W203 Lab 3 - Reducing Crime

Mikra Walekova | Dominik Graf

Introduction

In this paper, we are going to review the factors that drive local crime rate intending to identify factors that are within the control or influence of local government to propose policies that could lead to a reduction of local crime. These policy recommendations are intended for local political campaigns in North Carolina.

Data Loading

The following study is based on a single cross-section of data, a selection of North Carolina counties, of a multi-year panel, of which only data from 1987 are used in this study. Panel data is known to have limitations (i.e. unabalanced panels with non-randomly missing data), and therefore, our ability to make inferences about the population based on this data set is limited.

There are six blank records and one duplicate record in the source file which are removed.

```
raw_data <- raw_data[!is.na(raw_data$year), ]
raw_data <- raw_data[!duplicated(raw_data), ]
```

This reduces the dataset from 97 to 90 records. A second thing we notice when loading the dataset is that the **prbconv** variable is read in as a character string even though all the values are numbers. Therefore, we convert the variable type to numeric.

```
raw_data$prbconv <- as.numeric(raw_data$prbconv)
```

	Min.	Median	Mean	Max.	Explanation
county	1	103	100.6000	197	County identifier
year	87	87	87	87	Year
crmrte	0.0055	0.0300	0.0335	0.0990	Crimes committed per person
prbarr	0.0928	0.2715	0.2952	1.0909	Prob. of arrest
prbconv	0.0684	0.4517	0.5509	2.1212	Prob. of conviction
prbpris	0.1500	0.4222	0.4106	0.6000	Prob. of prison sentence
avgsen	5.3800	9.1100	9.6889	20.7000	Avg. prison sentence (days)
polpc	0.0007	0.0015	0.0017	0.0091	Police per capita
density	0.00002	0.9792	1.4357	8.8277	People per 100 sq. miles
taxpc	25.6929	34.9161	38.1610	119.7615	Tax tevenue per capita
west	0	0	0.2444	1	Is west
central	0	0	0.3778	1	Is central
urban	0	0	0.0889	1	Is urban
pctmin80	1.2837	24.8516	25.7129	64.3482	% minority
wcon	193.6432	281.1624	285.3532	436.7666	Wage: Construction
wtuc	187.6173	404.7800	410.9065	613.2261	Wage: Trns, util, commun
wtrd	154.2090	202.9879	210.9214	354.6761	Wage: Wholesale retail trade
wfir	170.9402	317.1257	321.6213	509.4655	Wage: Finance, insurance, real estate
wser	133.0431	253.1188	275.3379	2,177.0680	Wage: Service industry
wmfg	157.4100	321.0500	336.0327	646.8500	Wage: Manufacturing
wfed	326.1000	448.8550	442.6189	597.9500	Wage: Federal employees
wsta	258.3300	358.4000	357.7402	499.5900	Wage: State employees
wloc	239.1700	307.6500	312.2801	388.0900	Wage: Local government employees
mix	0.0196	0.1009	0.1290	0.4651	Face-to-face / other crime mix
pctymle	0.0622	0.0777	0.0840	0.2487	Young male %

The **county** variable represents a unique numeric identifier covering 90 unique counties. The **year** interval variable shows the year 1987 for all records. Looking at the rest, **crmrte**, **polpc**, **density**, **taxpc**, **mix** are ratio variables representing a ratio; **prbarr**, **prbconv**, **prbpris**, **pctmin80**, **pctymle** are ratio variables representing percentages; **west**, **central**, **urban** are 0-1 coded dummy variables; and **avgsen**, **wcon**, **wtuc**, **wtrd**, **wfir**, **wser**, **wmfg**, **wfed**, **wsta**, **wloc** are ratio variables representing an average (measure of central tendency). The wage variables are given as weekly wages. For the three dummy variables (**west**, **central**, **urban**), we read their 0-1 coding as false-true. The table above gives an explanation for each variable.

One important thing we notice is that **density** is not given as people per square mile. This becomes important in the next section when we clean our data as there is one anomalous point in this variable. When we look online, we see that North Carolina has a surface area of 53,818 square miles while the population of the state in 1987 was approximately 6,481,000. This implies that the average density in the state was around 120.4 people per square mile in 1987. When we look at our dataset though, we see an average density value of 1.44 and a maximum density of 8.83, therefore, something does seem right. We suspect that the **density** variable is actually in terms of 100 people per square mile.

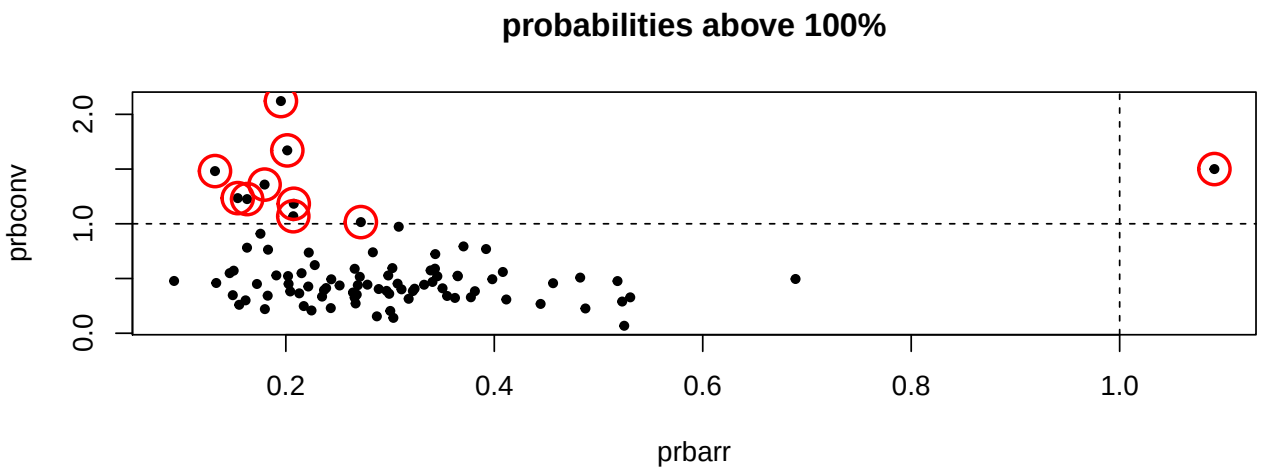
We notice that the **pctmin80** variable uses data from 1980 even though our cross-sectional data is from 1987. Thus, there is a bit of a mismatch here. It is possible that the percentage of minorities changed over seven years and therefore using this variable could lead to "garbage in, garbage out". We couldn't find minority population data for the state of North Carolina but we did look at the national change from the 1980 to 1990 census. We assume the change in the national percentage of minorities is indicative of the change in the state. From 1980 the percent of minorities increased from 20.4% to 24.4% in 1990. We do not believe this is a large enough change to be of significance. For this reason, we believe **pctmin80** (from 1980) is a fairly good representation of the actual percentages in 1987.

Furthermore, we see that our dataset has 90 counties, while the state of North Carolina has 100 counties. Therefore, we are missing 10%. Since we are missing some counties and do not know how to weight the importance of each county (e.g. by population) within the state, we do not feel comfortable making assertions at the state level, instead we keep the inferences we make at the county level. At the county level, we consider 90 our to 100 counties as a sufficient sample size.

Data Cleaning

In this section, we detect and correct corrupt or inaccurate records.

Starting with **prbarr** and **prbconv**, we notice values above 1, implying the probability of arrest or conviction is above 100%. This is not intuitive, and the reason for this lies in the proxy for these variables. The **prbarr** is proxied by the ratio of arrests to offences in a given year so it is possible that in a given year there could be a backlog of offence cases originating from previous years for which arrests are made in the current year. Therefore, the total offences for which arrests can be made is greater than or equal to the number of offences in a given year, resulting in the possibility of the number of arrests in a year being higher than the number of offences in a year. The same logic applies to the **prbconv**, which is proxied by the ratio of convictions to arrests. To correct this, we top-code the two variables by setting an upper bound of 1 (or 100%).



This affects 1 **prbarr** observation and 10 **prbconv** observations.

```
raw_data$prbarr <- pmin(1, raw_data$prbarr)
raw_data$prbconv <- pmin(1, raw_data$prbconv)
```

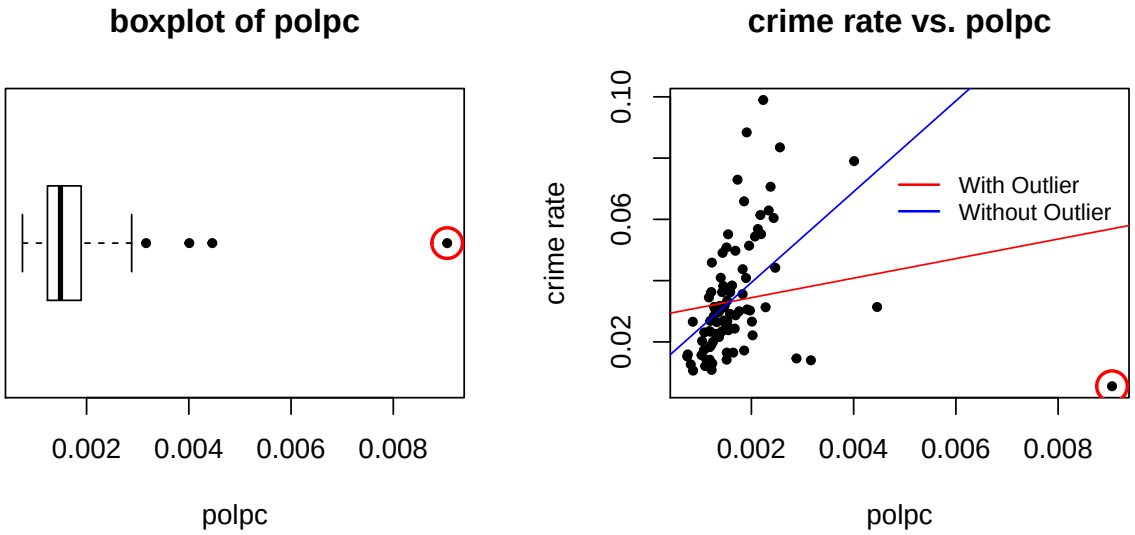
Furthermore, for easier readability and interpretation later in the report, we scale up the **prbarr**, **prbconv**, **prbpris** and **pctymle** percentage variables by 100 so that an increase of 1 corresponds to 1% (**pctmin80** is already scaled by 100).

Another anomalous value we see is in **density** where the smallest value is 0.00002, which is $0.00002 \times 100 = 0.00203$ people per square mile. This is a an extremely small number with the second smallest density being 0.30057 which is 14776x bigger. For there to be 1 person living in the county this record is representing, the surface area of the county would have to be $(1/(0.00203))$ 492 square miles. When we look online, we see that the surface areas of the counties in North Carolina range from 172 square miles (Chowan) to 949 square miles (Robeson). Therefore, if this density value were to represent the largest county (Robeson), it would mean only 2 people ($949 \times 0.00203 = 1.926$) were living there, which does not make sense and thus leads us to believe it is an error. Because the variable **urban** is strongly correlated with density (82.1%), which we can be seen in the correlation matrix later in this report, and because they are intuitively related (urban areas usually have higher densities), we replace the **density** outlier with the mean of the non-urban densities since our outlier's record is in a non-urban region.

```
index <- raw_data$density == min(raw_data$density)
raw_data$density[index] <- mean(raw_data$density[raw_data$urban == 0 & !index])
```

Therefore, the **density** outlier is replaced with 1.061.

When we look at the **polpc** variable we see one ostensibly significant outlier with a value of 0.00905. If we run a univariate regression on the crime rate, we can see this outlier/leverage point has significant influence.

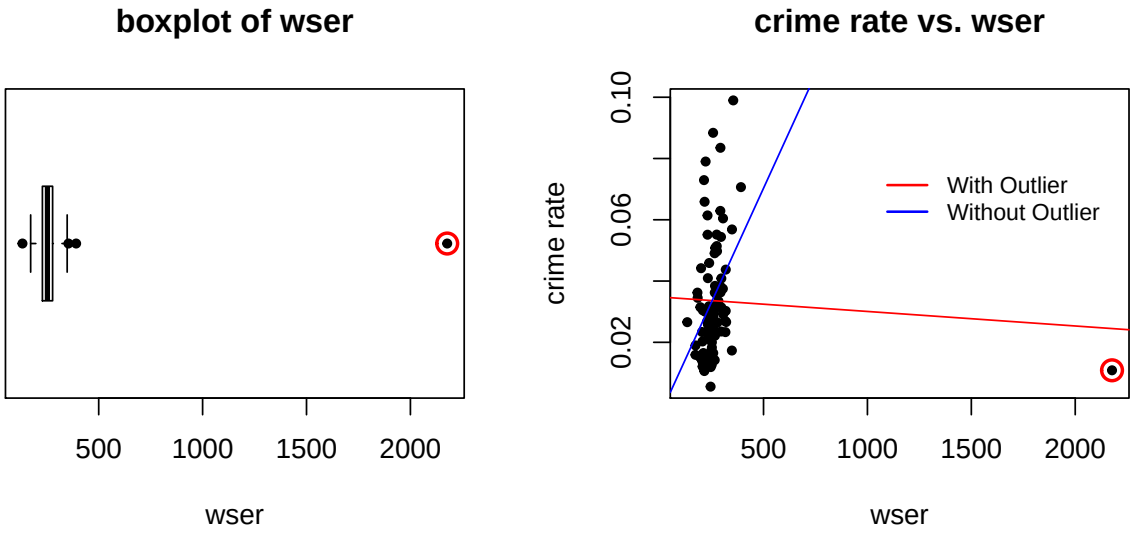


This outlier is 4.6 standard deviations aways from the second largest value, which is quite extreme. To avoid the high influence, we decide to replace this value. We use **prbarr** and **taxpc** to impute a value for this outlier since these variables are highly correlated (38.7% & 28.1%) and intuitively related to police per capita. For higher numbers of police per capita, we would expect higher probabilities of arrests and a higher tax revenue per capita to pay for them. We fit a linear model to impute a value for police per capita. The R^2 (32.2%) of this model is not very high but nevertheless, this approach retains more information than replacing with the mean or top-coding with an arbitrarily-chosen value.

```
index <- which.max(raw_data$polpc)
raw_data$polpc[index] <- predict(
  lm('polpc ~ prbarr + taxpc', data = raw_data[-index, ]), raw_data[index, ])
```

Therefore, the **polpc** outlier is replaced with 0.00111.

When we look at the **wser** variable, we see another significant outlier with a value of 2177.1. If we run a univariate regression on the crime rate, we can see this leverage point has significant influence.



This outlier is 8.6 standard deviations aways from the second largest value, which is quite extreme. Looking at

historical inflation rates, we can estimate what $1 in 1987 would be worth today (in 2019). This outlier implies a 1987 weekly wage of $2177.1 which is equivalent to a 2019 weekly wage (inflation adjusted) of $4909.3, or an annual salary of $255283. It could be the case that this particular county has only one business in the service industry consisting of only a few people, with the owner making a few times more than this wage, thus drastically skewing the distribution upwards. In any case, if this point is correct, we believe it to be an aberartion and not a good representation due to a small sample size within the county (assuming our thought process is correct). Therefore we decide to replace it. Since we do not know which other wage variables are most appropriate to use to impute a value, we use all of the wage variables along with K-Nearest-Neighbours to impute a value for this outlier. This KNN approach to impute a value for **wser** retains more information than simply replacing with the mean or top-coding with an arbitrarily-chosen value.

```
wage <- c('wcon', 'wtuc', 'wtrd', 'wfir', 'wser', 'wmfg', 'wfed', 'wsta', 'wloc')
index <- which.max(raw_data$wser)
raw_data$wser[index] <- NA
raw_data[, wage] <- DMwR::knnImputation(raw_data[, wage])
```

Therefore, the **wser** outlier is replaced with 220.8.

Lastly, we assume **west** and **central** are mutually exclusive in the sense that if **west** equals 0, the record represents the eastern or central region of the state and represents the west if **west** equals 1. However, there is one record where **west** and **central** are both equal to 1, which is not possible. Because **west** is highly correlated with **pctmin80** (-63.4%), we use this to see whether this record is likely in the western region or not.

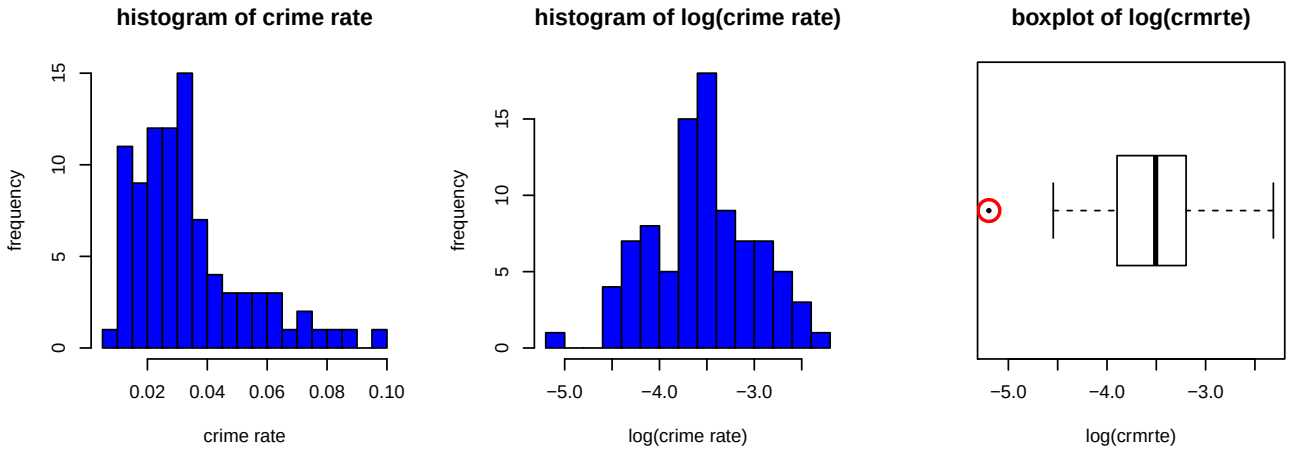
We can see the mean **pctmin80** in the west is 14.6% and 29.1% outside of the west while our anomalous record's **pctmin80** is 13.3%. Therefore, we classify this outlier as west.

```
raw_data$central[index] <- 0
```

Looking at the county split, we can see that of the 90 counties in our dataset, 22 (24.4%) are in the west, 33 (36.7%) are in the center and the remaining 35 (38.9%) we classify as east. We looked online to see what the actual regional split is, but there is no standard way of classifying west, central and eastern North Carolina. While we assume that the classification method used for our data was robust, any conclusions relating to the regionality will have limitations.

Before moving on, we would like to mention that we considered creating a new variable to proxy income inequality, which we believe is positively correlated with crime-rate. This proxy would have been computed using the nine wage variables by subtracting the highest wage from the lowest wage for each record. However, we did not believe this was a reasonably good proxy for this variable and prefer a more appropriate measure of income inequality such as the Gini-coefficient for each county.

Dependent Variable



To understand the determinants of crime and how to influence them through local government, it is necessary first to establish what we are aiming to explain - the dependent variable. Our conceptual definition of crime is the number of crimes committed per person in a given area. We operationalize this with the **crmrte** variable which represents the crime-rate or crimes committed per person in a given county. Since we are given county-level data, we find little difference between our conceptual and operational definition.

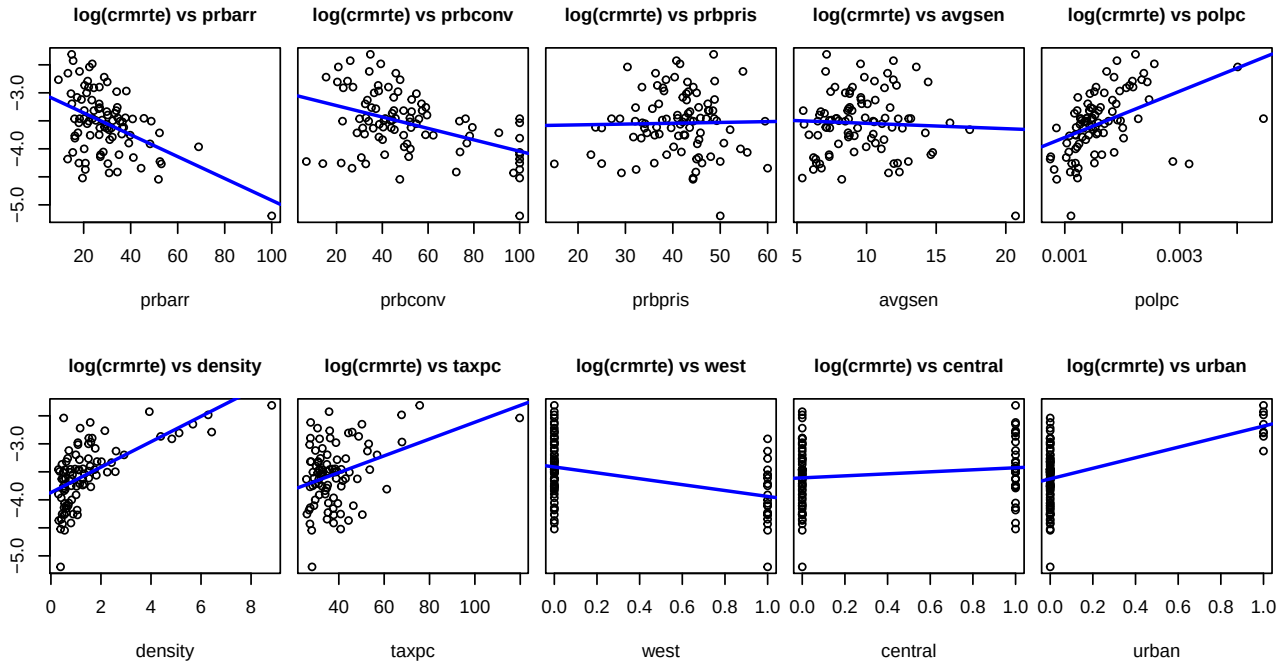
After the examination of the crime-rate (**crmrte**) histogram, we can see a positive skew. For our linear regression model, it is preferable that our variables are normally distributed, especially our dependent variable as any unexplained variation will show up in the residuals, which we assume to be normally distributed. We can see that this log-transformed distribution looks much more symmetric and normal. With this transformation, the interpretation changes to a percent change, which is easier to understand than an increase in crimes per person. Furthermore, the exact interpretation of the slope coefficients depends on whether we have a log-lin or log-log relationship. For a log-lin relationship, a one-unit increase in the independent variable results in a change of $[(\exp(\beta) - 1) \times 100]$ percent in **crmrte**. This will be the interpretation of any untransformed variables. For a log-log relationship, a 1% increase in the independent variable results in a change of β percent in **crmrte**. This will be the interpretation of any log-transformed variables. One assumption we have to make here is that **crmrte** is not equal to zero since the log of zero is undefined. Nevertheless, we feel comfortable asserting that in any given year, there will be at least one crime committed per county. Lastly, we want to mention that we see one **log(crmrte)** value that seems like an outlier in the boxplot above. We will pay attention to this leverage point (#51) later to see if it has influence.

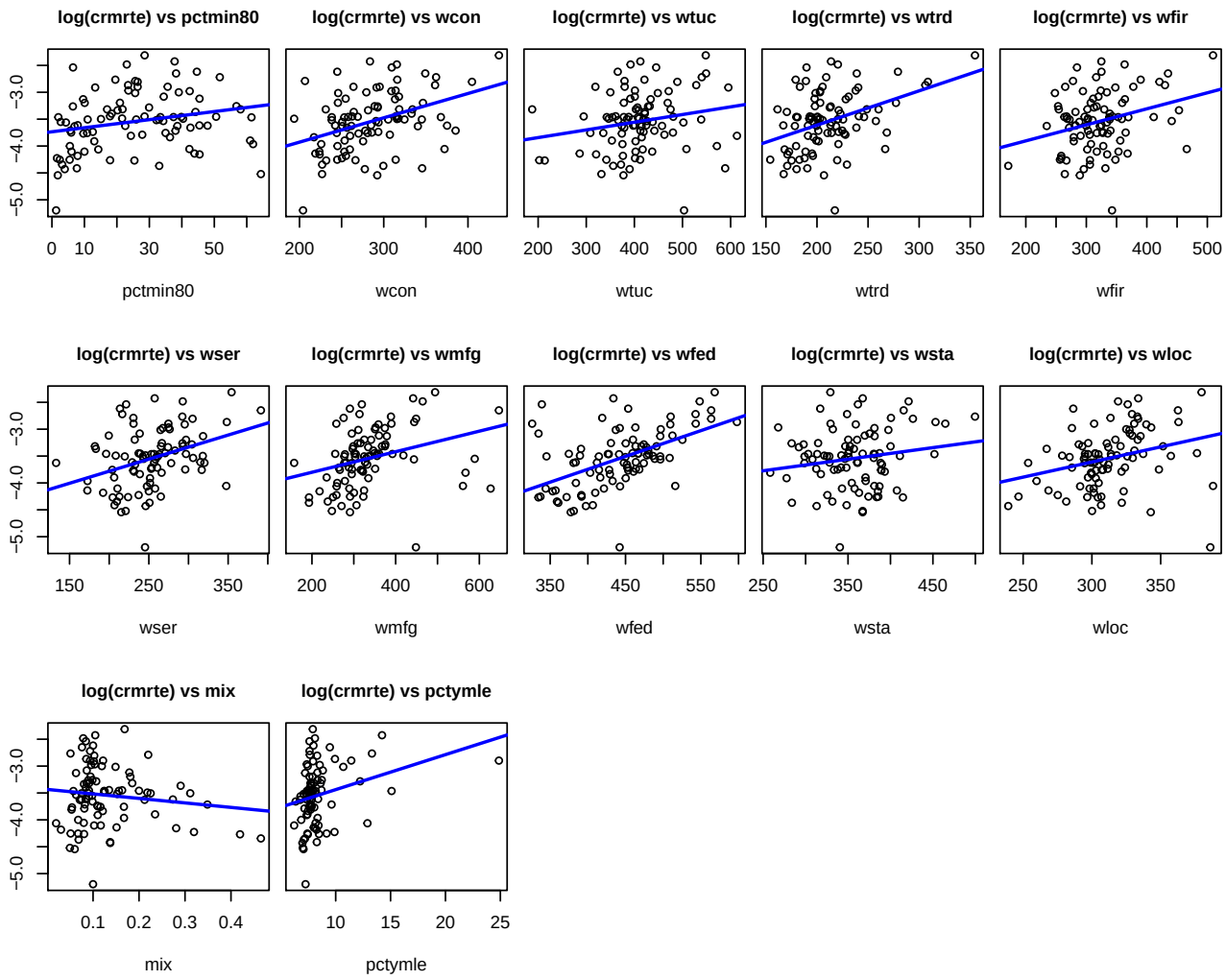
```
raw_data$crmrte <- log(raw_data$crmrte)
names(raw_data)[names(raw_data) == 'crmrte'] <- 'log(crmrte)'
```

Exploratory Data Analysis

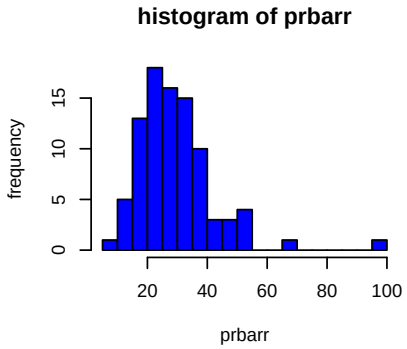
Before we jump into building the three models, we are interested in exploring the relationships between the possible independent variables and our dependent variable.

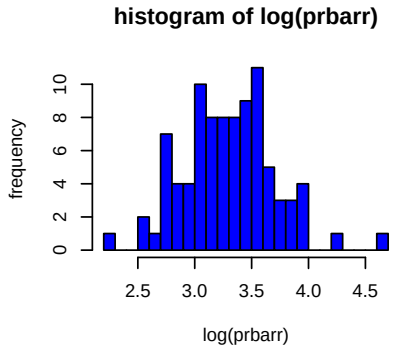
We begin by looking at the univariate linear regressions for all the possible independent variables (every y-axis represents **log(crmrte)**).

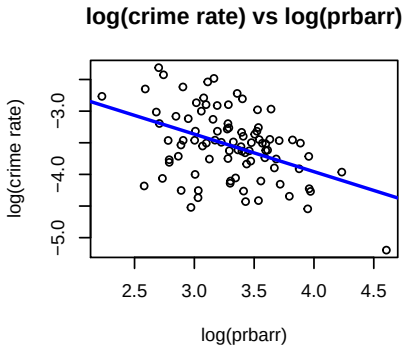


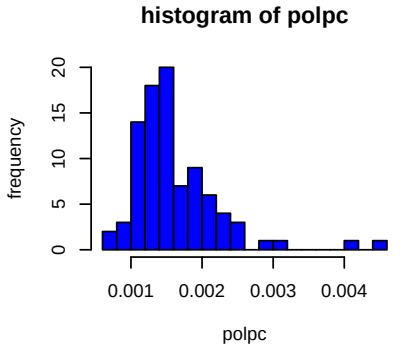

log(crmrte) vs pctmin80
log(crmrte) vs wcon
log(crmrte) vs wtuc
log(crmrte) vs wtrd
log(crmrte) vs wfir
log(crmrte) vs wser
log(crmrte) vs wmfg
log(crmrte) vs wfed
log(crmrte) vs wsta
log(crmrte) vs wloc
log(crmrte) vs mix
log(crmrte) vs pctymle
pctmin80
wcon
wtuc
wtrd
wfir
wser
wmfg
wfed
wsta
wloc
mix
pctymle

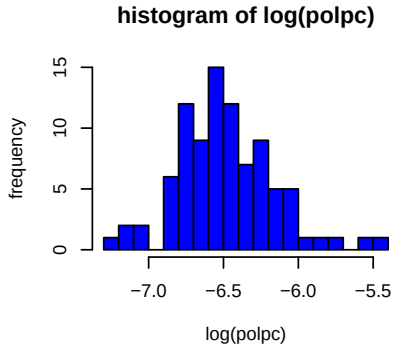
We notice a few non-linear patters in the scatterplots of **log(crmrte)** vs. **prbarr**, **polpc**, **density**, **taxpc**, **mix** and **pctymle** which leads us to believe a log-transformation may be appropriate for these variables. We explore this below in the histograms and scatterplots.

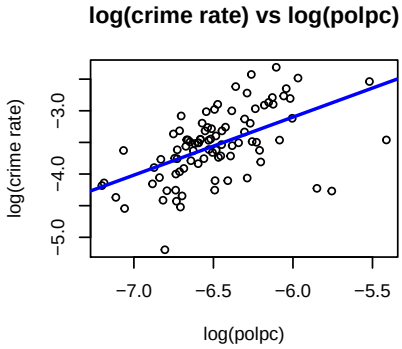


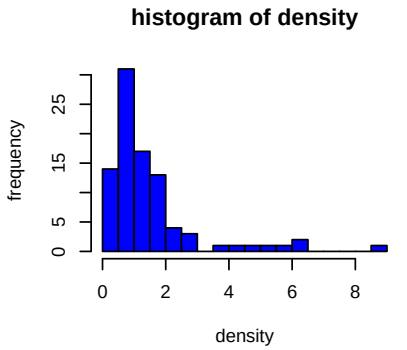


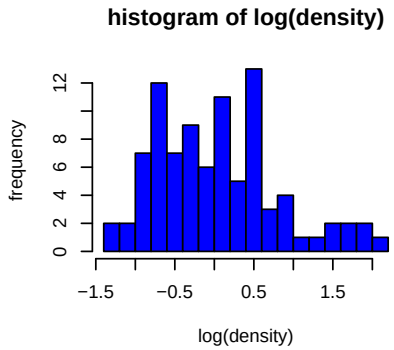


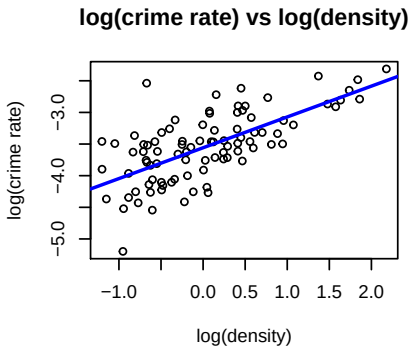


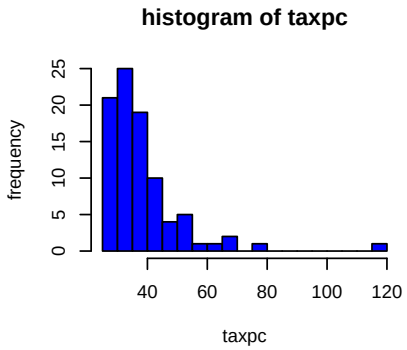


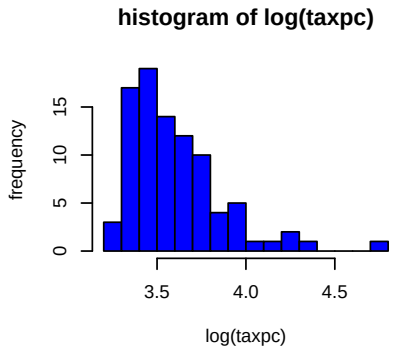


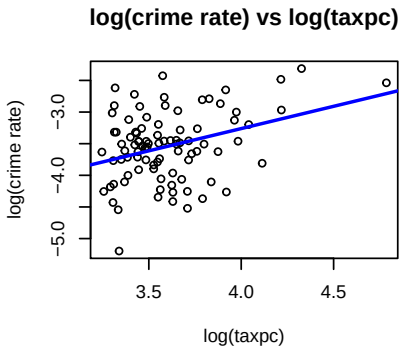


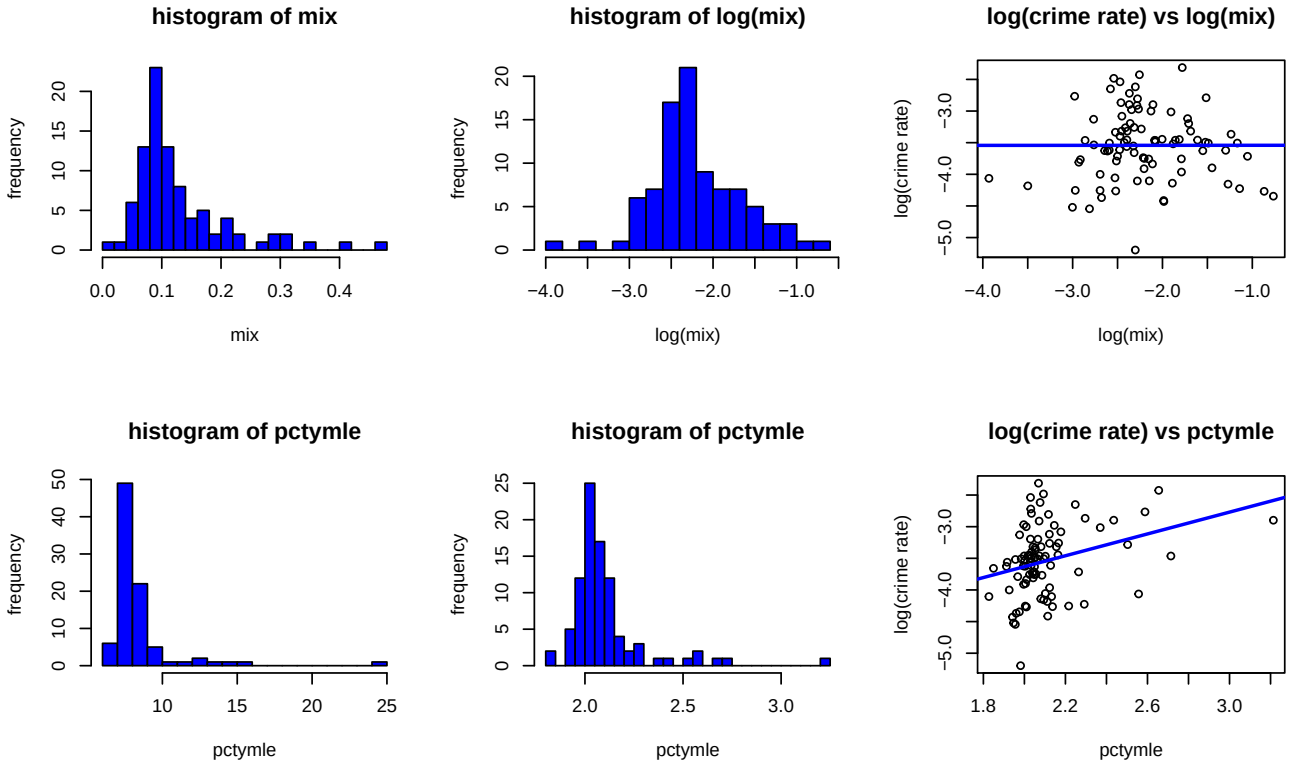








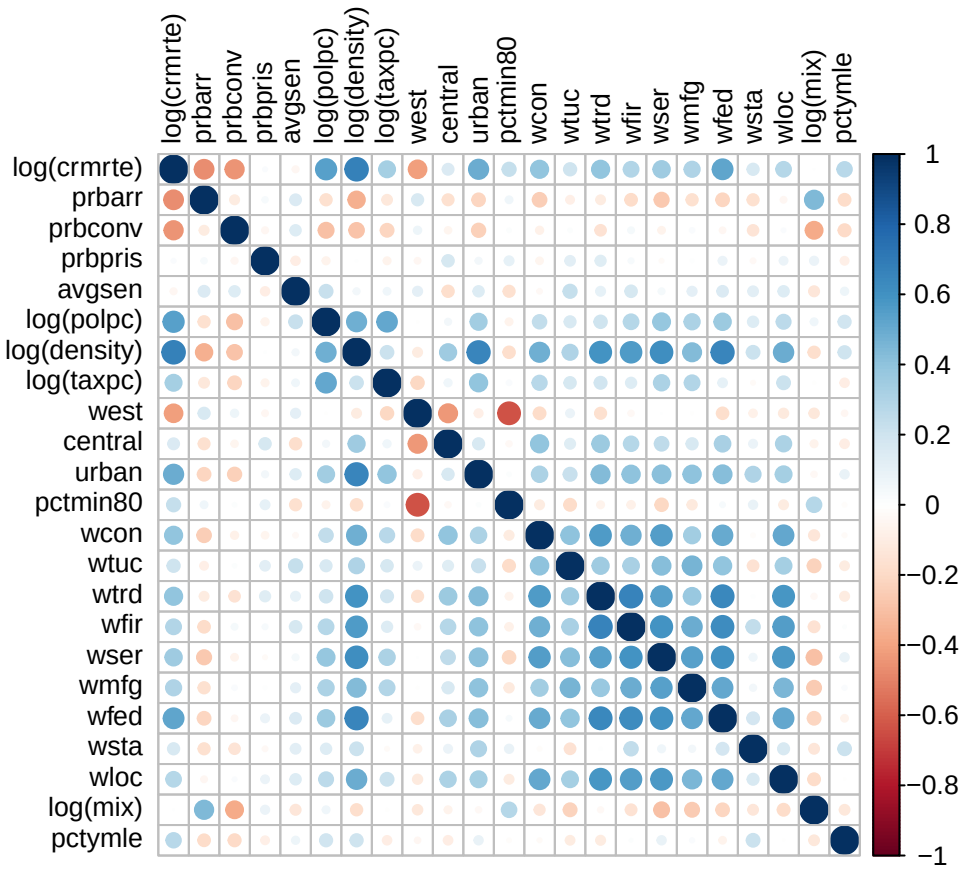




We can see the distributions of **prbarr**, **polpc**, **density**, **taxpc**, **mix** and **pctymle** all have a positive skew and after the log-transformation, the distributions all look more symmetric and normal. The univariate linear fits look much more appropriate as well after the log-transform. However, we do not feel comfortable log-transforming percentages that are bound between 0-100% because we don't find it appropriate to use a log-transformation on variables where an obvious maximum exists; in our case 100%. Therefore, despite the log-transformation helping the variables follow a more normal distribution, we decide against this. We also considered transforming these variables to a log-odds ratio but since we have some values at 100%, which would result in undefined values, we decide against this in addition to the difficult interpretation of the resulting slope coefficients. Therefore, we do not log-transform **prbarr** and **pctymle**. Moving to the other log-transformations, one assumption we have to make, despite all of these variables being ratios and thus having a true zero, is that they are not equal to zero since the log-transform of zero is undefined. For **polpc**, **density** and **taxpc** we feel comfortable making this assumption since we assume each county has at least one person, one police officer and some tax revenues. For **mix** we cannot be sure this variable is non-zero for other years and other counties not included our dataset. Nevertheless, we make this assumption.

The log-transform changes the interpretation of these three variables to a log-log relation: (**log(polpc)**) a 1% increase in **polpc** will result in a β_{polpc}% change in crime-rate; (**log(density)**) a 1% increase in **density** will result in a $\beta_{density}$% change in crime-rate; (**log(taxpc)**) a 1% increase in **taxpc** will result in a β_{taxpc}% change in crime-rate; and (**log(mix)**) a 1% increase in **mix** will result in a β_{mix}% change in crime-rate. The log-transformation helps with the interpretation as it changes to a percent change in these ratios, which is easier to grasp than an increase of one in police per capita, people per square mile, tax per capita and face-to-face vs other types of offences.

We now explore the correlation matrix of the dependent and candidate independent variables.



From the correlation matrix, we can see the variables with the strongest correlation with **log(crmrte)** are **log(density)** (67.7%), **log(polpc)** (54.5%), **wfed** (52.3%), **urban** (49.1%), **prbconv** (-44.4%), **prbarr** (-47.0%) and **west** (-41.4%). For these variables, we also note a strong linear relationship in the univariate scatterplots above, particularly for **log(density)**. The least correlated variables are **log(mix)** (0.0%), **prbpris** (2.1%) and **avgsen** (-4.9%). Looking at the univariate scatterplots of these three latter variables, we can comfirm that no discernable linear relation seems to exist.

Among the wage variables, we notice a particularily strong inter-variable positive correlations among all the nine wage variables (**wcon**, **wtuc**, **wtrd**, **wfir**, **wser**, **wmfg**, **wfed**, **wsta**, **wloc**), which can cause multicollinearity within a model if more than one of these variables are included. We addresss this concern later for model 3 when we include all variables. Furthermore, we notice three strong correlations among candidate independent variables: (1) **west** and **pctmin80** (-63.4%); (2) **log(polpc)** and **log(taxpc)** (51.1%); (3) **log(density)** and **urban** (65.8%); and (4) **prbarr** and **log(mix)** (44.3%). The negative correlation between **west** and **pctmin80** may imply that percent of minorities is higher in the central and eastern regions than the western region. The positive correlation between **log(polpc)** and **log(taxpc)** makes sense as higher tax revenues are needed to fund a higher police presence. And lastly, the positive correlation between **log(density)** and **urban** is expected since an urban county tends to have a higher population density than a rural county. The positive correlation between **prbarr** and **log(mix)** leads us to believe that there are more arrests for face-to-face offenses than for other types of offenses.

Model Building Process

We build three models in this report. The first (base) model will be built with only 1-2 key variables that are influenceable by the local government to reduce crime. The second model will add covariates to increase the accuracy of the base model by adding variables via a naunced, iterative approach. Care is given to ensure interpretation and intuition of results. The third model will then include all other remaining variables to test the robustness of our models 1 & 2.

The following report does not perform any validations such as K-folds cross-validation on the results. The aim of this report is to find preliminary results that can be explored further in subsequent study.

The model we find to be most important is model 2. Therefore, for this model, we will present a more detailed assessment of the CLM assumptions along with responses to any violations.

Model 1

We build our base model by identifying key variables that are (1) strongly correlated with the dependent variable, (2) within the control or influence of local government to reduce crime and (3) intuitively related with crime.

We let the data speak to us and begin by looking at the variables most correlated with our dependent variable. From the variables **log(crmrte)** is most correlated with (**log(density)**, **log(polpc)**, **wfed**, **urban**, **prbconv**, **prbarr**, **west**), only **log(polpc)**, **prbconv** and **prbarr** are influenceable by local government. Local government can influence these variables by: (1) prioritizing local law enforcement to focus on making arrests (↑ **prbarr**); (2) implement more robust procedures for evidence collection (↑ **prbconv**); (3) speed up court process for convictions (↑ **prbconv**); and (4) increase spending on law enforcement to increase police presence (↑ **polpc**). These present actionable policy recommendations, which we are interested in.

We expect the relation of **prbarr** and **prbconv** to both be negative with the crime-rate since these two variables are a risk to someone considering committing a crime as they represent the probability of "getting caught" and "not getting away with it". Our expectations are in-line with the correlations of these two variables. For **log(polpc)** we expect the relationship to also be negative with the crime-rate since we argue that a more concentrated police presence (**polpc**) should reduce crime-rate. However, this explanation is not in-line with the correlation of **log(polpc)** with **log(crmrte)**, which is positive. This is counterintuitive. The issue here, we believe, is that we have a snapshot of data for one year (1987) and cannot see how these variables evolved through time to assess causation. We believe a time-dimension is playing a key role here. For example, we can think that police per capita is a decision that is taken ex-post, after many crimes have happened. This implies the causal relationship ↑ **crmrte** → ↑ **polpc**. Based on this, and for our snapshot, we could assume that police presence is proportional to crime-rate with no other relationship. However, we do not feel comfortable with this assertion and therefore drop this variable from consideration.

This leaves us with **prbarr** and **prbconv** as our key variables. Because the correlation between these two variables is fairly low (-10.8%), we are not worried about any multicollinearity being injected into our model and are therefore comfortable using both for our base model. The model we fit is:

$$log(\text{crmrte}_i) = \beta_0 + \beta_{\text{prbarr}} \cdot \text{prbarr}_i + \beta_{\text{prbconv}} \cdot \text{prbconv}_i + u_i$$

	Dependent variable:	
	log(crmrte)	
	Unadjusted SE	HC-Adjusted SE
	(1)	(2)
prbarr	-0.022^{***}	-0.022^{***}
	(0.003)	(0.005)
prbconv	-0.012^{***}	-0.012^{***}
	(0.002)	(0.002)
Constant	-2.316^{***}	-2.316^{***}
	(0.146)	(0.219)
Observations	90	90
R^2	0.468	0.468
Adjusted R^2	0.456	0.456
Residual Std. Error (df = 87)	0.405	0.405
F Statistic (df = 2; 87)	38.289^{***}	38.289^{***}
Note:	*p<0.05; **p<0.01; ***p<0.001	

Our base model explains (R^2) 46.8% of the variation in **log(crmrte)**, the adjusted-R^2 is 45.6% and the AIC is 97.56. The way we interpret this model is as follows: (**prbarr**) for an increase of 1 in **prbarr** (which is scaled by 100%), *ceteris paribus*, crime-rate decreases by 2.16% and (**prbconv**) for an increase of 1 in **prbconv** (which is scaled by 100%), *ceteris paribus*, crime-rate decreases by 1.14%. Both coefficients are highly statistically significant (<0.1%) regardless if the unadjusted or HC-adjusted standard errors are used. Furthermore, the signs of our two covariates are in-line with what we expect them to be.

Since $\hat{\beta}_{\text{prbarr}}$ and $\hat{\beta}_{\text{prbconv}}$ are both highly statistically significant, we can see evidence that local government can reduce the crime-rate by increasing the risk to potential criminals from committing a crime by increasing the probability of arrest and probability of conviction. This can be achieved by (1) prioritizing local law enforcement to focus on making arrests (↑ **prbarr**); (2) implement more robust procedures for evidence collection (↑ **prbconv**); and (3) speed up court process for convictions (↑ **prbconv**). For local political campaigns in North Carolina looking to reduce crime as a campaign promise, this gives evidence on how they can achieve this. However, does increasing **prbarr** have the same effect on reducing crime as does increasing **prbconv**? We test for this with the hypothesis that $H_0 : \beta_{prbarr} - \beta_{prbconv} = 0$ and $H_A : \beta_{prbarr} - \beta_{prbconv} \neq 0$. We choose a two-sided test here since it is harder to reject the null hypothesis, but we note that $\hat{\beta}_{prbarr} > \hat{\beta}_{prbconv}$ and will thus make a conclusion based on this inequality. We do this test by setting

$$\theta_1 = \beta_{prbarr} - \beta_{prbconv} \ ,$$

re-writing the hypothesis as

$$H_0 : \theta_1 = 0 \text{ and } H_A : \theta_1 \neq 0 \ ,$$

and re-writing our population model as

$$\log(crmrte) = \beta_0 + (\theta_1 + \beta_{prbconv})prbarr + \beta_{prbconv} \cdot prbconv + u \ .$$

$$\log(crmrte) = \beta_0 + \theta_1 prbarr + \beta_{prbconv}(prbarr + prbconv) + u$$

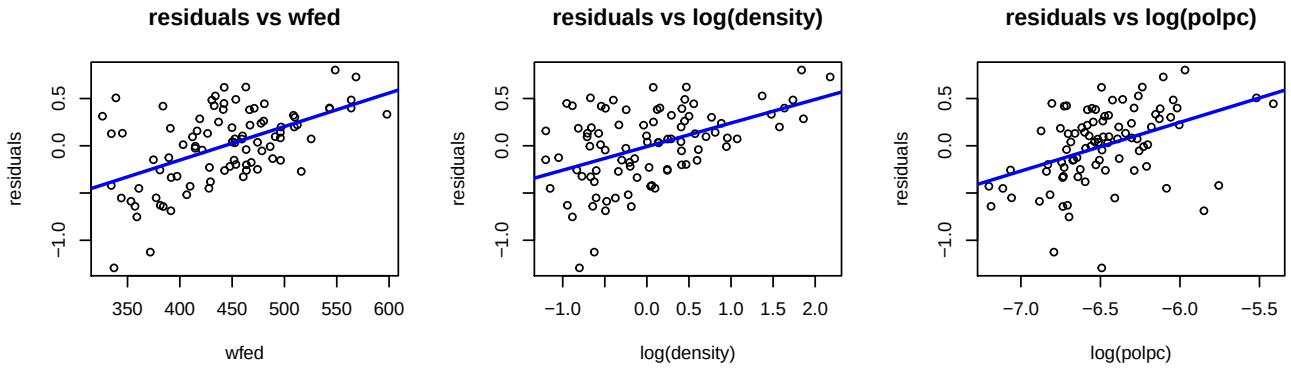
When we run this regression, we reject the null hypothesis that $\theta_1 = 0$ at the 5% significance level as the t-statistic (using HC-adjusted SE) for the slope coefficient θ_1 is -2.32 and the p-value is 2.278%. Therefore, we can say that there is a statistically significant difference between β_{prbarr} and $\beta_{prbconv}$ and the data suggests that reducing **prbarr** has a stronger effect on reducing crime than **prbconv**. This is intuitive as from the standpoint of a potential criminal; they need to first worry about the probability of getting caught (arrested) and then conditional on them being caught, the probability of them being convicted. Therefore, since the arrest comes before the conviction, it is a stronger short-term risk factor and will most likely impact a criminal's decision more.

Even though from a statistical perspective these two factors are significant, are they practical? Looking at **prbarr**, we can see a 2.16% reduction in crime-rate for a 1% increase in **prbarr** is quite important. Likewise, a 1.14% reduction in crime-rate for a 1% increase in **prbconv** is also quite important, but as confirmed in our t-test above ($H_0 : \beta_{prbarr} - \beta_{prbconv} = 0$), the effect of **prbconv** is less than **prbarr**'s (almost half). Nevertheless, both are practical from the standpoint of reducing crime.

Model 2

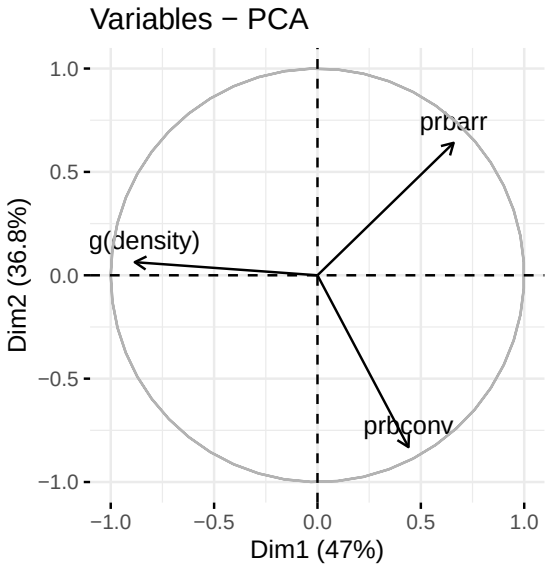
Our goal for model 2 is to increase the accuracy of our base model with additional covariates. We assume that the quadratic curve in both the "Residuals vs Fitted" and "Scale-Location" plots of model 1, pictured later in the report (CLM Assumptions section), is a result of omitted variables in our model 1. We considered a forward-stepwise regression based on AIC but decided against this due to the risk of blind data dredging and the chance of obtaining a model that may have unintuitive coefficients. We, therefore, take a more nuanced, iterative approach of adding covariates by considering (1) correlation with model residuals, (2) intuition on the relationship with crime-rate, (3) correlation with the crime-rate and (4) correlation with existing covariates.

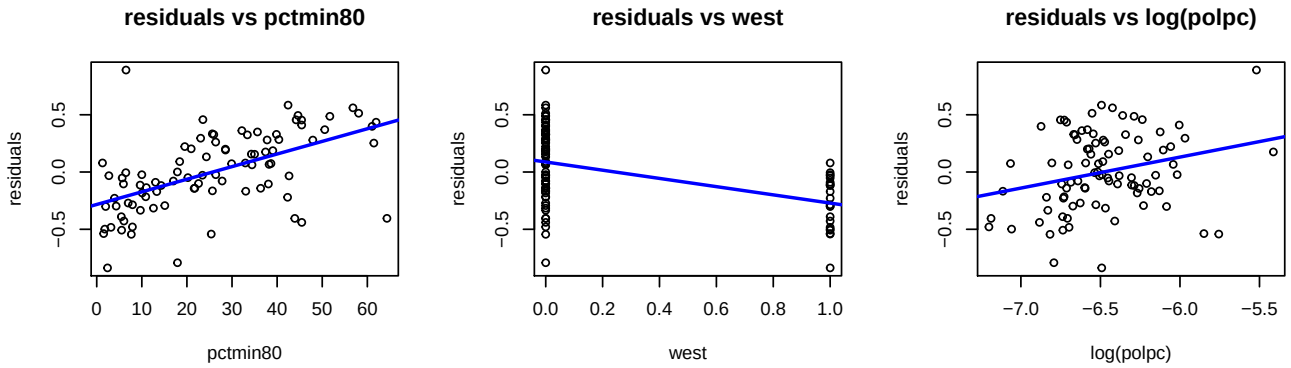
We begin by looking at which variables are most correlated with the base model's residuals.




```
wfed log(density)   log(polpc)
0.53         0.48         0.42
```

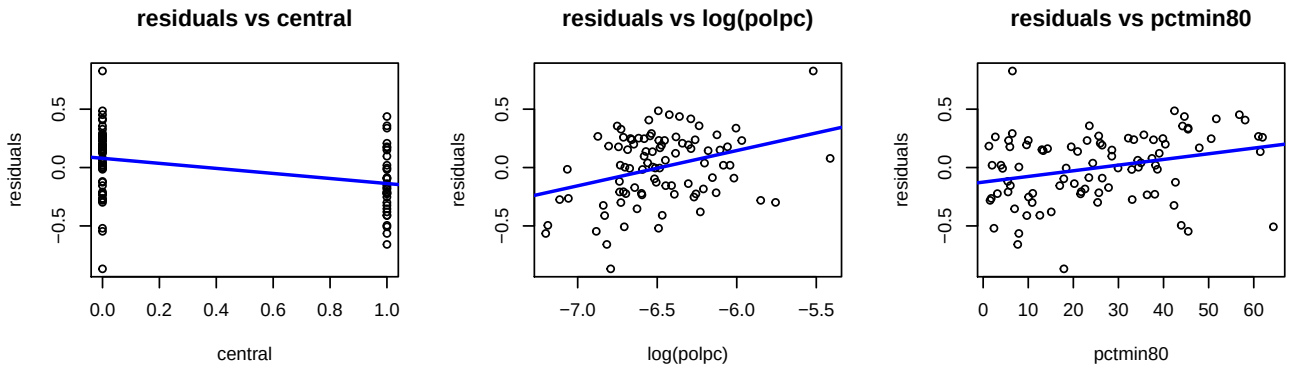
The three most correlated variables with the base model's residuals are **wfed** (53.2%), **log(density)** (47.6%) and **log(polpc)** (42.0%). From the univariate plot above, we can can see that there is a strong linear fit with residuals with **wfed** and **log(density)**, and a moderate fit with **log(polpc)**. Of these variables, we see that **wfed** has a positive correlation with **log(crmrte)** (52.3%), which we cannot explain intuitively since we don't see how nominal wage is related to crime-rate. We could rationalize this as a higher wage implying higher income inequality and thus a higher crime-rate but this explanation is stretched; therefore, we drop this variable. Next we have **log(density)** which, as we saw in the EDA section, has the highest correlation with **log(crmrte)** among all variables (67.7%). This makes sense as an area with a higher concentration of people offers exponentially more opportunities for criminal offenses. Furthermore, we can see from the loading plot of the first two principal components of the three variables that the angles between the vectors of **prbarr** and **log(density)** and **prbconv** and **log(density)** are quite large, implying that **log(density)** is fairly uncorrelated with the two other existing covariates, which means this shold be a good uncorrelated combination. Therefore we add **log(density)** to our base model which increases our adjusted-R^2 from 46.8% to 62.5% and decreases our AIC from 97.56 to 68.04.






```
pctmin80       west log(polpc)
    0.56      -0.46       0.26
```

Moving to the next iteration, we find the three most correlated variables with the new residuals are **pctmin80** (55.7%), **west** (-46.2%) and **log(polpc)** (26.3%). **pctmin80** has a (23.3%) correlation with **log(crmrte)** while **west** has a -41.4% correlation with **log(crmrte)**. The correlation of **pctmin80** is plausible as areas with higher proportions of minorities may have more difficultly integrating with society and resort to crime. The negative correlation with **west** is also sensible since we note from our EDA that crime-rate seems to be higher in the central and eastern regions. We hypothesize that once we add one of these two variables, the other's correlation with the residuals will drop due to the the high correlation between **pctmin80** and **west**. The correlation of **pctmin80** with the existing covariates are low and range from -17.7% to 6.2%, while the correlation of **west** with the existing covariates are also low and range from -10.7% to 16.8%. The univariate fits above show a strong linear relationship for both **pctmin80** and **west** with the residuals, but particularly for **west**, where we can see that for **west** equal to 1, all the residuals are either near 0 or below 0 with no residuals above around 0.2. There definitely seems to be some benefit in adding this variable. For this reason, as well as **west**'s stronger correlation with **log(crmrte)** than **pctmin80**, and since **west** and **pctmin80** have similar correlations with existing covariates, it is added to the model. Adding this variable increases our adjusted-R^2 from 62.5% to 70.8% and decreases our AIC from 68.04 to 47.51.




```
central log(polpc)   pctmin80
  -0.35       0.33       0.28
```

Moving to the next iteration, we find the three most correlated variables with the new residuals are **central** (-35.3%), **log(polpc)** (33.2%) and **pctmin80** (27.8%). We can see the univariate plots above with the residuals. The slopes are in line with the correlations and we see a particularly strong linear fit with **log(polpc)**. We first notice that the correlation of **pctmin80** with the residuals dropped from 55.7% to 27.8%, as expected, from adding **west** to the model. Because of **pctmin80**'s strong correlation with **west** (-63.4%), we do not consider it to avoid multicollinearity. Furthermore, we note that **log(polpc)** has a 54.5% correlation with **log(crmrte)**. As mentioned in our model 1 building process, we expected a negative relationship with **log(crmrte)** but because we only have a snapshot of the data from 1987 and not from previous years, we cannot assess its causation. We could assume **polpc** is proportional to crime-rate but we do not feel comfortable making this assumption; therefore we drop **log(polpc)**. This leaves us

with **central** which has a 15.9% correlation with **log(crmrte)**. The signs makes sense as we know from our EDA, the western region has a lower crime-rate than the central and eastern regions, with the eastern region having a crime-rate more or less in-line the central region. Moving to the correlations of **central** with the existing covariates, we note some covariation with correlations ranging from -43.3% to 35.1%. However, we notice that when we add this covariate to the model, the VIF scores range from 1.17 to 1.47, which is nothing to worry about and therefore, we add this variable to our model. This increases our adjusted-R^2 from 70.8% to 75.9% and decreases our AIC from 47.51 to 32.23.

log(polpc)	wfed	wtuc	pctmin80	wmfg
0.32	0.22	0.18	0.18	0.16

Now when we look at the variables most correlated with our new residuals, we see that the five most correlated variables are **log(polpc)**, **wfed**, **wtuc**, **pctmin80** and **wmfg**. As mentioned before, we do not feel comfortable adding wage variables due to their positive correlation with **log(crmrte)**, which we cannot explain. Furthermore, **pctmin80** is -63.4% correlated with **west** and therefore do not feel comfortable adding this variable; and lastly, since the correlation of **log(polpc)** is positive with **log(crmrte)**, and we do not feel comfortable with this positive relation, we drop this variable as well. Therefore we end our iterative covariate search and stick with the model composed of **prbarr**, **prbconv**, **log(density)**, **west** and **central**. The model we fit is:

$$log(\text{crmrte}_i) = \beta_0 + \beta_{\text{prbarr}} \cdot \text{prbarr}_i + \beta_{\text{prbconv}} \cdot \text{prbconv}_i + \beta_{log(\text{density})} \cdot log(\text{density}_i) + \beta_{\text{west}} \cdot \text{west}_i + \beta_{\text{central}} \cdot \text{central}_i + u_i$$

	Dependent variable:	
	log(crmrte)	
	Unadjusted SE	HC-Adjusted SE
	(1)	(2)
prbarr	−0.012***	−0.012***
	(0.002)	(0.003)
prbconv	−0.007***	−0.007***
	(0.001)	(0.002)
log(density)	0.387***	0.387***
	(0.047)	(0.051)
west	−0.511***	−0.511***
	(0.077)	(0.071)
central	−0.303***	−0.303***
	(0.072)	(0.074)
Constant	−2.609***	−2.609***
	(0.117)	(0.173)
Observations	90	90
R^2	0.759	0.759
Adjusted R^2	0.745	0.745
Residual Std. Error (df = 84)	0.277	0.277
F Statistic (df = 5; 84)	52.976***	52.976***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Our final model (depicted on the next page) explains (R^2) 75.9% of the variation in **log(crmrte)**. The adjusted-R^2 is 74.5% and the AIC is 32.23. The way we interpret this model is as follows: (**prbarr**) for an increase of 1 in **prbarr** (which is scaled by 100%), *ceteris paribus*, crime-rate decreases by 1.18%; (**prbconv**) for an increase of 1 in **prbconv** (which is scaled by 100%), *ceteris paribus*, crime-rate decreases by 0.71%; (**log(density)**) for a 1%

increase in **density**, *ceteris paribus*, crime-rate increases by 0.39%; (**west**) when the county is in the western region, *ceteris paribus*, crime-rate is 40.00% lower than in the eastern region; and (**central**) when the county is in the central region, *ceteris paribus*, crime-rate is 26.11% lower than the eastern region. The signs of our five covariates are in-line with what we expect them to be and that they are all statistically significant at the 0.1% significance level regardless if we use the unadjusted or HC-adjusted White standard errors.

To ensure that our model 2 has statistical significance over our base model 1, we test the joint hypothesis that

$$H_0 : \beta_{log(density)} = 0, \beta_{west} = 0, \beta_{central} = 0$$

$$H_A : \text{atleast one coefficient is not equal to 0 .}$$

We do this by comparing the RSS from our base model and our model 2 using the Wald test that follows an F-distribution. With this test we reject the null hypothesis (using HC-adjusted SE) that $\beta_{log(density)}$, β_{west} and $\beta_{central}$ are jointly equal to 0 at the 0.1% significance level with an F-statistic of 37.95 and a very small p-value close to 0%. This gives us comfort that the variables we added increased statistical significance to the base model.

Our two key variables, **prbarr** and **prbconv**, are still statistically significant after adding the new covariates, however, we notice the slope coefficients' magnitudes have been reduced from -0.022 to -0.012 for **prbarr** and from -0.012 to -0.007 for **prbconv**. Despite this, we can still see evidence that local government can reduce crime by increasing the risk to potential criminals from committing a crime by increasing the probability of arrest and probability of conviction. As mentioned in model 1, this can be achieved by (1) prioritizing local law enforcement to focus on making arrests (↑ **prbarr**); (2) implement more robust procedures for evidence collection (↑ **prbconv**); and (3) speed up court process for convictions (↑ **prbconv**). For local political campaigns in North Carolina looking to reduce crime as a campaign promise, this gives evidence on how they can achieve this.

As we did in model 1, we are interested in testing whether $\beta_{prbarr} > \beta_{prbconv}$. We do this with the two-sided hypothesis

$$H_0 : \beta_{prbarr} - \beta_{prbconv} = 0 \text{ and } H_A : \beta_{prbarr} - \beta_{prbconv} \neq 0 .$$

Again, we choose a two-sided test here since it is harder to reject the null hypothesis, but we note that $\hat{\beta}_{prbarr} > \hat{\beta}_{prbconv}$ and will thus make a conclusion based on this inequality (if we reject the two-sided hypothesis, we will also reject the one-sided hypothesis that $\beta_{prbarr} > \beta_{prbconv}$. We first define

$$\theta_1 = \beta_{prbarr} - \beta_{prbconv}$$

and re-write our population model as

$$\log(crmrte) = \beta_0 + \theta_1 prbarr + \beta_2(prbarr + prbconv) + \beta_3 log(density) + \beta_4 west + \beta_5 central + u .$$

When we run this regression, we reject the null hypothesis that $\theta_1 = 0$ at the 10% significance level as the t-statistic (using HC-adjusted SE) for the slope coefficient θ_1 is -1.75 and the p-value is 8.40%. This implies there is marginal statistical significance that the two coefficients are statistically different from each other and that reducing **prbarr** has a stronger effect on reducing crime than **prbconv**.

Even though from a statistical perspective, these two factors are marginally significant, we are interested in whether they are practical. Looking at **prbarr**, we can see a 1.18% reduction in crime-rate for a 1% increase in **prbarr** is reasonably important. Likewise, a 0.71% reduction in crime-rate for a 1% increase in **prbconv** is also fairly important. We see the magnitudes of these two coefficients slowly converging together going from our base model to this model. In any case, both are practical from the standpoint of reducing crime and thus of interest to local government.

Lastly, we believe there is some benefit for political campaigns that can be gleaned from the coefficients for **west** and **central** dummy variables. Since we have three regions (west, central, east), when both **west** and **central** are equal to 0, the model gives the crime-rate for the eastern region. In this case the crime-rate is 40.00% lower in the west than in the east and the crime-rate is 26.11% lower in the center than in the east. This implies that crime-rate is higher in the east than in the west and center. Therefore, for any action, targeting the reduction of crime in

the east is likely to have higher impact on the local constituents and electorate than in the center or west. The next question we ask is whether there is a statistically significant difference between the west (**west**) and the center (**central**). As above, we test this by defining $\theta_2 = \beta_{west} - \beta_{central}$, writing our hypothesis as

$$H_0 : \theta_2 = 0 \text{ vs. } H_A : \theta_2 \neq 0 \ ,$$

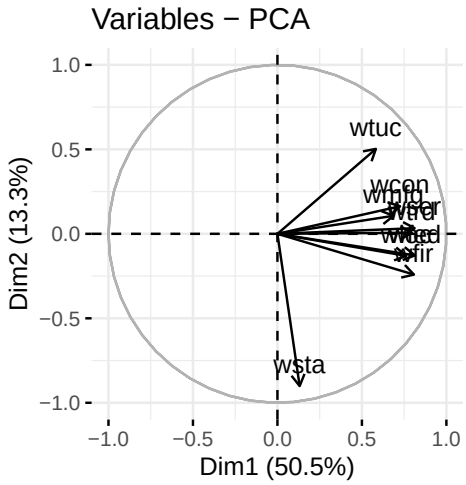
and re-writing our population model as

$$\log(crmrte) = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + \beta_3 log(density) + \theta_2 west + \beta_5 (west + central) + u \ .$$

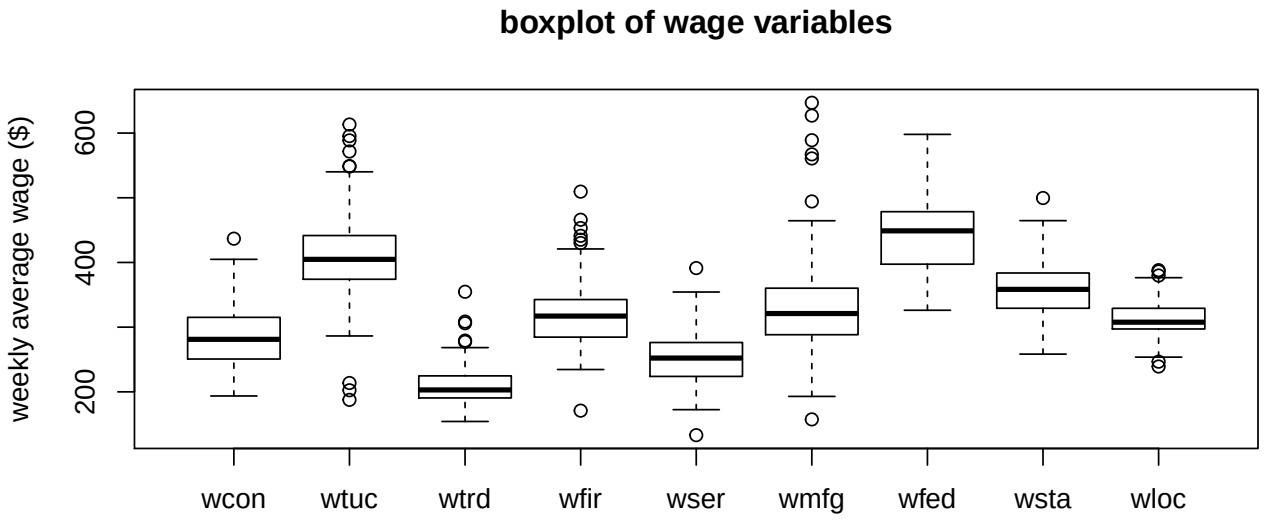
Again, we choose a two-sided test here since it is harder to reject the null hypothesis, but we note that $\hat{\beta}_{west} < \hat{\beta}_{central}$ and will thus make a conclusion based on this inequality. When we run this regression, we reject the null hypothesis at the 1% significance level with a t-statistic (using HC-adjusted SE) for the slope coefficient θ_2 of -2.82 and a p-value is 0.59%. This implies there is high statistical significance that the two coefficients are statistically different from each other and that there is evidence there is less crime in the west than in the central region. Therefore, a political campaign and any associated action in the eastern region of North Carolina is likely to be more effective than in the central or western region.

Model 3

We now move on to our third model, which we use to test the robustness of our models 1 & 2. We add all the remaining variables; however before we do this, we have to solve an issue. As discovered in the EDA section, we notice strong positive correlations among the nine wage variables. This can be seen in the loading plot below of the first two principal components of these nine variables.



Including all of these covariates in a model would inevitably result in quite a lot of multicollinearity. Therefore, we consider creating an aggregate wage variable by averaging across the nine wages. Looking at the boxplot below; we notice the underlying distributions of the different wages vary widely, which is an issue: wages with more substantial variances will drown out the variation of wages with smaller variances.



We solve this problem by first standardizing all wage variables to have zero-mean and unit-variance and then compute the mean across the standarized wages. This new variable, **aggwage**, remains positively correlated with **log(crmrte)** (47.5%). Since this is a standardized variable, the interpretation changes: for a 1 standard deviation increase in aggregate wage, crime-rate will change by $[(\exp(\beta_{aggwage}) - 1) \times 100]\%$.

```
wage <- c('wcon', 'wtuc', 'wtrd', 'wfir', 'wser', 'wmfg', 'wfed', 'wsta', 'wloc')
raw_data$aggwage <- apply(scale(raw_data[, wage]), 1, mean)
```

With this, we run our regression with the wage variables replaced by **aggwage**.

Our final model explains (R^2) 81.4% of the variation. The adjusted-R^2 is 77.9% and the AIC is 26.94. The way we interpret this model is as follows: (**prbarr**) for a 1% increase in **prbarr**, *ceteris paribus*, crime-rate decreases by 0.01%; (**prbconv**) for an increase of 1 in **prbconv** (which is scaled by 100), *ceteris paribus*, crime-rate decreases by 0.65%; (**log(density)**) for a 1% increase in **density**, *ceteris paribus*, crime-rate increases by 0.28%; (**west**) when the county is in the western region, *ceteris paribus*, crime-rate is 30.10% lower than in the eastern region; (**central**) when the county is in the central region, *ceteris paribus*, crime-rate is 21.59% lower than the eastern region; (**prbpris**) for an increase of 1 in **prbpris** (which is scaled by 100), *ceteris paribus*, crime-rate increases by 0.22%; (**avgsen**) for an increase of 1 year in **avgsen**, *ceteris paribus*, crime-rate decreases by 0.41%; (**log(polpc)**) for a 1% increase in **polpc**, *ceteris paribus*, crime-rate increases by 0.36%; (**log(taxpc)**) for a 1% increase in **taxpc**, *ceteris paribus*, crime-rate decreases by 0.04%; (**urban**) when the county is in an urban area, *ceteris paribus*, crime-rate is 2.48% lower than in rural areas; (**pctmin80**) for an increase of 1 in **pctmin80** (which is scaled by 100), *ceteris paribus*, crime-rate increases by 0.51%; (**aggwage**) for an increase of 1 standard deviation in **aggwage**, *ceteris paribus*, crime-rate increases by 9.82%; (**log(mix)**) for a 1% increase in **mix**, *ceteris paribus*, crime-rate increases by 0.01%; and (**pctymle**) for a 1% increase in **pctymle**, *ceteris paribus*, crime-rate decreases by 0.00%. It is reasurring to see the signs for our three main covariates in model 2 (**prbarr**, **prbconv**, **log(density)**, **west** and **central**) all remain the same and their statistical significance remain at the 5% significance level (using HC-adjusted White SE), while none of the new covariates added in model 3 are significant.

Since we use model 3 as a robustness check on model 2, we are interested in whether all the additional covariates in model three have slope coefficients that are jointly equal to 0.

$$H_0 : \beta_{prbpris} = \beta_{avgsen} = \beta_{log(polpc)} = \beta_{log(taxpc)} = \beta_{urban} = \beta_{pctmin80} = \beta_{aggwage} = \beta_{log(mix)} = \beta_{pctymle} = 0$$

$$H_A : \text{atleast one coefficient is not equal to 0}$$

We do this by comparing the RSS from our model 2 and our model 3 using the Wald test that follows an F-distribution. We reject the null hypothesis (using HC-adjusted SE) of all the additional slope coefficients being jointly equal to 0

with an F-statistic of 1.54 and a p-value of 15.08%. This makes us confident in our model 2 covariate selection since the additional regressors do not jointly add statistical significance to the model.

In line with our narrative for model 2, we check whether $\beta_{prbarr} > \beta_{prbconv}$ and whether $\beta_{west} < \beta_{central}$ still holds for our model 3. Again, we use a two-sided test here since it is harder to reject than a one-sided test. We expect the strength of these tests to be reduced due to multicollinearity of all the additional covariates increasing the standard errors of our coefficients. We first set $\theta_1 = \beta_{prbarr} - \beta_{prbconv}$ and define the hypothesis test as $H_0 : \theta_1 = 0$ vs. $H_A : \theta_1 \neq 0$. After we re-write our population model to account for this new definition, we run the regression and fail to reject the null hypothesis with a test-statistic (using HC-adjusted SE) for the slope coefficient θ_1 of -1.54 and a p-value of 12.66%. Likewise, by setting $\theta_2 = \beta_{west} - \beta_{central}$, defining the hypothesis test as $H_0 : \theta_2 = 0$ vs. $H_A : \theta_2 \neq 0$, re-writing the population model and running the regression, we fail to reject the null hypothesis of $\theta_2 = 0$ with a test-statistic for the slope coefficient for θ_2 of -1.12 and a p-value of 26.72%. Unfortunately our statistically significant differences vanish with our all-inclusive model 3 but this shouldn't be of too much concern since as we mentioned earlier, there is considerable multicollinearity added to the model to increase the standard errors which makes a lot of these tests lose their strength.

Lastly, it is interesting to see that **prbpris** does not affect the crime rate. We can view **prbarr**, **prbconv** and **prbpris** as three different levels of risk to criminals, all conditional on a state occuring. For example, **prbarr** is the probability of getting arrested conditional on committing a crime, **prbconv** is the probability of being convicted conditional on being arrested, and **prbpris** is the probability of going to prison conditional on being convicted. These all present risks to someone looking to commit a crime. The probability of arrest is the first risk, followed by the probability of conviction and lastly, the probability of going to prison. It makes sense that **prbarr** tends to have a more substantial effect on reducing crime-rate than **prbconv**, and **prbconv** has a more substantial impact than **prbpris**. **prbpris** is so far away that potential criminals may not even consider this since they first need to be caught and then once they have been caught, they have to be convicted. Therefore, this leads us to recommend that local government should focus less on putting people in prison and more on catching and convicting them, which seems to be a stronger deterrent to crime. A second added benefit of this is economics. It costs on average around $31,000 to incarcerate someone for one year. Reducing this, will most likely allow government to spend taxpayer's money more effectively in other areas.

	Dependent variable:	
	log(crmrte)	
	Unadjusted SE	HC-Adjusted SE
	(1)	(2)
prbarr	-0.012^{***}	-0.012^{**}
	(0.003)	(0.004)
prbconv	-0.007^{***}	-0.007^{*}
	(0.001)	(0.003)
log(density)	0.278^{***}	0.278^{**}
	(0.074)	(0.095)
west	-0.358^{**}	-0.358^{*}
	(0.120)	(0.175)
central	-0.243^{**}	-0.243^{*}
	(0.082)	(0.114)
prbpris	0.002	0.002
	(0.004)	(0.006)
avgsen	−0.004	−0.004
	(0.012)	(0.015)
log(polpc)	0.360^{**}	0.360
	(0.130)	(0.238)
log(taxpc)	−0.044	−0.044
	(0.152)	(0.232)
urban	−0.025	−0.025
	(0.151)	(0.146)
pctmin80	0.005	0.005
	(0.003)	(0.004)
aggwage	0.094	0.094
	(0.064)	(0.084)
log(mix)	0.013	0.013
	(0.069)	(0.091)
pctymle	0.005	0.005
	(0.014)	(0.017)
Constant	−0.375	−0.375
	(1.325)	(1.974)
Observations	90	90
R^2	0.814	0.814
Adjusted R^2	0.779	0.779
Residual Std. Error (df = 75)	0.258	0.258
F Statistic (df = 14; 75)	23.464^{***}	23.464^{***}
Note:	*p<0.05; **p<0.01; ***p<0.001	

Summary Regression Table

Looking at the juxtaposition of our three models, we can see that our two key variables - **prbarr** and **prbconv** - have a reasonably constant factor loading across the three models. The factor loadings from model 1 to model 2 do change a bit with the magnitudes being reduced; however, going from model 2 to model 3, the coefficients do not vary significantly. This suggests that our key variable coefficients are fairly robust and provides greater confidence in the key effects on reducing crime-rate. The practical significance of these two variables remains stable for all three models.

Looking at **log(density)**, we can see that by adding all remaining variables in model 3 does reduce its slope coefficient from 0.387 to 0.278. Furthermore, the slope coefficient for **west** is also reduced by quite a bit from model 2 to model 3 from -0.511 to -0.358, while **central** remains fairly constant moving from -0.303 to -0.243.

CLM Assumptions

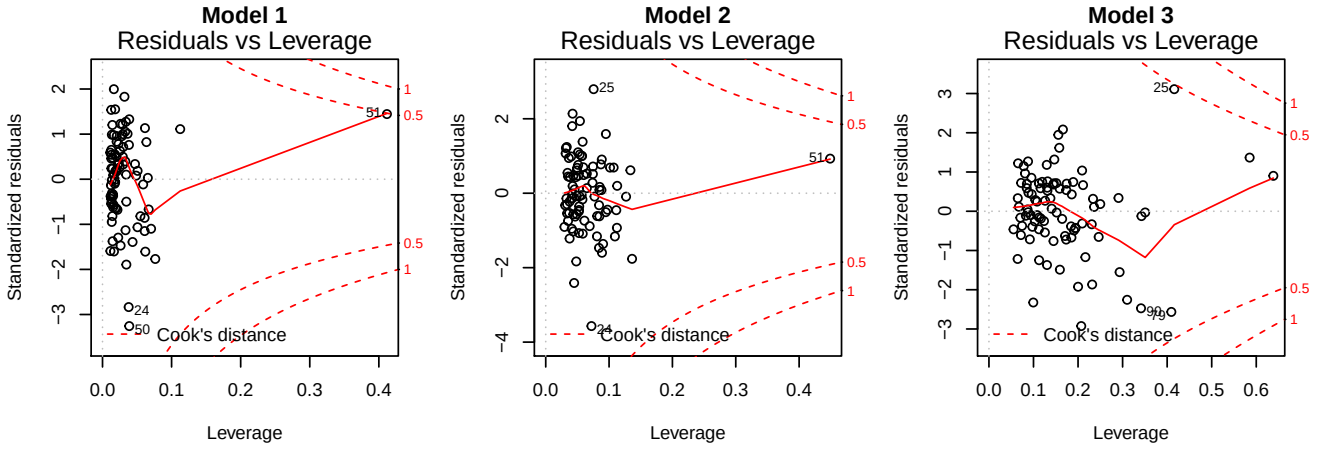
In this section we will examine the six CLM (**C**lassical **L**inear **M**odel) assumptions for the three models. Our main focus will be model 2 as this is the model we find of most importance.

CLM1: Linearity in Parameters

All three linear models are linear in their parameters as each parameter is either a constant or multiplied by an independent variable.

CLM2: Random Sampling

We are using a single cross-section of data of a multi-year panel. Panel data is known to have limitations; however, as this is a multi-year panel, we assume that it is a balanced panel. Also, as the unit of measurement is a county, even though some crime characteristics in one county may influence the crime characteristics of another county for the purposes of this report we assume the individual counties are independent and identically distributed (i.i.d.). However, due to the outlined limitations of our data, our ability to make causal inferences about the population is also limited. Therefore we focus on the descriptive statistics analyzing the effects of certain variables on crime rates. Therefore, we do not see evidence of non-random sampling and believe having 90% of the state's counties in our dataset is a good representation of the population.



When we look at potential outliers in our models, we don't see any evidence of this with no residuals falling outside the 0.5 and 1.0 Cook's distance boundary. We can, however, identify the **log(crmrte)** outlier we defined earlier in our report as record #51. We can see this point in all three plots, and despite this point having leverage, it has no influence. Particularly for model 2, there does not seem to be any points near the 0.5 boundary, which is reassuring. In any case, there does not seem to be evidence of non-random data from these plots.

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
prbarr	−0.022***	−0.012***	−0.012**
	(0.005)	(0.003)	(0.004)
prbconv	−0.012***	−0.007***	−0.007*
	(0.002)	(0.002)	(0.003)
log(density)		0.387***	0.278**
		(0.051)	(0.095)
west		−0.511***	−0.358*
		(0.071)	(0.175)
central		−0.303***	−0.243*
		(0.074)	(0.114)
prbpris			0.002
			(0.006)
avgsen			−0.004
			(0.015)
log(polpc)			0.360
			(0.238)
log(taxpc)			−0.044
			(0.232)
urban			−0.025
			(0.146)
pctmin80			0.005
			(0.004)
aggwage			0.094
			(0.084)
log(mix)			0.013
			(0.091)
pctymle			0.005
			(0.017)
Constant	−2.316***	−2.609***	−0.375
	(0.219)	(0.173)	(1.974)
Observations	90	90	90
R^2	0.468	0.759	0.814
Adjusted R^2	0.456	0.745	0.779
Residual Std. Error	0.405 (df = 87)	0.277 (df = 84)	0.258 (df = 75)
F Statistic	38.289*** (df = 2; 87)	52.976*** (df = 5; 84)	23.464*** (df = 14; 75)
Note:			*p<0.05; **p<0.01; ***p<0.001

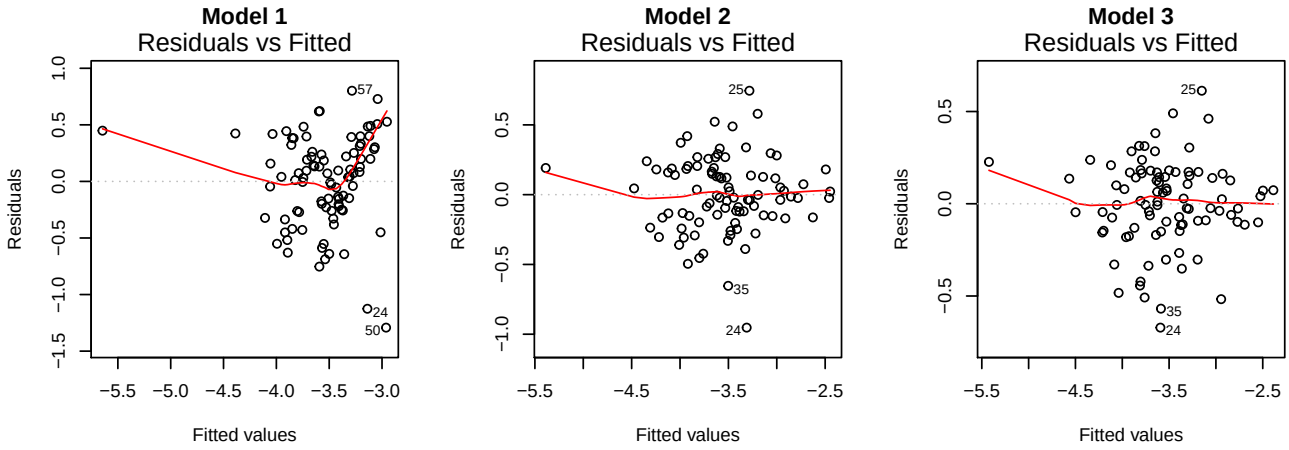
CLM3: No Perfect Multicollinearity

We do not see any perfect multicollinearity for the three models. Below are the VIF values for model 2.

```
     prbarr      prbconv `log(density)`        west     central
   1.245670     1.166213     1.468995     1.270595     1.400409
```

The VIF values for model 2 range from 1.01 to 1.01 and are well below the thresholds of 5 and 10, therefore, we have no strong or perfect multicollinearity. The VIF value for model 1 is 1.01 which also means we do not have any evidence of perfect multicollinearity. For model 3 the VIF values range from 1.10 to 4.24. Here we can see the highest VIF value (4.24), which is for **log(density)**, is high but still below the thresholds of 5 and 10, and therefore, there is no evidence of perfect or problematic multicollinearity. This VIF value is expected as **log(density)** has strong correlations with **aggwage** (70.8%), **urban** (65.8%) and **log(polpc)** (48.6%). High wages and a high density of police officers is expected in a high density area; and urban by definition is usually a place with a high density of people.

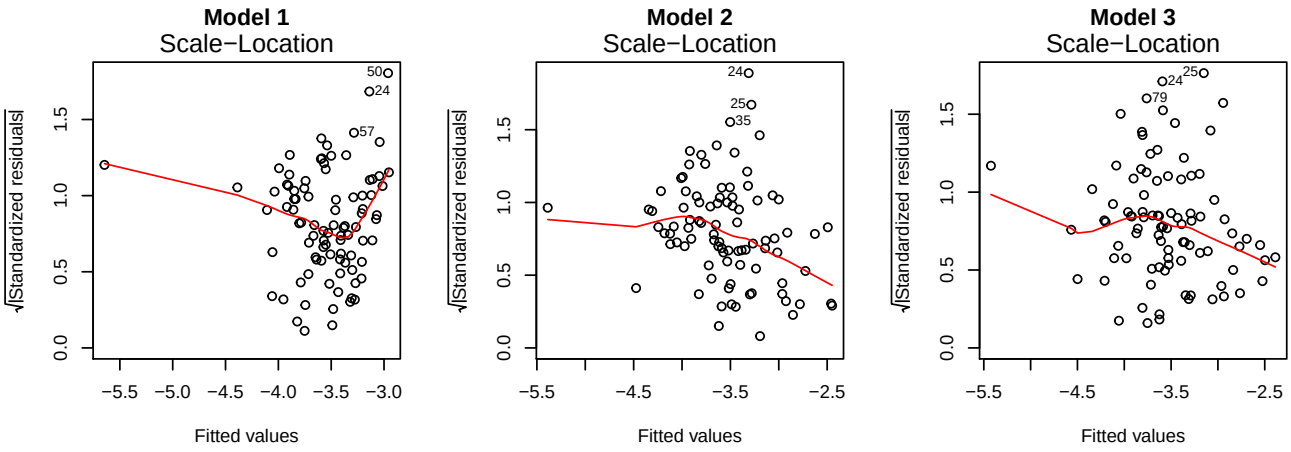
CLM4: Zero-Conditional Mean



We first notice that on the left side of the plot, we have a downward sloping LOESS curve, but this is due to one (outlier) point pulling the curve up in this area. Because we have too few data points on the far left side of the graph, the LOESS curve could be randomly high here because of this one fitted value. This fitted value is for the record with the outlier in **log(crmrte)** that we identified in the ealier section. For these plot diagnostics, we focus on the LOESS curve from fitted values to the right of fitted values with value -4.5. For model 1 we see zero-conditional mean is violated as there is an upward sloping curve on the right side of the plot. For model 2 & 3 we do not see evidence of zero-conditional mean being violated with both curves reasonably flat.

Since we see a violation in this assumption for model 1, we can still claim exogeneity since we do not see any reason for any of the variables to be endogenous.

CLM5: Homoscedasticity

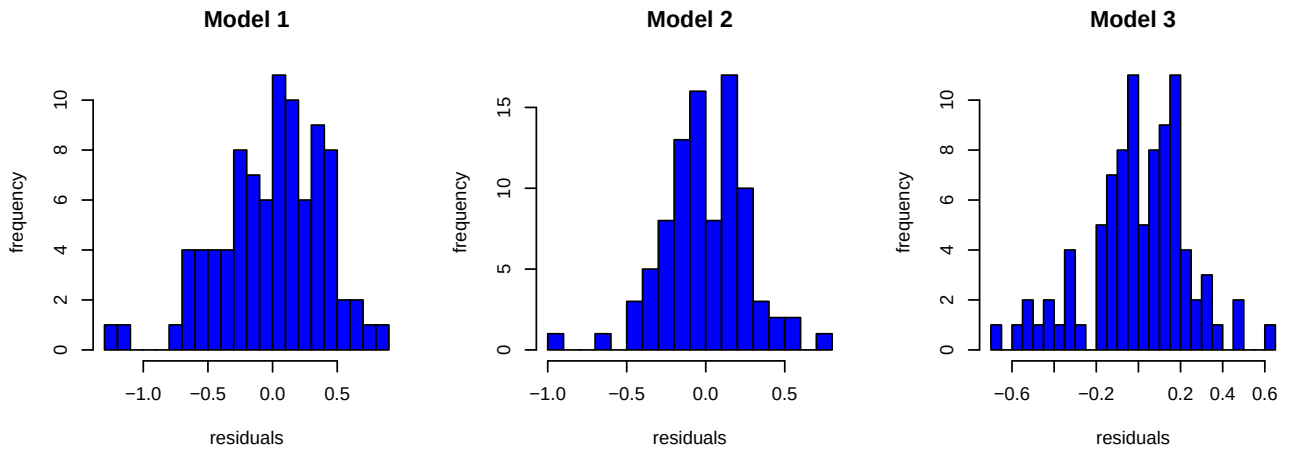


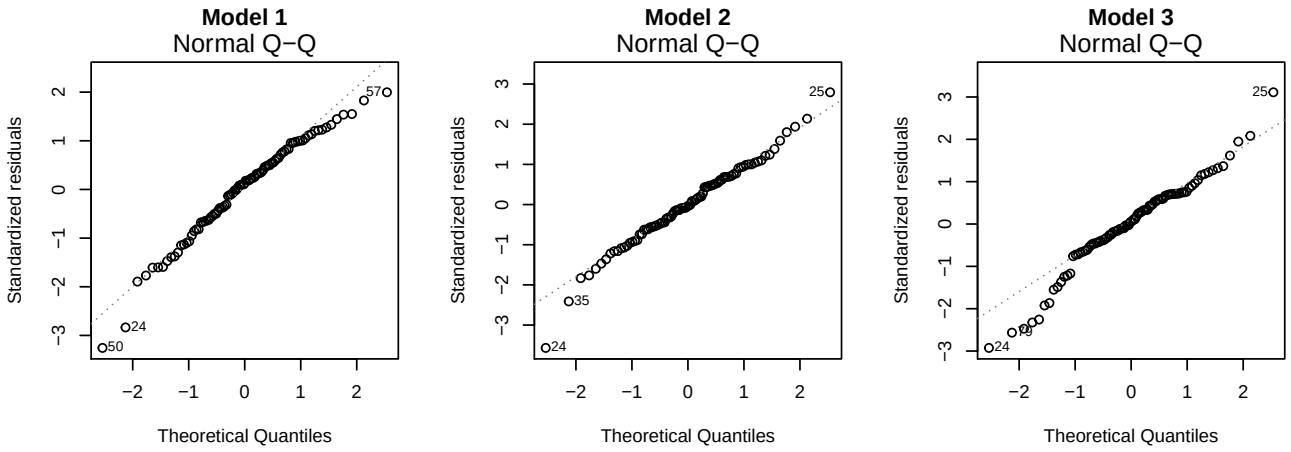
Similar to CLM4, we notice on the left side of the plot one point that could randomly pull the curve up or down. For these plot diagnostics, we focus on the LOESS curve from fitted values to the right of fitted values with value -4.5. For model 1, we see evidence of heteroscedasticity with a sharp upward sloping LOESS curve on the right side of the plot. For models 2 & 3 we also see evidence of heteroscedasticity as the LOESS curves are sloping downward.

This is confirmed with the Breush-Pagan test rejecting the null hypothesis of homoskedacticity for all three models at the 10% significance level with test statistics of 4.68 (model 1), 14.84 (model 2) and 32.63 (model 3) and p-values of 9.618% (model 1), 1.108% (model 2) and 0.325% (model 3).

Since we see evidence of heteroscedasticity, we need to adjust the standard errors of our slope coefficients for our three models using homoscedastic-adjusted (White) standard errors. These errors tend to be more conservative and will, therefore, reduce the significance (t-statistics) of the slope coefficients.

CLM6: Errors are Normally Distributed





For the three models, we can see the histograms of the errors have a shape that simulates a normal distribution. It seems that the errors from model 2 are most in-line with the normal distribution.

For model 1 we can see a diagonal line in the QQ-plot with a tail on the upper end and two outliers on the bottom end. Therefore, we see some evidence of non-normality here. When we check the Shapiro-Wilk test, we see it rejecting the null hypothesis of normality at the 10% significance level with a test statistic of 0.974 and a p-value of 6.404%.

For model 2, we see a diagonal line in the QQ-plot with two outliers in the bottom end. Therefore, we don't see any evidence of non-normality here. This is confirmed with the Shapiro-Wilk test which fails to reject the null hypothesis of normality with a test statistic of 0.985 and a p-value of 36.559%.

For model 3, we see a diagonal line in the QQ-plot with a heavy tail in the bottom end. Therefore, we see evidence of non-normality here. This is confirmed with the Shapiro-Wilk test which rejects the null hypothesis of normality at the 10% significance level with a test statistic of 0.974 and a p-value of 7.263%.

CLM Conclusion

For model 1, CLM assumption of zero-conditional mean (CLM4), homoscedasticity (CLM5) and normality (CLM6) were violated. Even though we do not have a zero-conditional mean, we claim exogeneity since we do not see any reason to believe there are any endogenous variables. This means that model 1's estimators are consistent (for large samples, $\lim_{n\to\infty} \hat{\beta} = \beta$). Furthermore, we do not have evidence that our sample is non-random, and because our sample is large (>30) we can still use the CLT to assume our sampling distributions follow a normal distribution. Therefore, our conclusions from model 1 are valid. The violation of CLM5 (homoscedasticity) implies that there might be some omitted variables that are showing up in the residuals. This was addressed in model 2, which has less drastic heteroscedasticity. Furthermore, for this model, instead of adding variables, we rely on HC-adjusted White standard errors (which are more conservative) for our test statistics.

For model 2, only homoscedasticity (CLM5) was violated. There is clearly an improvement in the heteroscedasticity going from model 1 to model 2. This was most likely due to omitted variables being present in model 1, which were partially addressed in the model 2. Instead of looking for omitted variables though, we rely on HC-adjusted White standard errors for our test statistics. Since only CLM5 is violated, we can say that our model estimates are unbiased and consistent. Once we HC-adjust (White) our standard errors though, we can make inferences on our data. We can always resort to the CLT here - since our sample is large (>30) and there is no evidence of our dataset being a non-random sample - and with this, we can make inferences.

For model 3, we see a violation in homoscedasticity (CLM5) and normality of errors (CLM6). The heteroscedasticity is improved in this model over model 2 as we added new variables that potentially limited any omitted variables. Nevertheless, as mentioned in our next section, there are some omitted variables not covered in our dataset that could be still affecting the heterogeneity of errors. With these violations, we can say the estimators of this model are unbiased and consistent. We again resort to the CLT here - since our sample is large (>30) and there is no evidence of our dataset being a non-random sample - and with this, we can make inferences.

5.0 The Omitted Variables Discussion

Unemployment

Numerous studies examine the causal impact of unemployment on crime rate, yet our models do not include the unemployment variable.

An unemployed person is more likely to engage in criminal activity. We know that unemployment is generally higher among young adults. In our third model, this would indicate that the **pctymle** may include a noticeable amount of the unemployment bias.

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{pctymle} + \beta_2 \cdot \text{unemployment} + u$$

$$\text{unemployment} = \alpha_0 + \alpha_1 \cdot \text{pctymle} + u$$

We expect both the correlation between the **pctymle** and unemployment as well as with crime rate and unemployment to be positive, and therefore we estimate that the omitted variable bias is also positive. Equally, the beta for the **pctymle** is positive as such OLS estimates away from zero, gaining statistical significance.

This omitted variable could be proxied by the unemployment rate for each county.

Poverty

Generally, where poverty is prevalent in a community, crime may be seen as an opportunity for less-fortunate people to access goods, they may not be able to afford it. In a way, the prize may outweigh the risk or arrest or conviction.

We expect that unemployment and poverty are closely interrelated. Therefore we would expect that the impact of omitting poverty variable from our models to be consistent with unemployment.

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{unemployment} + \beta_2 \cdot \text{poverty} + u$$

$$\text{poverty} = \alpha_0 + \alpha_1 \cdot \text{unemployment} + u$$

This omitted variable could be proxied by the percent of people with an income below a certain threshold that could be considered poverty in the county.

Substance abuse

An increase in crime rate may be tied to an increase in substance abuse. A person with a substance abuse problem may be unable to support their addiction without crime ($\beta_2 > 0$).

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{density} + \beta_2 \cdot \text{substance\_abuse} + u$$

Higher population density is likely to create greater number of opportunities, increase stress and temptation to turn to addictive substances. We, therefore, hypothesize that any bias resulting from substance abuse is likely to be related to **density** in our model.

$$\text{substance\_abuse} = \alpha_0 + \alpha_1 \cdot \text{density} + u$$

We believe that with higher population density the substance abuse is also higher $\alpha_1 > 0$. Given that the omitted variable bias is positive $\beta_2 \cdot \alpha_1 > 0$, the OLS estimates would over estimate the marginal effect of density on crime rate. Furthermore it would scale the coefficient away from zero gaining statistical significance.

This omitted variable could be proxied by the percent of people in substance abuse treatment programs in the county.

Inequality of Income

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{wage} + \beta_2 \cdot \text{income\_inequality} + u$$

$$\text{income\_inequality} = \alpha_0 + \alpha_1 \cdot \text{wage} + u$$

We believe a positive relation between income inequality and wages $\alpha_1 > 0$. Given that the omitted variable bias is positive $\beta_2 \cdot \alpha_1 > 0$ and the $(\beta_1 > 0)$, the ols estimates would over estimate the marginal effect of wage on crime rate. However the magnitude of wage impact in our third model is relatively low and insignificant.

This omitted variable could be proxied by the Gini Coefficient for each county.

Education

We think that the levels of education and unlawful behaviour may be correlated. More specifically, individuals with lower attained education are more likely to turn to criminal activity to earn their living $(\beta_2 < 0)$.

The closest proxy variable to education would be the wage variables in our third model. Our model shows $(\beta_1 > 0)$ and we expect that higher education would lead to higher wages $(\alpha_1 > 0)$. The resulting bias therefore would be negative $\beta_2 \cdot \alpha_1 < 0$ and this would cause the OLS coefficient of wage to be scaled towards zero losing its statistical significance.

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{wage} + \beta_2 \cdot \text{education} + u$$

$$\text{education} = \alpha_0 + \alpha_1 \cdot \text{wage} + u$$

This omitted variable could be proxied by the literacy rate for each county as well as average years of education attained.

Family Conditions

Neglect, abuse or lack of stability are likely to increase the chance of young people to act out and engage in unlawful activities $(\beta_2 > 0)$. Some studies suggest that convicted criminals have experienced four to five times as many adverse formative events than non-criminal adults.

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{density} + \beta_2 \cdot \text{family\_neglect} + u$$

$$\text{family\_neglect} = \alpha_0 + \alpha_1 \cdot \text{density} + u$$

Any bias related to family neglect is likely to be embedded within our density variable in our main model. Higher population density $(\beta_1 > 0)$ is likely to increase the pace of life which then leads to parents to be either more stressed and have less time to spend with their offspring. This then may lead to abuse or neglect respectively $(\alpha_1 > 0)$. The resulting bias, therefore, would be $\beta_2 \cdot \alpha_1 > 0$ and this would cause the OLS coefficient of density to be scaled away from zero gaining statistical significance.

This omitted variable could be proxied by the divorce rate and percent of single parents for each county.

Mental Health

Mental health may influence individual decision-making ability and morality; it is. Therefore, we hypothesize that elevated levels of mental health problems within a society are likely to contribute to an increased crime rate $(\beta_2 < 0)$.

$$\text{crmrte} = \beta_0 + \beta_1 \cdot \text{density} + \beta_2 \cdot \text{mental\_instability} + u$$

$$\text{mental\_health} = \alpha_0 + \alpha_1 \cdot \text{density} + u$$

In models 1, 2 and 3, we would expect density to be positively related to crime-rate $(\beta_1 > 0)$ variable. We would expect that the increased stress, pace, noise, pollution of living in an area of high population density would contribute

to decreased mental health (depression, anxiety etc.) ($\alpha_1 < 0$). The resulting bias, therefore, would once again be positive $\beta_2 \cdot \alpha_1 > 0$ and this would cause the OLS coefficient of density to be scaled away from zero gaining statistical significance.

This omitted variable could be proxied by the percent of people in a county looking for professional help for mental issues.

Effect on Main Variables

Regarding our main variables **prbarr**, **prbconv** and **west**, we believe these variables are orthogonal to all of the listed omitted variables above, and therefore we expect little bias and thus see their effects as real. This implies that the recommendations we make in the next section are not impacted by the omitted variable bias discussed.

6.0 Conclusion

Following the analysis of the North Carolina county data, we propose that the local political campaigns:

1. Focus on improving probability of arrest, for example - prioritize local law enforcement to focus on making more arrests to increase the probability of arrest. Based on the data, we conclude that this is the most significant deterrent from committing an offence (**prbarr** ↑).
2. Focus on improving probability of conviction, for example - implement stronger procedures on evidence collected by local law enforcement for each arrest to increase the probability of conviction. We recommend this as a way to increase the risk to potential criminals from committing an offence (**prbconv** ↑). Or take action to speed up the court process for convicting criminals. This should increase the number of convictions and as a result, act as a deterrent to potential criminals (**prbconv** ↑).
3. Do not focus on probability of prison sentence as based on our analysis, unlike increasing arrest and conviction rate, an increase in the probability of prison sentence does not seem to have the desired effect on the crime rate. Therefore, we recommend not to focus on policies specifically aimed at increasing the probability of prison (non-significant **prbpris**).
4. The priority of reducing crime as a campaign is likely to have highest impact in the eastern region of North Carolina, followed by the central region and finally the western region. This is because a message of reducing crime should resonate more in areas with higher crime-rates.

Based on these recommendations, we would like to add a few remarks. Regarding recommendation 1, this policy must be correctly executed. Law enforcement should still retain their high standards for making an arrest and not arrest people without sufficient evidence. Arresting too many people who didn't commit a crime can backfire as public opinion of law enforcement would inevitably sour. What we imply with recommendation 1 is that law enforcement should prioritize their time to maximize the chance of making arrests. This can be done by having more police officers on the street at any point in time instead of them being in the office, or having law enforcement in general speed up the process required to make arrests. The same applies to recommendation 3; the process quality should not be put into jeopardy.

The following report did not perform any validations such as K-folds cross-validation on our results. Our aim for this report was to find some preliminary results that we can delve into more deeply in future research. Furthermore, we suggest a new study in which we add proxy variables for the omitted variables mentioned in our earlier study. This would allow us to make stronger inferances and test the robustness of the results of this report.

We hope you will find this report's analysis and recommendations actionable and we look forward to working with you again.

Walekova & Graf Consulting

Mikra Walekova

Dominik Graf