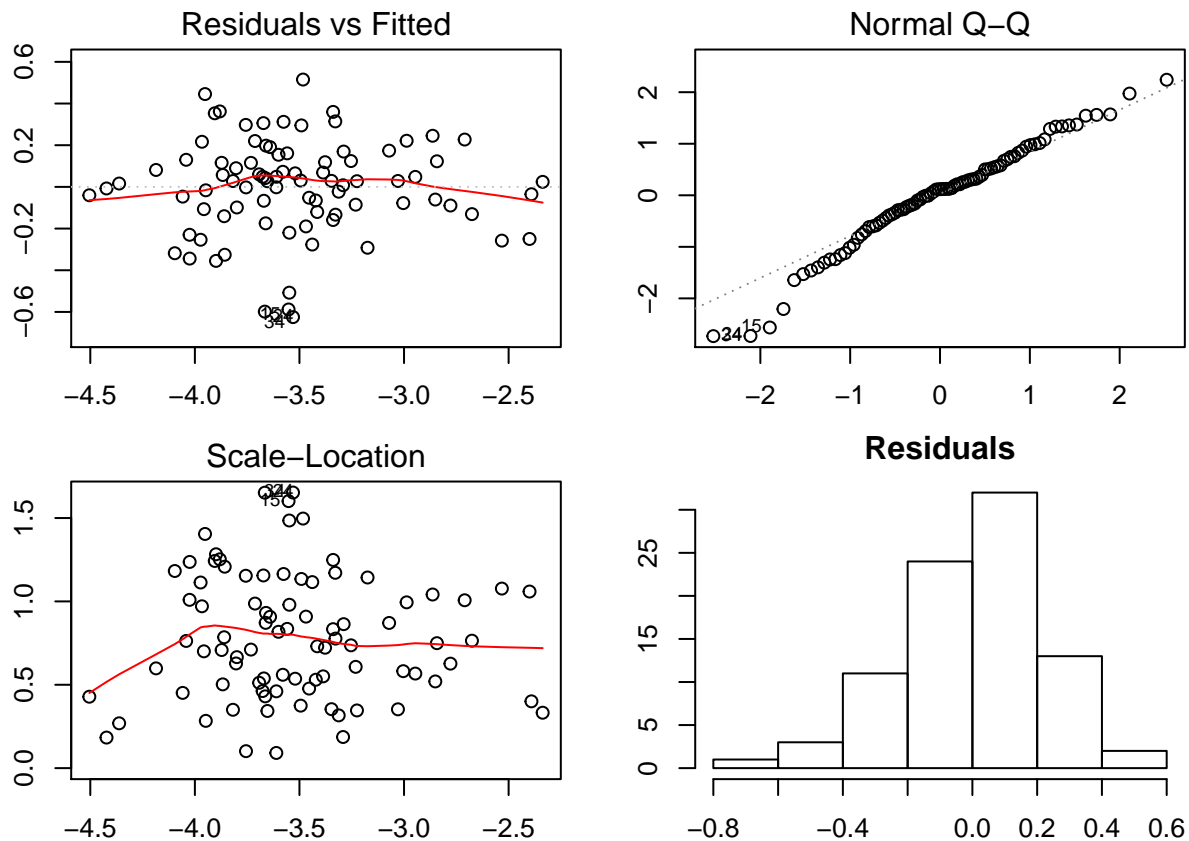
CLM.1: Linear in parameters

Any population distribution could be represented as a linear model plus some error (error might be poorly behaved). We chose the above model such that the dependent variable is a linear function of the explanatory variables. Therefore the CLM.1 assumption is met for our all of our models.

CLM.2: Random Sampling

For this assumption, the data needs to be a random sample drawn from the population. In general, we need to know about the experiment setup and procedure.

We first note that North Carolina is divided into 100 counties. Our dataset contained information from only 90 different counties. Though we didn't use data from all counties, there was no indication of non-random sampling during our analysis. We therefore assume that the counties sampled are random and presume our criterion is met for all the models we created.

CLM.3: No Perfect Multicollinearity

From our EDA it was apparent that none of our variables had constant values (since we had already removed the *year* variable). In addition, inspection of the correlation plot we generated indicates there are no perfectly correlated variable pairs.

Next, we check for high degree of collinearity.

```
vif(model1)
```

```
##      prbarr log_prbconv log_density   pctmin80    pctymle
##    1.542042    1.348490    1.351237   1.074870   1.125765
```

17

The variance inflation factor does not provide any evidence of multicollinearity as well. We therefore assume CLM.3 condition is satisfied.

CLM.4: Zero Conditional Mean

By examining the residuals-vs-fitted plot for model, we conclude that the assumption of zero conditional mean is mostly met. The red spline curve is close to zero for the most part, but it does deviate towards the edges.

Although we violate ZCM, we note that our sample size is large (>35). Hence we will apply the weaker exogeneity condition, by checking if our independent variables have significant covariance with the residuals:

```r
cov(model1$residuals, crime_data$prbarr)
```

```
## [1] -1.460867e-16
```

```r
cov(model1$residuals, crime_data$log_prbconv)
```

```
## [1] 1.141665e-17
```

```r
cov(model1$residuals, crime_data$log_density)
```

```
## [1] -6.871928e-20
```

```r
cov(model1$residuals, crime_data$pctmin80)
```

```
## [1] -7.430984e-15
```

```r
cov(model1$residuals, crime_data$pctymle)
```

```
## [1] -1.825238e-17
```

All of them are close to zero. Hence we can argue that our estimators are consistent, therefore our bias is close to zero.

There are two other corrections possible. We may have omitted variable(s). We certainly have many such variables that we list out later in this report. We could also correct our regression model specification, possibly adding higher-order functions or interaction terms.

CLM.5: Homoskedasticity

When we examine the residuals-vs-fitted plot, it is apparent that the variance of errors in the middle of the plot is higher than the variance of errors in the edges of the plot. This suggests heteroskedasticity. Additionally, the spline curve on the scale-location plot is curved rather than flat, indicating heteroskedasticity.

```r
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 8.6976, df = 5, p-value = 0.1218
```

We also conducted the studentized Breusch-Pagan test where the null hypothesis states that the model supports homoskedasticity. The p-value obtained indicates statistical significance that rejects the null hypothesis (=our model violates homoskedasticity).

To address this violation of CLM, we will use heteroskedasticity-robust standard errors in our analyses.

CLM.6: Normality of error terms

Looking at the histogram of the residual term indicates that we do *not* have a normal distribution. This can further be confimed by the Shapiro test. The null hypothesis of this test is that residuals have a normal distribution. The p-value indicates we reject this null hypothesis.

```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.97586, p-value = 0.1072
```

However, because we have a large sample, we again make asymptotic argument. The CLT states that our estimators will have a normal sampling distribution. Since the skew is not too much, we can still use the model.

Other possibilities include transformations of dependent and/or independent variables, when they exhibit skewed distribution.

Model 2

Next, let us include some more variables, as model2.

In this model we also include three variables that show positive correlation with crime rate, albeit not causal. We think they are all the result of wealthy, urban demographics: high police per capita, high tax revenue and high federal wages. It is an omitted variable.

We do not include outcome variables that absorb causal effect (by having negative correlation).

```
model2 = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
            pctmin80 + pctymle +
          + log_polpc + log_taxpc + wfed,
          data=crime_data)
summary(model2)$adj.r.squared
```

```
## [1] 0.8008852
```

The fit improves to 0.80 (adjusted R^2). This is because some of the newly added variables are more normal in distribution (=less skew).

Statistical significance:

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

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -1.5207e+00  1.8534e+00 -0.8205 0.4144807
## prbarr      -1.3462e-02  3.3149e-03 -4.0610 0.0001165 ***
## log_prbconv -2.3172e-01  9.0871e-02 -2.5499 0.0127605 *
## log_density  3.2115e-01  5.9437e-02  5.4031 7.068e-07 ***
## pctmin80     1.4124e-04  1.7726e-05  7.9681 1.172e-11 ***
## pctymle      1.4927e-02  1.8740e-02  0.7965 0.4281722
## log_polpc    2.2899e-01  2.0451e-01  1.1197 0.2663103
## log_taxpc   -5.3036e-02  1.4841e-01 -0.3574 0.7217935
## wfed         9.9763e-04  8.0714e-04  1.2360 0.2202151
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now see that *pctymle* is not significant. In addition, $log_t axpc$, *polpc* and *wfed* are also not statistically significant. It seems as if none of the variables we added matter, statistically speaking.

Let us take a look at its plots and check if we violate any of the CLM assumptions:

```
par(mfrow=c(2,2), mai=c(0.35,0.35,0.35,0.35))
plot(model2, which=c(1,2,3))
hist(model2$residuals, main="Residuals")
```
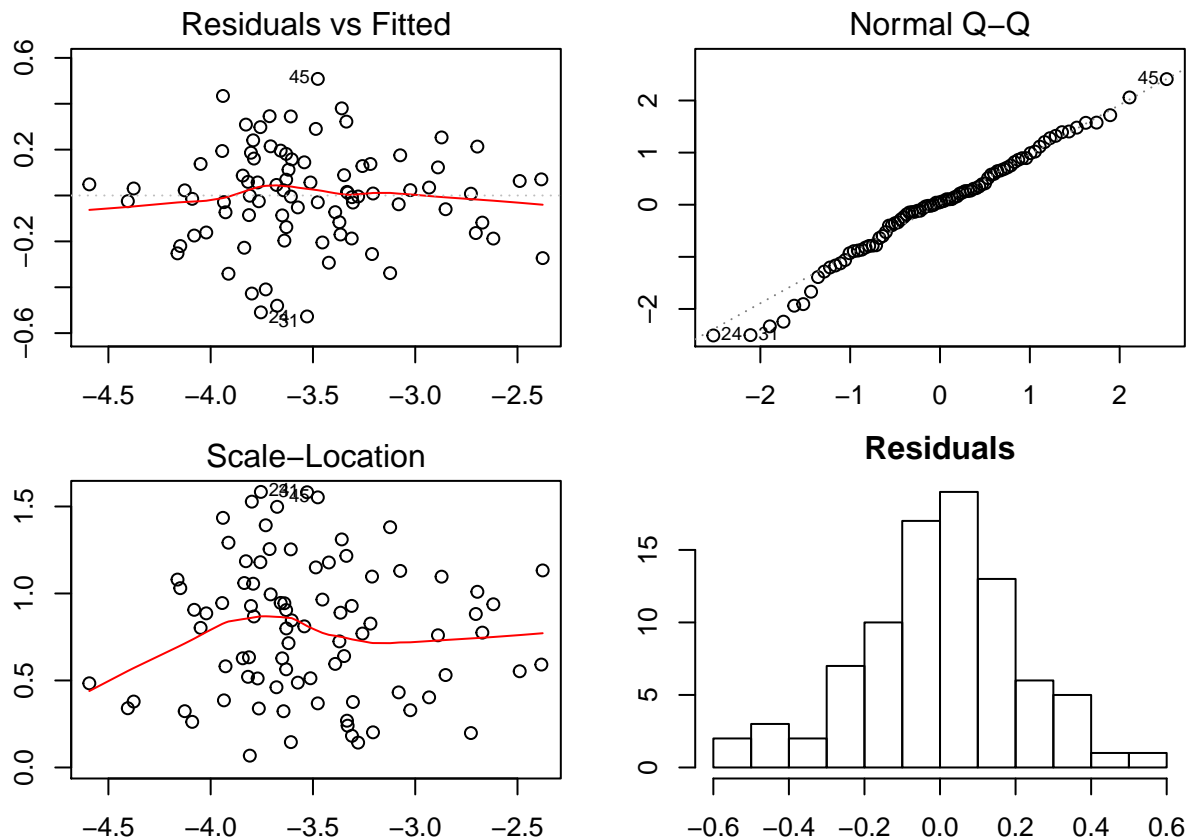

The model plots look very similar to the ones in Model 1.

Model 3

We will now build a third model that includes almost all the variables, for the sake of completeness and comparison. We only exclude wages because their distribution is highly alike.

```
model3 = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
                pctmin80 + pctymle
            + log_polpc + log_taxpc
            + prbpris + avgsen + wfir + wfed + mix,
            data=crime_data)
summary(model3)$adj.r.squared
```

```
## [1] 0.8244144
```

This tops fit at about 0.824 (adjusted R^2). However, it is worth noting that the benefit is not much: all those extra variables improved the fit by about 4%.

```r
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -8.3484e-01  1.6498e+00 -0.5060   0.61435
## prbarr      -1.3845e-02  3.1111e-03 -4.4501 3.021e-05 ***
## log_prbconv -1.8107e-01  9.2515e-02 -1.9572   0.05415 .
## log_density  3.5488e-01  6.7576e-02  5.2516 1.430e-06 ***
## pctmin80     1.3618e-04  1.6775e-05  8.1181 8.447e-12 ***
## pctymle      1.6627e-02  1.4015e-02  1.1863   0.23933
## log_polpc    2.8971e-01  1.8112e-01  1.5995   0.11402
## log_taxpc   -4.8620e-02  1.4773e-01 -0.3291   0.74302
## prbpris     -3.3788e-01  3.4205e-01 -0.9878   0.32650
## avgsen      -2.2069e-02  1.0841e-02 -2.0356   0.04542 *
## wfir        -1.5789e-03  8.1055e-04 -1.9479   0.05527 .
## wfed         1.7220e-03  8.0120e-04  2.1493   0.03492 *
## mix          3.5130e-01  4.6870e-01  0.7495   0.45596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In addition to what we saw in model2, only *avgsen*, *wfir* and *wfed* are seen to be practically significant. We can ignore *prbpris*, *mix*, *taxpc*, and *pctymle*.

```r
par(mfrow=c(2,2), mai=c(0.35,0.35,0.35,0.35))
plot(model3, which=c(1,2,3))
hist(model3$residuals, main="Residuals")
```
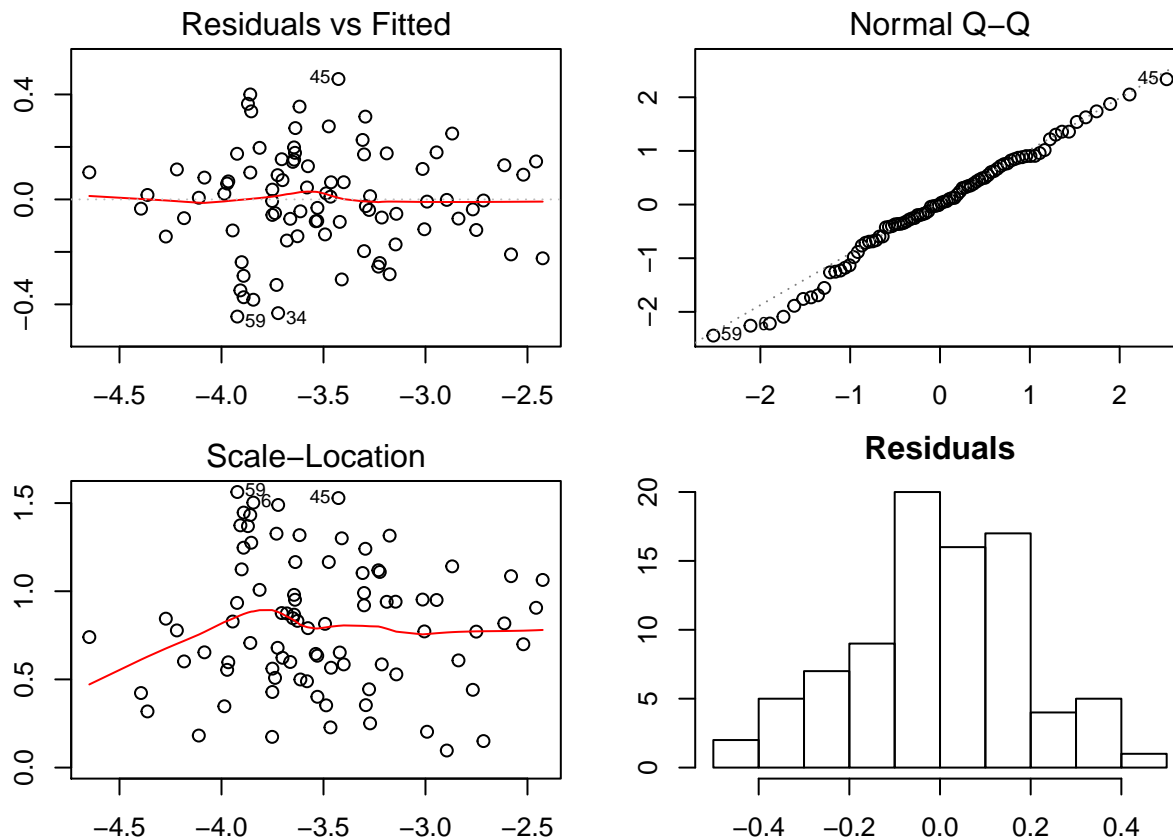
This model comes closest to satisfying all 6 CLM assumptions. We see a nice spline on 0.

Variation to Model 3

We will now build a variation of model3 that includes all the wage variables in addition to the model3 variables. We want to see if the wages are jointly significant.

```
model3b = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
             pctmin80 + pctymle
           + log_polpc + log_taxpc
           + prbpris + avgsen
           + wcon + wloc + wtrd + wtuc + wfir + wser + wmfg + wfed + wsta
           + mix,
           data=crime_data)
summary(model3b)$adj.r.squared
```

```
## [1] 0.8185732
```

```
#par(mfrow=c(2,3), mai=c(0.35,0.35,0.35,0.35))
#plot(model3b, which=c(1,2,5))
coeftest(model3b, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -5.3340e-01  1.9385e+00 -0.2752   0.78405
## prbarr      -1.4337e-02  3.1072e-03 -4.6142 1.871e-05 ***
## log_prbconv -1.8377e-01  1.0687e-01 -1.7196   0.09019 .
## log_density  3.4374e-01  7.3305e-02  4.6892 1.422e-05 ***
## pctmin80     1.3434e-04  1.7413e-05  7.7152 8.562e-11 ***
## pctymle      2.2401e-02  9.5690e-03  2.3410   0.02226 *
## log_polpc    3.2635e-01  2.2317e-01  1.4623   0.14839
## log_taxpc   -3.2255e-02  1.7298e-01 -0.1865   0.85265
## prbpris     -4.0212e-01  3.6261e-01 -1.1090   0.27147
## avgsen      -2.3623e-02  1.2256e-02 -1.9274   0.05824 .
## wcon         1.2810e-04  7.0972e-04  0.1805   0.85732
## wloc         3.8380e-04  1.7515e-03  0.2191   0.82723
## wtrd         1.0253e-03  1.2273e-03  0.8354   0.40650
## wtuc         1.8438e-04  4.9261e-04  0.3743   0.70938
## wfir        -1.5605e-03  9.3759e-04 -1.6644   0.10077
## wser        -1.3247e-03  1.0028e-03 -1.3210   0.19107
## wmfg         4.7941e-05  3.9550e-04  0.1212   0.90389
## wfed         1.7027e-03  9.0298e-04  1.8856   0.06375 .
## wsta        -5.8974e-04  7.2093e-04 -0.8180   0.41628
## mix          2.3734e-01  4.9780e-01  0.4768   0.63510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coeftest confirms that none of the wage variables individually have any significant effect.

We finally perform a similar experiment to check if sum of all wages has any effet:

```
crime_tmp = sqldf("SELECT *, wcon+wloc+wtrd+wtuc+wfir+wser+wmfg+wfed+wsta AS 'sum_wages'
               FROM crime_data" )

sqldf('SELECT sum_wages from crime_tmp LIMIT 5')
```

```
##   sum_wages
## 1  3054.890
## 2  2652.879
## 3  2553.648
## 4  2823.133
## 5  2759.238
```

```
crime_data_tmp = crime_tmp
model3c = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
             pctmin80 + pctymle
           + log_polpc + log_taxpc
           + prbpris + avgsen
           + sum_wages
             +mix,
           data=crime_data_tmp)
summary(model3c)$adj.r.squared
```

```
## [1] 0.8011389
```

```
coeftest(model3c, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  4.4192e-01 1.6235e+00  0.2722   0.78623
## prbarr      -1.3665e-02 3.2455e-03 -4.2104 7.089e-05 ***
## log_prbconv -2.0639e-01 1.0475e-01 -1.9703   0.05255 .
## log_density  3.7020e-01 7.4823e-02  4.9477 4.571e-06 ***
## pctmin80     1.4654e-04 1.6476e-05  8.8943 2.634e-13 ***
## pctymle      7.3485e-03 1.2566e-02  0.5848   0.56047
## log_polpc    3.7807e-01 2.0025e-01  1.8880   0.06294 .
## log_taxpc   -1.0089e-01 1.5434e-01 -0.6537   0.51532
## prbpris     -3.4859e-01 3.2485e-01 -1.0731   0.28671
## avgsen      -2.3496e-02 1.0439e-02 -2.2509   0.02736 *
## sum_wages   -2.0079e-05 1.9984e-04 -0.1005   0.92024
## mix          5.4113e-02 4.8271e-01  0.1121   0.91105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model3c, c("sum_wages = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## sum_wages = 0
##
## Model 1: restricted model
## Model 2: log_crmrte ~ prbarr + log_prbconv + log_density + pctmin80 +
##     pctymle + log_polpc + log_taxpc + prbpris + avgsen + sum_wages +
##     mix
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     75
```

```
## 2     74  1 0.0101 0.9202
```

We can see from the coeftest that the sum of wages has no significant effect on the model. A similar conclusion can be drawn from the linearHypothesis test.

Measure of fit of three different models

```r
AIC(model1, model2, model3)
```

```
##        df        AIC
## model1  7  1.973512
## model2 10 -2.464560
## model3 14 -9.867252
```

The AIC scores also reflect the fit: the last model has the best (least) score, whereas the first model has the worst.

CLM observations for other models (2 and 3)

For both our other models (2 and 3), we observed that most of the CLM assumptions were satisfied and were mostly in line with what we observed in model 1. They both supported: linear in parameters, random sampling, multicollinearity and had normal residuals.

Both models 2 and 3 also supported heteroskedasticity, so we used robust standard errors while performing our analysis. We believe that by adding all the variables, as in model3, we are likely canceling out some of the variances in errors.

Model 4

We will now attempt to develop a custom model based on geographical region. This is to show that different regions can be modeled differently for better fit and analysis.

```r
crime_west = crime_data %>% filter(west==1)
crime_central = crime_data %>% filter(central==1)
crime_other_region = crime_data %>% filter(central==0 & west==0)

formula = log_crmrte ~ log_density + pctymle + pctmin80 +
  prbarr + log_prbconv + log_taxpc

logcrmrte.west.lm4a = lm(formula, data=crime_west)
summary(logcrmrte.west.lm4a)$r.squared
```

```
## [1] 0.8948345
```

```r
logcrmrte.central.lm4b = lm(formula, data=crime_central)
summary(logcrmrte.central.lm4b)$r.squared
```

```
## [1] 0.8311449
```

```r
logcrmrte.other_rgn.lm4c = lm(formula, data=crime_other_region)
summary(logcrmrte.other_rgn.lm4c)$r.squared
```

```
## [1] 0.7467755
```

The models for West and Central regions are better than the general model 2 we have in terms of adjusted R^2.

However, for other regions, our model has lower adjusted R^2 than West or Central. This shows that the observations for other regions need to be analyzed with a model different from the generic one. It could also be that our model is influenced heavily by the observations from the urban areas in West and Central.

Regression Table

	Dependent variable:		
	log_crmrte		
	(1)	(2)	(3)
prbarr	−0.014***	−0.013***	−0.014***
	(0.003)	(0.003)	(0.003)
log_prbconv	−0.236**	−0.232*	−0.181
	(0.075)	(0.091)	(0.093)
log_density	0.420***	0.321***	0.355***
	(0.046)	(0.059)	(0.068)
pctmin80	0.0002***	0.0001***	0.0001***
	(0.00002)	(0.00002)	(0.00002)
pctymle	0.014*	0.015	0.017
	(0.007)	(0.019)	(0.014)
log_polpc		0.229	0.290
		(0.205)	(0.181)
log_taxpc		−0.053	−0.049
		(0.148)	(0.148)
prbpris			−0.338
			(0.342)
avgsen			−0.022*
			(0.011)
wfir			−0.002
			(0.001)
wfed		0.001	0.002*
		(0.001)	(0.001)
mix			0.351
			(0.469)
Constant	−2.744***	−1.521	−0.835
	(0.382)	(1.853)	(1.650)
Observations	86	86	86
R^2	0.796	0.820	0.849
Adjusted R^2	0.784	0.801	0.824
Residual Std. Error	0.234 (df = 80)	0.224 (df = 77)	0.211 (df = 73)
F Statistic	62.566*** (df = 5; 80)	43.736*** (df = 8; 77)	34.258*** (df = 12; 73)

Note: *p<0.05; **p<0.01; ***p<0.001

5. Omitted variables

We have talked about omitted variables as part of our EDA. Here we continue the discussion and also talk about the direction of bias.

The important omitted variables that come to mind are:

- Median level of education in the county: The median level of education in the county may have a negative correlation to crime rate. This OV will also have a negative correlation with prbarr, prbconv, prbpris, avgsen. So, if a model uses prbarr, prbconv, prbpris, avgsen, the OVB bias would be positive.

- Median ratio of total household income to number of family members: The median ratio of total household income to number of family members in the county may have a negative correlation to crime rate. This OV will also have a negative correlation with prbarr, prbconv, prbpris, avgsen, and a positive correlation with taxpc. So, if a model uses prbarr, prbconv, prbpris, avgsen, the OVB bias would be positive. For a model using taxpc, the OVB bias would be negative

- Total number of neighbourhood crime watch groups in the county: The total number of neighbourhood crime watch groups in the county may have a negative correlation to crime rate. This OV will also have a negative correlation with prbarr, prbconv, prbpris, avgsen. So, if a model uses prbarr, prbconv, prbpris, avgsen, the OVB bias would be positive

- Family cohesiveness such as divorce rate, domestic violence: This maybe difficult to measure.The family cohesiveness measure may have a positive correlation to crime rate. This OV will also have a positive correlation with prbarr, prbconv, prbpris, avgsen. So, if a model uses prbarr, prbconv, prbpris, avgsen, the OVB bias would be positive.

- Any extreme climate change when the crime occured. This can be a significant change compared to the expected climate in that month. Higher temperatures lead to high tempers and may have a positive correlation to crime rate. This OV will also have a positive correlation with prbarr, prbconv, prbpris, avgsen. So, if a model uses prbarr, prbconv, prbpris, avgsen, the OVB bias would be positive.

- Alcoholism and substance abuse: Patient data from hospitals and police data can be a good marker to judge this. High number of cases of drug abuse correlates with higher crime rate to a large extent. This OV will also have a positive correlation with prbarr, prbconv, prbpris, avgsen. So, if a model uses prbarr, prbconv, prbpris, avgsen, the OVB bias would be positive.

- Repeat crimes: The data does not have information about serial criminals as against first-time offenders. If most of the crime is due to repeat offenders, then we will have to accommodate our policies accordingly.

- Underreporting and unreported crimes: It is also worth checking via other sources how much crime goes unreported in North Carolina, and whether data is often "corrected" as time goes by.

- Tourism: As we outlined in the initial analysis, tourism is an omitted variable that can affect our analysis. If a county depends on tourism, it may have mostly floating population. This causes problems when modeling using density of resident population.

- Geography: Geography also tends to affect crime, in either aiding or hindering it: Western regions of North Carolina are hilly, whereas the eastern ones are plain.

6. Conclusion

Based on the above study, here are the conclusions we would like to offer the political campaign:

- Crime rate is most correlated (0.7) with population density. Median crime in urban areas is more than double those in rural areas. Our policies should make our cities safer, in order to reduce crime significantly.

- As probability of arrest and/or conviction increase, crime decreases. Law enforcement is a good deterrent to crime.
- Prison by itself does not correlate with crime, nor does the length of prison sentence. We should use this data to argue for shorter sentences and alternatives to prison, such as reform and counselling. We should be careful to continue to provide strong disincentives to crime, however.
- Police per capita correlates positively with crime: this may mean we have not improved the effectiveness of our police in crime-infested areas. This may also be due to omitted variables and is worth exploring further.
- Similarly, wealth as proxied by tax revenue begets crime. This may also suggest that we have the money and should be able to reroute tax dollars better to fight crime in high-crime areas.
- Minorities have moderate correlation with crime. Our policies should address integration of minorities into the mainstream and reduce segregation.
- Similarly, we should investigate correlation of crime with young men. If these men are driven to crime due to lack of education or unemployment (omitted variables in this dataset), we should pay attention to reforming our education or job market.