Lab 3: Reducing Crime

Ram Iyer & Daniel Olmstead

16 April 2018

```
# Load relevant libraries
library(tidyverse)
library(car)
library(corrplot)
library(ggExtra)
library(GGally)
library(knitr)
library(kableExtra)
library(reshape2)
library(stargazer)
library(broom)
library(ggfortify)
library(lmtest)
library(sandwich)
```

INTRODUCTION AND APPROACH

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

In generating policy suggestions for a political campaign, our approach is to group the variables into three categories which would imply a particular policy response, in hopes of determining which categories have the most influence on crime, thereby suggesting which policy approach would be most effective. These categories are:

1. Deterrence (ie, "Tough on Crime")

These variables describe policing, crime and punishment, under the philosophy that stricter policing and more punishment leads to less crime. These include **prbarr, prbconv, prbpris, avgsen, mix, and polpc.**

2. Economic (ie, "Prosperity and Peace")

These variables revolve around the argument that crime is a function of economics, and if you can raise the standard of living for the poor, they will have less incentive to commit crime. These include **taxpc, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, and wloc.**

3. Demographic/Structural

These variables focus on the demographic and geographic makeup of the state, and while they may not have a direct policy response (ie, a campaign cannot run on mass redistribution of the population), they can help with targeting policies. These include **density, west, central, urban, pctmin80, pctymle.**

Load the data and format the variables

There were a few issues with the data that we cleaned up on import: removing a duplicate row, pruning out some empty records at the end of the CSV, and recasting *prbconv* as a numeric (rather than a factor).

```
Crime.data <- data.frame(unique(read.csv("crime_v2.csv", header = TRUE)))
Crime.data <- na.omit(Crime.data)
Crime.data$prbconv <- as.numeric(as.character(Crime.data$prbconv))
cols <- sapply(Crime.data,class)
# Variables with their types
kable(t(cols[1:12]),format="latex",booktabs=T);kable(t(cols[11:25]),format="latex",booktabs=T)
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central
integer	integer	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	integer	integer

west	central	urban	pctmin80	wcon	wtuc	wtrd	wfir	wser	wmfg	wfed	wsta
integer	integer	integer	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeri

EDA, SINGLE-VARIABLE ANALYSIS

The dataset consists of county-level crime-rate information from 90 counties in North Carolina in 1987.

county and year

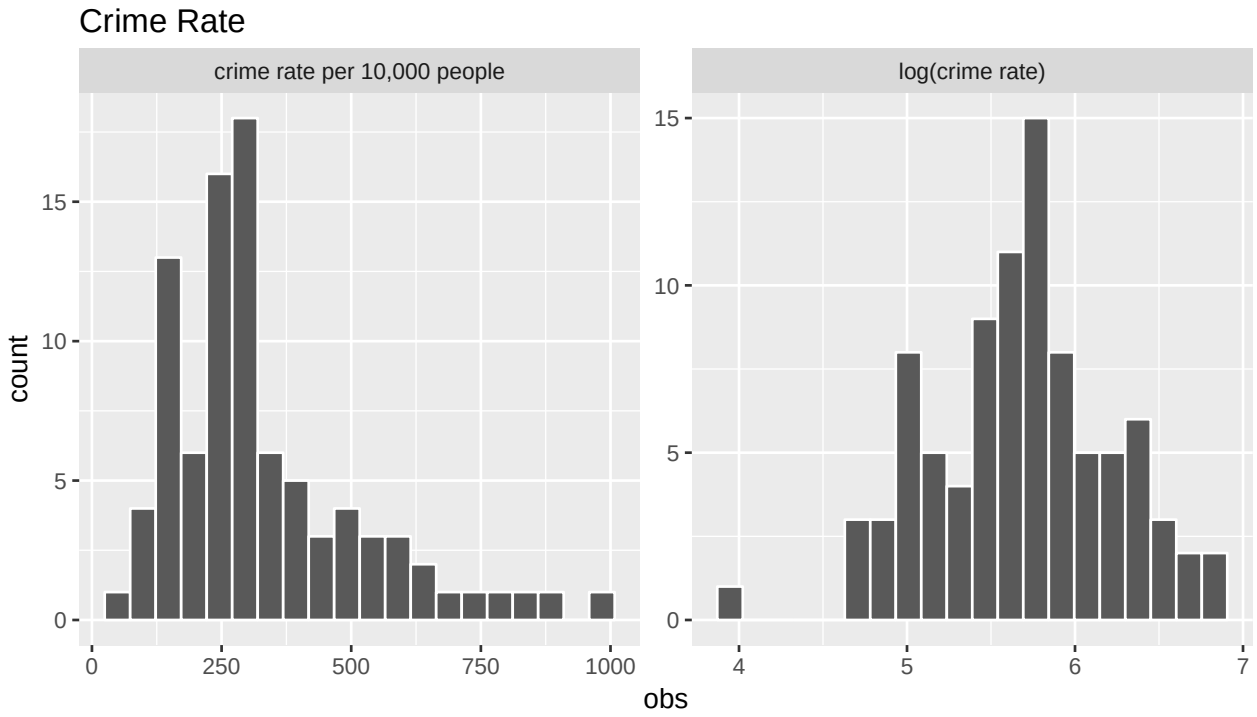
county is simply an integer identifier for each county. All of them are from 1987, so the only value in **year** is '87.'

crmrte

Per-capita crime rate is the dependent variable for our analyses. We convert it from per capita to per 10,000 people for simpler interpretation, and compare the raw numbers to a log transformation.

```
Crime.data <- mutate(Crime.data,crmrte_orig = crmrte)
crdf <- rbind(data.frame(wcat="crime rate per 10,000 people", obs=10000 * Crime.data$crmrte),
              data.frame(wcat="log(crime rate)", obs=log(10000 * Crime.data$crmrte)))

ggplot(crdf, aes(x=obs)) + geom_histogram(bins=20, colour='white') +
  ggtitle("Crime Rate") +
  facet_wrap( ~ wcat, scales="free")
```



Crime rate is largely between 150-300 crimes per ten thousand people, with a large positive skew covering about a half-dozen higher-crime counties. The highest crime rate observed is 1000 in 10000 people, for unfortunate county 119. The average crime rate is 335.1. There is an unusual dip at 200, the shape of which is very similar to the histogram for *avgsen*, as described later. A log transformation, as seen on the right, makes for a more normal distribution and, as a dependent variable, lets us interpret changes in independent variables as predicting a percentage change in crime risk. Therefore we consider the transformed *crmrte*, henceforth referred to as *log.crmrte* for all further analysis.

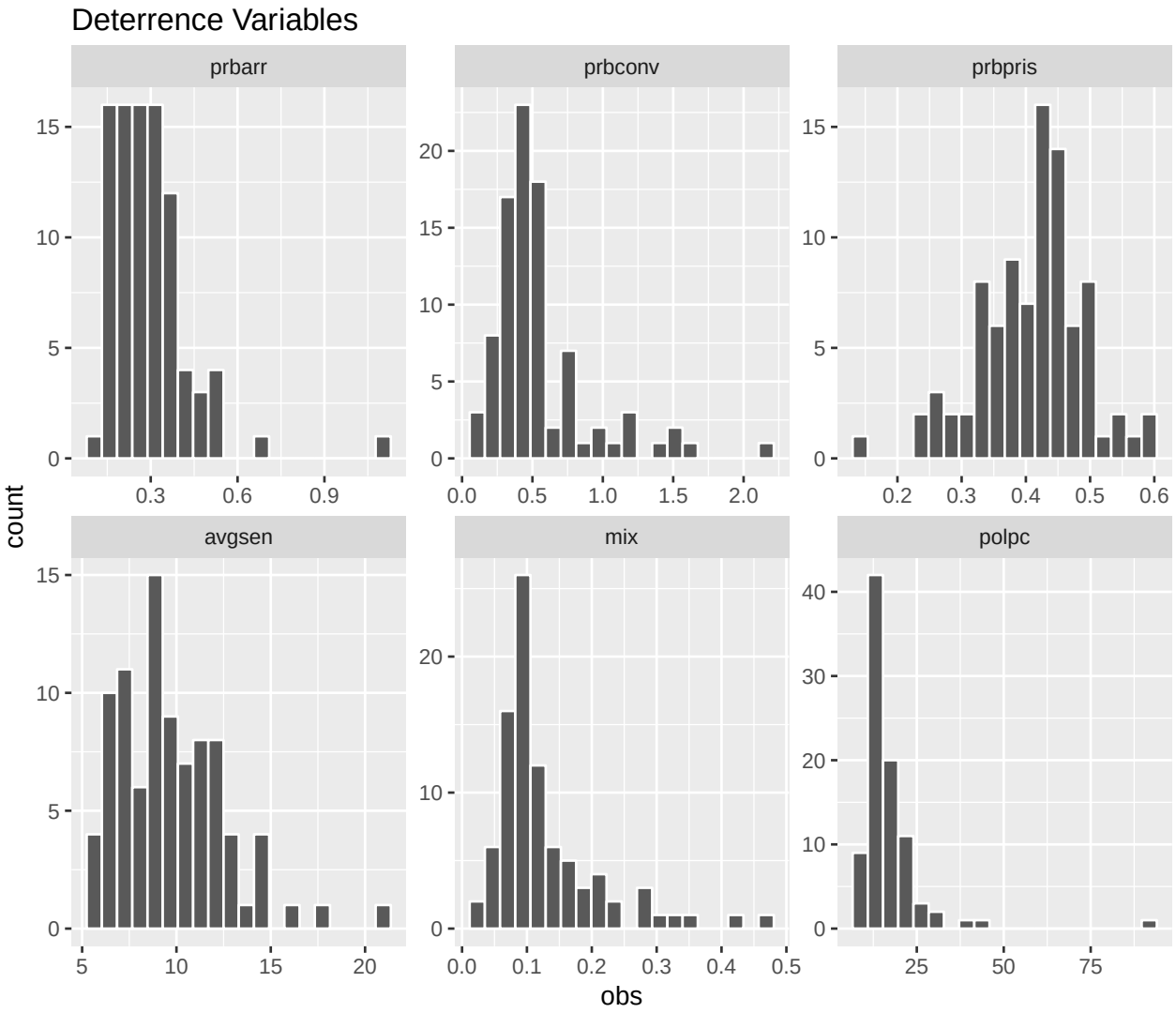
```
Crime.data$log.crmrte <- log(Crime.data$crmrte_orig * 10000)
```

Deterrence variables

```
# translate to police per 10000 people
Crime.data <- mutate(Crime.data,polpc_orig = polpc); Crime.data$polpc <- 10000 * Crime.data$polpc_orig

# Make a master dataframe for deterrence vars
wdf <- rbind(data.frame(wcat="prbarr", obs=Crime.data$prbarr),
             data.frame(wcat="prbconv", obs=Crime.data$prbconv),
             data.frame(wcat="prbpris", obs=Crime.data$prbpris),
             data.frame(wcat="avgsen", obs=Crime.data$avgsen),
             data.frame(wcat="mix", obs=Crime.data$mix),
             data.frame(wcat="polpc", obs=Crime.data$polpc))

ggplot(wdf, aes(x=obs)) + geom_histogram(bins=20, colour='white') +
  ggtitle("Deterrence Variables") +
  facet_wrap( ~ wcat, scales="free")
```



prbarr, prbconv and prbpris

```
outlier.prob <- Crime.data %>% filter(prbarr > 1 | prbconv > 1 | prbpris > 1) %>%
  select(log.crmrte,prbarr,prbconv,prbpris,avgsen,polpc,density)
```

These variables represent the probability of an arrest given an offense, the probability of conviction given arrest, and the probability of imprisonment given a conviction. There is some suspect data here, as one of the *prbarr* records and **ten** of the *prbconv* records have probabilities greater than one, which is theoretically impossible. These are the records:

```
kable(outlier.prob,format="latex",booktabs=T) %>%  kable_styling(latex_options = "striped")
```

log.crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density
5.027374	0.132029	1.48148	0.450000	6.35	7.4588	1.0463320
5.400725	0.162860	1.22561	0.333333	10.34	20.2425	0.5767442
5.146709	0.153846	1.23438	0.556962	14.75	18.5912	0.5478615
4.013351	1.090910	1.50000	0.500000	20.70	90.5433	0.3858093
5.675026	0.179616	1.35814	0.335616	15.99	15.8289	1.3388889
4.841522	0.207143	1.06897	0.322581	6.18	8.1426	0.3167155
5.105867	0.271967	1.01538	0.227273	14.62	15.1871	0.6092437
4.688619	0.195266	2.12121	0.442857	5.38	12.2210	0.3887588
5.749307	0.201397	1.67052	0.470588	13.02	44.5923	1.7459893
4.955320	0.207595	1.18293	0.360825	12.23	11.8573	0.8898810

Despite the *prima facie* implausible numbers, we have decided to retain these data points for the following reasons:

- They might reflect arrests that are made from a county different from where offenses took place.
- They might reflect arrests made for offenses committed in a previous year.
- There could be other reasons for underreporting offenses or having arrests outnumber offenses.

Given the limited contextual information provided with the data, we felt we did not have enough information to reject these data points as abnormalities. We decided to retain them as-is for our analysis.

Of the three, the chance of getting arrested is lowest while the chance of conviction is highest, although the variance of convictions is quite high. The probability of going to prison appears the most normally-distributed, with nearly a third of counties sending convicted criminals to prison a little less than half the time.

Note

There is a difficulty in attempting to use a linear model to describe a feedback loop, whereby an increase in crime rate will naturally lead to more arrests, while an increase in arrest rate due to stronger policing would (hopefully) lower crime rate. As we cannot accurately describe this model without time series data, we assume a static, linear relationship for this analysis.

mix

Percent of crime offenses that were face-to-face (ie, stick-ups, assault, rape, etc). Although imperfect, this variable can act as proxy for the ratio of violent crime to property crime–given that the majority of violent crime happens with the victim present, while property crime is the inverse. The shape is in line with other distributions, with county 5 the outlier with nearly half the offenses face-to-face. Given that the crime rate in county 5 is one of the lowest in the state, as is its density, this is probably just a function of the small number of offenses there. Average mix is 0.13.

avgsen

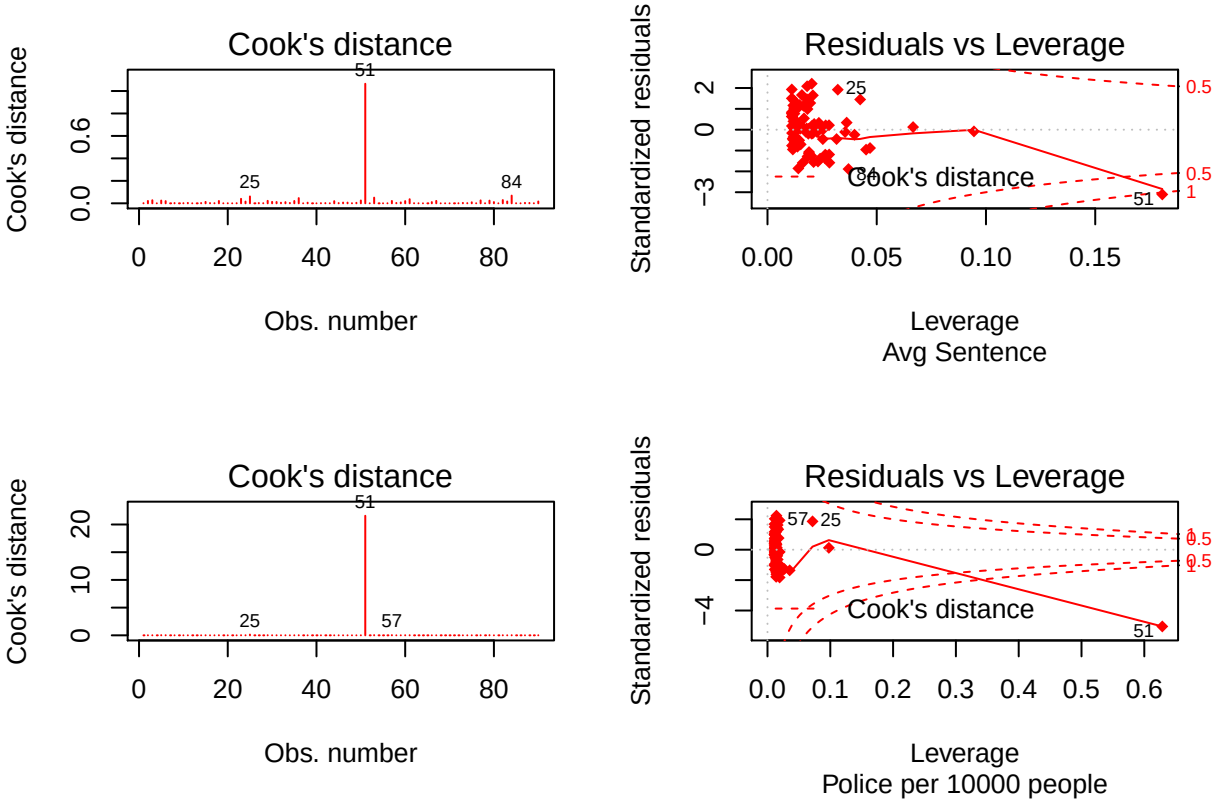
The average sentence, in days, for a prison conviction. Again there is a positive skew, with county 115 the outlier at 20.7 days. The average length of stay is 10, which does not seem very long. Maybe this variable includes county jails and local "lockups," and therefore includes drunk and disorderly overnights and other 1-2 night stays that bring down the average considerably. It is also possible that after this time period prisoners are transferred to state prisons. If that is the case, then the sentencing variable might not accurately reflect the true consequences of crime.

polpc

Average police per capita. For consistency, we perform a similar scaling as crime rate to reflect police per 10,000 people. Another positive skew distribution, again with county 115 (observation 51) a far outlier at 90.5. While this county also has a high probability of arrest and average length of sentence, in all other aspects the number of police is way out of line - for a county of similar density (based on 5 similar datapoints), race and affluence, one would expect a police force 1/10th the size. The average police per capita without this datapoint is 16.25.

Given suspicious outliers for two variables in county 115, we take a closer look and investigate the potential influence of these datapoints to the Crime rate using the cook's distance plots.

```
par(mfrow=c(2,2))
plot(lm(log.crmrte~avgsen,data=Crime.data),pch=18, col="red", which=c(4,5))
title(sub="Avg Sentence")
plot(lm(log.crmrte~polpc,data=Crime.data),pch=18, col="red", which=c(4,5))
title(sub="Police per 10000 people")
```



As the plot above indicates, the cook's distace for observation 51 has an abnormal influence for our main output variable of crime rate with both *polpc* and *avgsen* - although it is dramatically greater with *polpc*. To adjust for this, we impute these outlying values to the mean.

```
mean.polpc <- round(mean(Crime.data$polpc[Crime.data$polpc < 80]),2)
Crime.data$polpc <- ifelse(Crime.data$polpc > 80, mean.polpc, Crime.data$polpc)
mean.avgsen <- round(mean(Crime.data$avgsen [Crime.data$avgsen  < 19]),2)
Crime.data$avgsen <- ifelse(Crime.data$avgsen  > 19, mean.avgsen, Crime.data$avgsen)
```

Economic variables

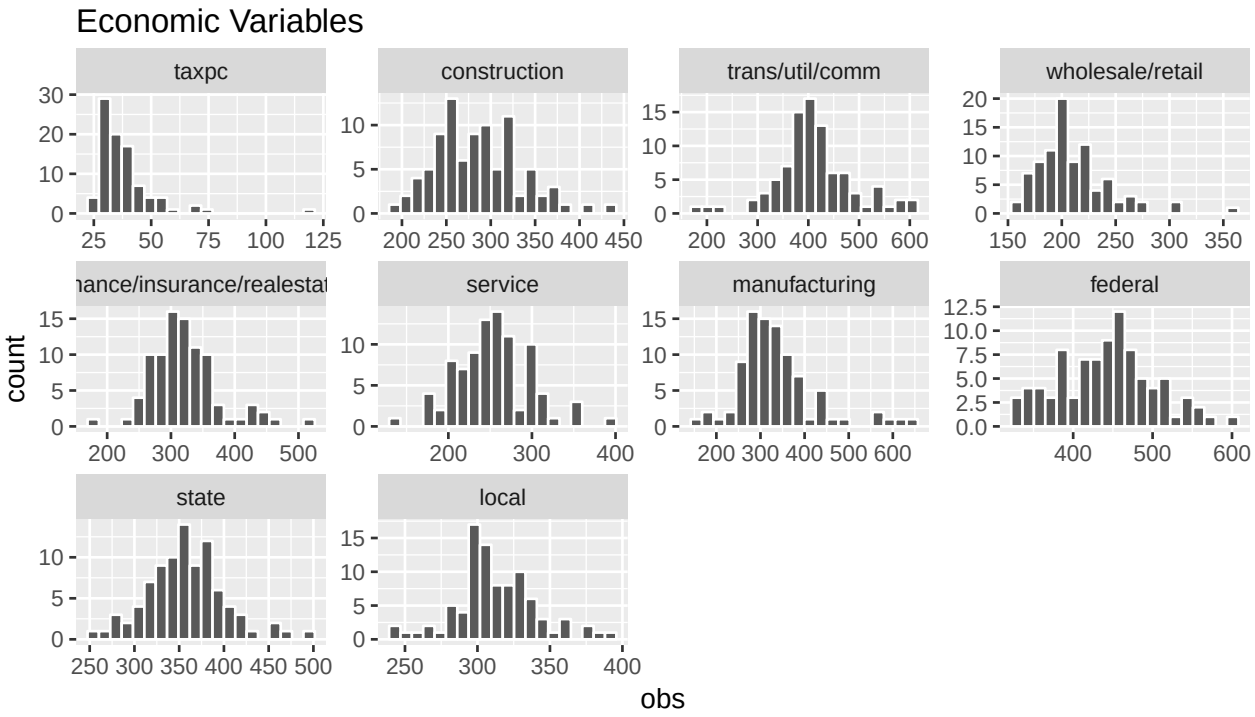
```
# Summary of data
kable(summary(Crime.data[,c("taxpc","wcon","wtuc","wtrd","wfir")]),format="latex",booktabs=T) %>%
    kable_styling(latex_options = "striped"); kable(summary(Crime.data[,c("wser","wmfg","wfed","wsta","
    kable_styling(latex_options = "striped")
```

taxpc	wcon	wtuc	wtrd	wfir
Min. : 25.69	Min. :193.6	Min. :187.6	Min. :154.2	Min. :170.9
1st Qu.: 30.73	1st Qu.:250.8	1st Qu.:374.3	1st Qu.:190.7	1st Qu.:285.6
Median : 34.92	Median :281.2	Median :404.8	Median :203.0	Median :317.1
Mean : 38.16	Mean :285.4	Mean :410.9	Mean :210.9	Mean :321.6
3rd Qu.: 41.01	3rd Qu.:315.0	3rd Qu.:440.7	3rd Qu.:224.3	3rd Qu.:342.6
Max. :119.76	Max. :436.8	Max. :613.2	Max. :354.7	Max. :509.5

wser	wmfg	wfed	wsta	wloc
Min. : 133.0	Min. :157.4	Min. :326.1	Min. :258.3	Min. :239.2
1st Qu.: 229.3	1st Qu.:288.6	1st Qu.:398.8	1st Qu.:329.3	1st Qu.:297.2
Median : 253.1	Median :321.1	Median :448.9	Median :358.4	Median :307.6
Mean : 275.3	Mean :336.0	Mean :442.6	Mean :357.7	Mean :312.3
3rd Qu.: 277.6	3rd Qu.:359.9	3rd Qu.:478.3	3rd Qu.:383.2	3rd Qu.:328.8
Max. :2177.1	Max. :646.9	Max. :598.0	Max. :499.6	Max. :388.1

```
Crime.data <- mutate(Crime.data,wser_orig = wser)
Crime.data$wser = ifelse(Crime.data$wser > 1000, mean(Crime.data$wser[Crime.data$wser < 1000],
                                                      na.rm = TRUE), Crime.data$wser)
# Make a master dataframe for wages
wdf <- rbind(data.frame(wcat="taxpc", obs=Crime.data$taxpc),
             data.frame(wcat="construction", obs=Crime.data$wcon),
             data.frame(wcat="trans/util/comm", obs=Crime.data$wtuc),
             data.frame(wcat="wholesale/retail", obs=Crime.data$wtrd),
             data.frame(wcat="finance/insurance/realestate", obs=Crime.data$wfir),
             data.frame(wcat="service", obs=Crime.data$wser),
             data.frame(wcat="manufacturing", obs=Crime.data$wmfg),
             data.frame(wcat="federal", obs=Crime.data$wfed),
             data.frame(wcat="state", obs=Crime.data$wsta),
             data.frame(wcat="local", obs=Crime.data$wloc))

ggplot(wdf, aes(x=obs)) + geom_histogram(bins=20, colour='white') +
  ggtitle("Economic Variables") +
  facet_wrap( ~ wcat, scale="free")
```



taxpc

Average annual tax revenue per capita. Similar shape to **polpc**, but this time the outlier is county 55 at $119.76/person. This is significantly higher than the next highest county, but county 55 is also one of the least populated, so it's possible that a handful of very high net worth individuals could pull up the average revenue of $38.16. Without a persuasive reason why it doesn't belong, we cannot discard it.

wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc

Average weekly wages for: construction, transportation/utility/communication, wholesale/retail, finance/insurance, service, manufacturing, federal, state, and local governments. There is a clear outlier in service wages, where county 185 reports an average wage over $2000, more than three times the next highest county and in contradiction of its tax per capita, which is clearly middle-of-the-pack. We have imputed that value to the mean.

Breaking out the individual distributions, a few things are notable: federal employees seem to be the most highly-paid in the state, even over the finance sector. Wholesale/retail are the lowest-paying jobs. Manufacturing has some outliers on the upper end of the distribution.

Demographic variables

```
# Summary of data
kable(summary(Crime.data[,c("density","west","central","urban","pctmin80","pctymle")]),format="latex",b
    kable_styling(latex_options = "striped")
```

density	west	central	urban	pctmin80	pctymle
Min. :0.00002	Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. : 1.284	Min. :0.06216
1st Qu.:0.54718	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:10.024	1st Qu.:0.07437
Median :0.97925	Median :0.0000	Median :0.0000	Median :0.00000	Median :24.852	Median :0.07770
Mean :1.43567	Mean :0.2444	Mean :0.3778	Mean :0.08889	Mean :25.713	Mean :0.08403
3rd Qu.:1.56926	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:0.00000	3rd Qu.:38.183	3rd Qu.:0.08352
Max. :8.82765	Max. :1.0000	Max. :1.0000	Max. :1.00000	Max. :64.348	Max. :0.24871

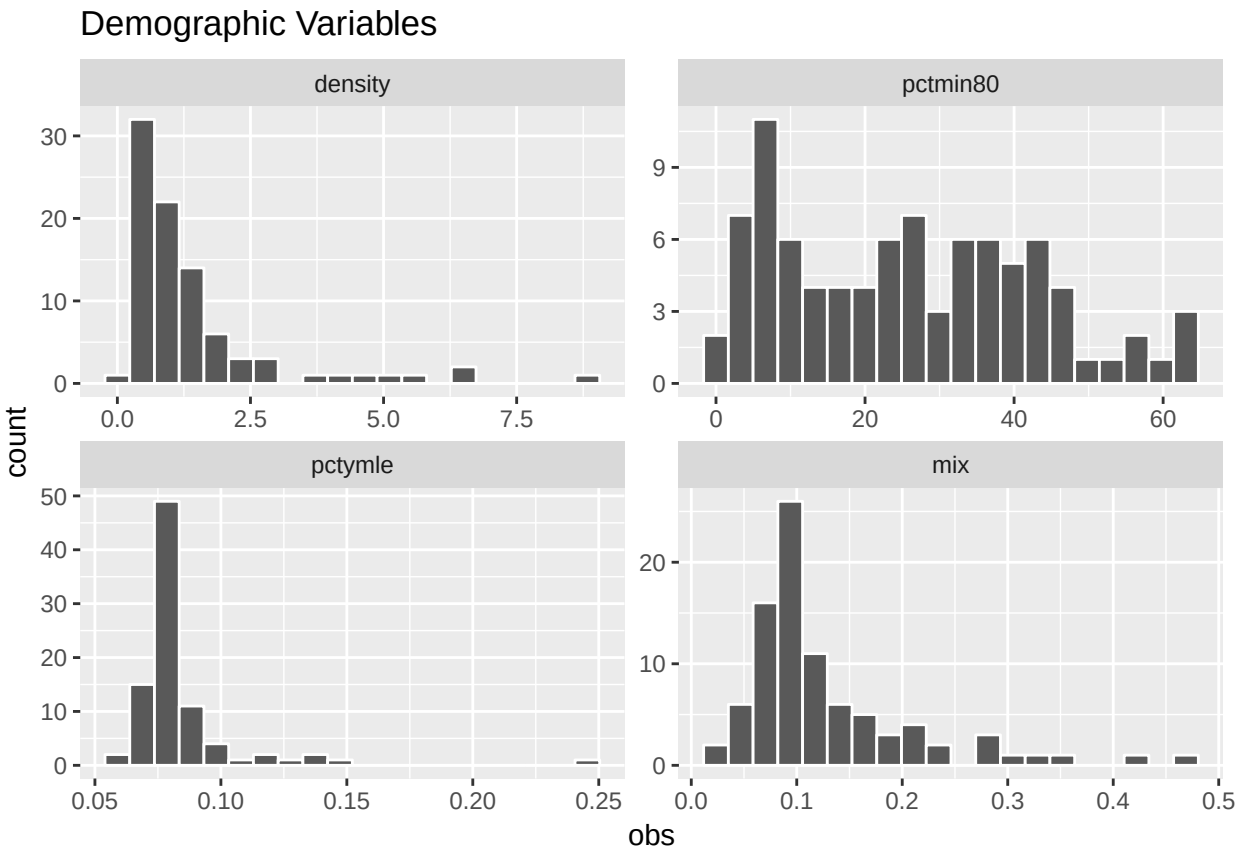
west, central, urban

These are dummy variables with binary identifiers to label a county as belonging in the west or central portions of the state, and/or whether the county qualifies as "urban." The geographic data is of marginal utility, as it isn't comprehensive - many, if not most, of the counties are neither west nor central, but we have no way of knowing if they are northern, southern or eastern. There is one county (71) that is both western and central, which we omit as we can't know which category it belongs to. Similarly, while it is useful to know if a county is urban, we find the "density" variable to be a much better descriptor here. As such, we omit *urban* from our consideration.

```
# Drop county that is both western and central
Crime.data <- Crime.data[-71,]

wdf <- rbind(data.frame(wcat="density", obs=Crime.data$density),
             data.frame(wcat="pctmin80", obs=Crime.data$pctmin80),
            data.frame(wcat="pctymle", obs=Crime.data$pctymle),
            data.frame(wcat="mix", obs=Crime.data$mix))

ggplot(wdf, aes(x=obs)) + geom_histogram(bins=20, colour='white') +
  ggtitle("Demographic Variables") +
  facet_wrap( ~ wcat, scale="free")
```



density

People per square mile. There is significant positive skew here, but that makes sense: most counties are rural, while a few in the cities will be denser. None of the high counties seem implausibly dense (indeed, the highest value is only 8 people per square mile). County 119, the positive outlier here, was also the outlier for *log.crmrte*, suggesting there might be a correlation between density and crime rate. Average density is 1.43 people per square mile, which is very sparse. County 173 (observation 79), appears to be a low outlier with a value of 0.00002 people per square mile, which seems impossible given humans' integral nature (the county would need to be 200,000 square miles–larger than California–and contain a single person to make this true). We impute this value to the mean.

```
mean.density <- round(mean(Crime.data$density[Crime.data$density > .02]),2)
Crime.data$density <- ifelse(Crime.data$density < .02, mean.density, Crime.data$density)
```

pctmin80

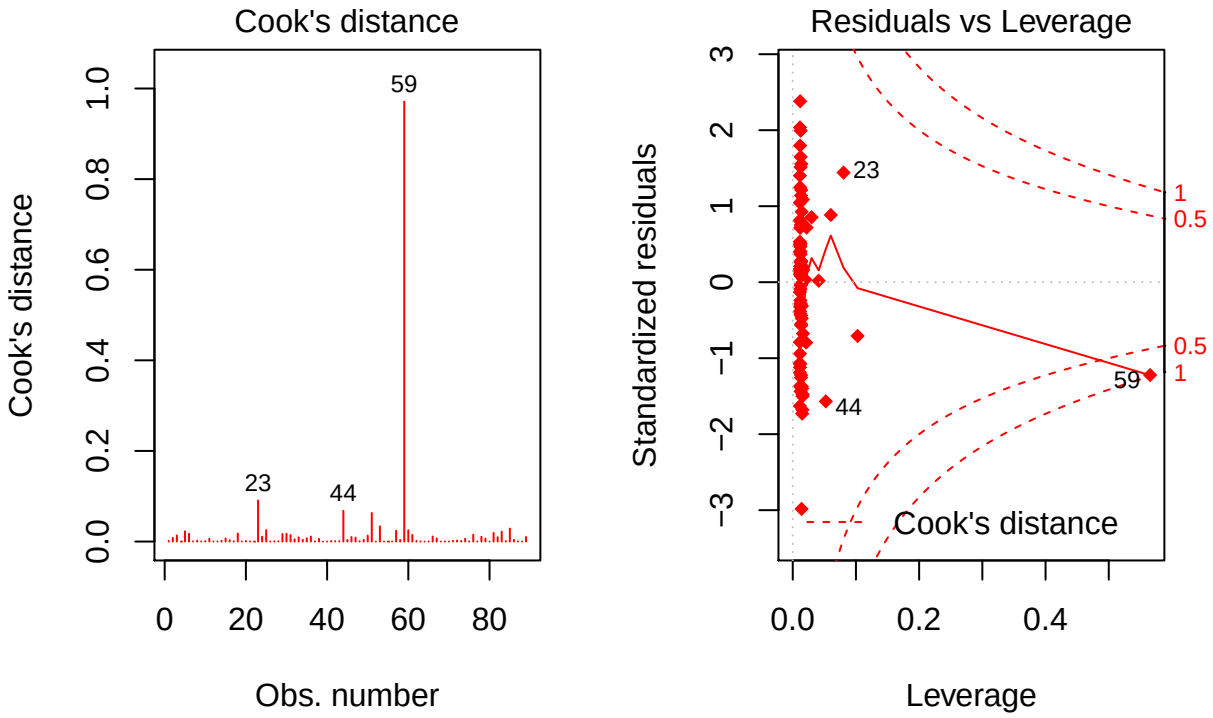
Minority percentage of the population (in 1980). While every other percentage variable in the dataset is presented as a fraction of 1, this value is between 0 and 100. For the sake of consistency in interpretation, we divide this variable by 100. Unlike most of the other variables, this is a broad distribution without a particularly dramatic positive skew or any outliers. Average share is 25.76%.

```
Crime.data$pctmin80 = Crime.data$pctmin80/100
```

pctymle

Young male percentage of the population. Tight clustering around the average of 8.41% young males, while county 133 (observation 59) has nearly a quarter the population, far more than any other county. This could possibly be explained by an army base located in the county. (Another explanation, a penitentiary, is belied by the fact that the average sentence length for this county is below average.) To gain a better perspective, we examine the cook's distance for this variable.

```
par(mfrow=c(1,2))
plot(lm(log.crmrte~pctymle,data=Crime.data),pch=18, col="red", which=c(4,5))
```



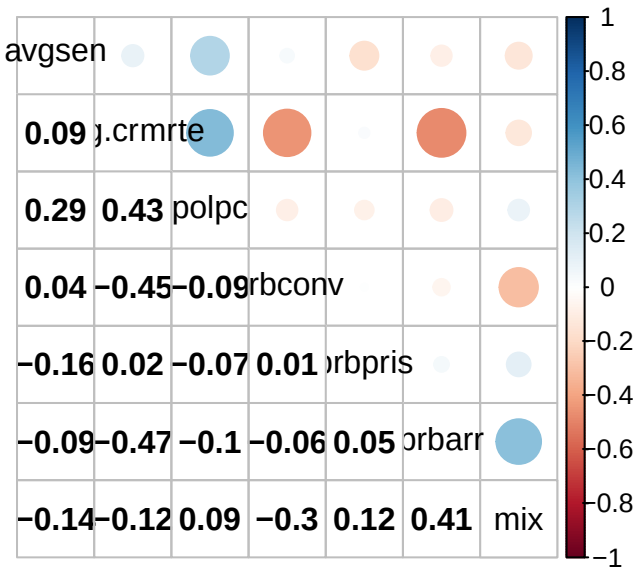
Cook's distance of 1 is right on the border of acceptability, but as the other key variables (*density*, *log.crmrte*, *polpc*) are not in line with the other counties with high *pctymle*, we impute this observation to the mean.

```
mean.pctymle <- round(mean(Crime.data$pctymle[Crime.data$pctymle < .2]),2)
Crime.data$pctymle <- ifelse(Crime.data$pctymle > .2, mean.pctymle, Crime.data$pctymle)
```

EDA, MULTIVARIATE/CORRELATION ANALYSIS

Deterrence Variables

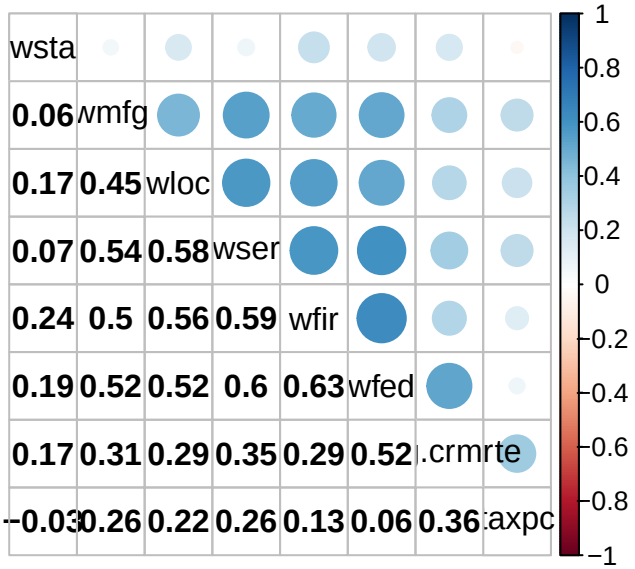
```
prob.data <- Crime.data[,c("log.crmrte","prbarr","prbconv","prbpris","avgsen","polpc", "mix")]
res <- cor(prob.data)
corrplot.mixed(res, order = "hclust", lower.col = "black", tl.col = "black", tl.srt = 45)
```



As per the correlation matrix, the only significant correlations to crime rate is with the Probabilty of Arrest and Probability of Conviction. In addition *polpc* and *prbarr* show positive correlation, as discussed in our note about the feedback loop between these variables. Sentencing does not seem to have much correlation, perhaps because sentencing does not seem particularly harsh (avg of 10 days). *mix* is correlated with *prbarr*, probably because the crimes are higher priority and the criminals are more identifiable, but it has a negative correlation to *prbconv*, which is puzzling–why are counties with a higher mix of face-to-face (and by proxy, violent) crime less likely to convict? It may be harder to get a conviction for a more serious crime which carries a more serious punishment, or it may be that perpetrators of in-person crime are generally more capable of hiring better lawyers.

Economic Variables

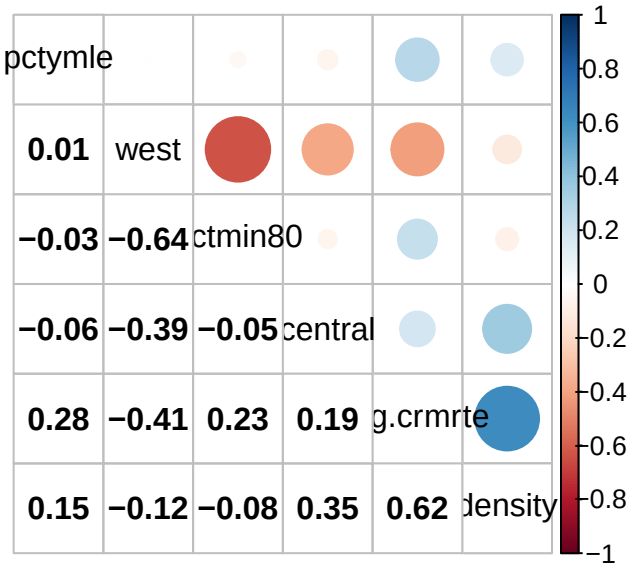
```
wage.data <- Crime.data[,c("log.crmrte","taxpc","wfir","wser","wmfg","wfed","wsta","wloc")]
res <- cor(wage.data)
corrplot.mixed(res, order = "hclust", lower.col = "black", tl.col = "black", tl.srt = 45)
```



Most wage components show a positive correlation with crime, and with each other. This is perhaps an indication of the affluence of the population as an indicator of more monetary compensation and hence incentive for crime. Although *wfed* has the highest correlation to crime rate, the proportion of federal employees in a given county must be very small, and given their wages are the highest of the group, this is likely more a proxy for affluence than a meaningful indicator of wages. Since *wmfg* is a high-employment sector, shows similar correlation to crime as the other wage variables, and shows a high degree of intercorrelation with other wage variables, we keep it as a proxy for wages in general and omit the rest from our analysis.

Demographic Variables

```
pdtmy.data <- Crime.data[,c("log.crmrte","pctmin80","density","pctymle", "west", "central")]
res<- cor(pdtmy.data)
corrplot.mixed(res, order = "hclust", lower.col = "black", tl.col = "black", tl.srt = 45)
```

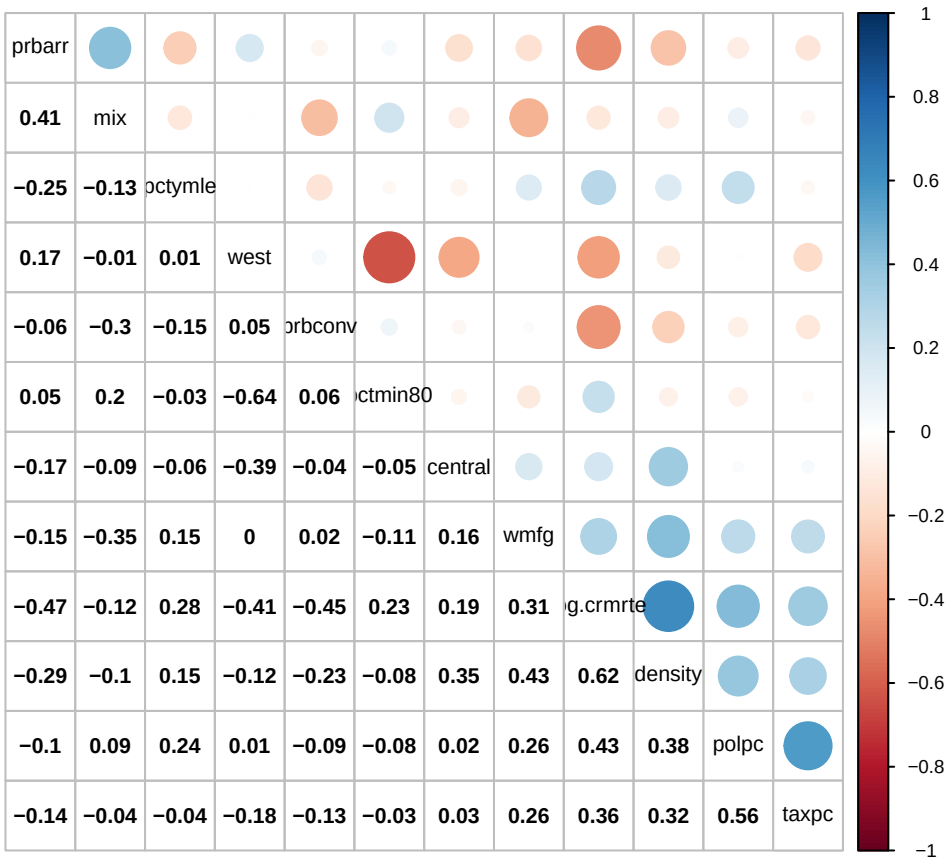


Density has the highest correlation to crime. Percent of young male (*pctymle*) and Percent of minority (*pctmin80*) also show some correlation. There is a strong negative correlation between both *west* and *log.crmrte* and *west* and *pctmin80*, suggesting that the western counties are both less diverse and less prone to crime than the rest of the state. *central* appears to be a bit more dense, but its relationship to *log.crmrte* is pretty weak.

Implications and intercorrelation

Based on these findings we find the variables with highest correlation to crime rate to be *prbarr*, *prbconv*, *polpc*, *density*, *taxpc*, *wmfg*, *pctmin80*, *west*, and *pctymle*. We can therefore remove from further analysis the variables *prbpris*, *central*, *urban* and all wage variables besides *wmfg*. Now we want to check intercorrelation between these prime variables.

```
demo.data <- Crime.data[,c("log.crmrte","prbarr","prbconv","polpc","density","taxpc","wmfg","pctmin80",
res<- cor(demo.data)
cex.before <- par("cex")
par(cex = 0.7)
corrplot.mixed(res, order = "hclust", lower.col = "black", tl.col = "black", tl.srt = 45)
```




```
par(cex = cex.before)
```

Density has high positive correlation with economic variables *taxpc* and *wmfg* and negative correlation with *prbarr* and *prbconv*, which makes sense as *polpc* has very weak positive correlation with density, therefore the higher crime rate in urban counties is being met with approximately the same number of police (per person), implying a criminal is less likely to be caught and punished the more dense his environment. *polpc* correlates with *prbarr*, which is to be expected: more cops, more arrests. Less expected is the correlation of *polpc* with *wmfg* and *taxpc*, but this may just be a function of more affluent counties being able to afford more police. *mix* has a weak negative correlation with *log.crmrte*, and a negative relationship with *wmfg*, which is puzzling but probably not relevant. Somewhat surprising is the lack of any correlation between *density* and *pctmin80*, as typically cities tend to be more diverse than rural areas.

Although none of the variables are collinear, we want to be careful about including variables that are highly correlated with each other in our models. Therefore if we pair *west* with *pctmin80*, *prbarr* and *mix*, *polpc* and *taxpc*, or *density* and *wmfg* in any of our models, we will want to double-check the Variance Inflation Factors to be sure the multicollinearity is not unacceptably high.

MODELING

Going by our research question on the main predictors of crime rate and the insights gained by the subsequent EDA, we now transition to the modeling phase. As seen during the correlation analysis the deterrence variables stood out as the most promising candidates for further modeling. It is also clear that density has a sizable impact on crime rate. Although it is hard to directly influence this variable as part of a political campaign strategy, we need to keep it as a control variable in our modeling so as to get reliable estimates for the coefficients of our variables of interest.

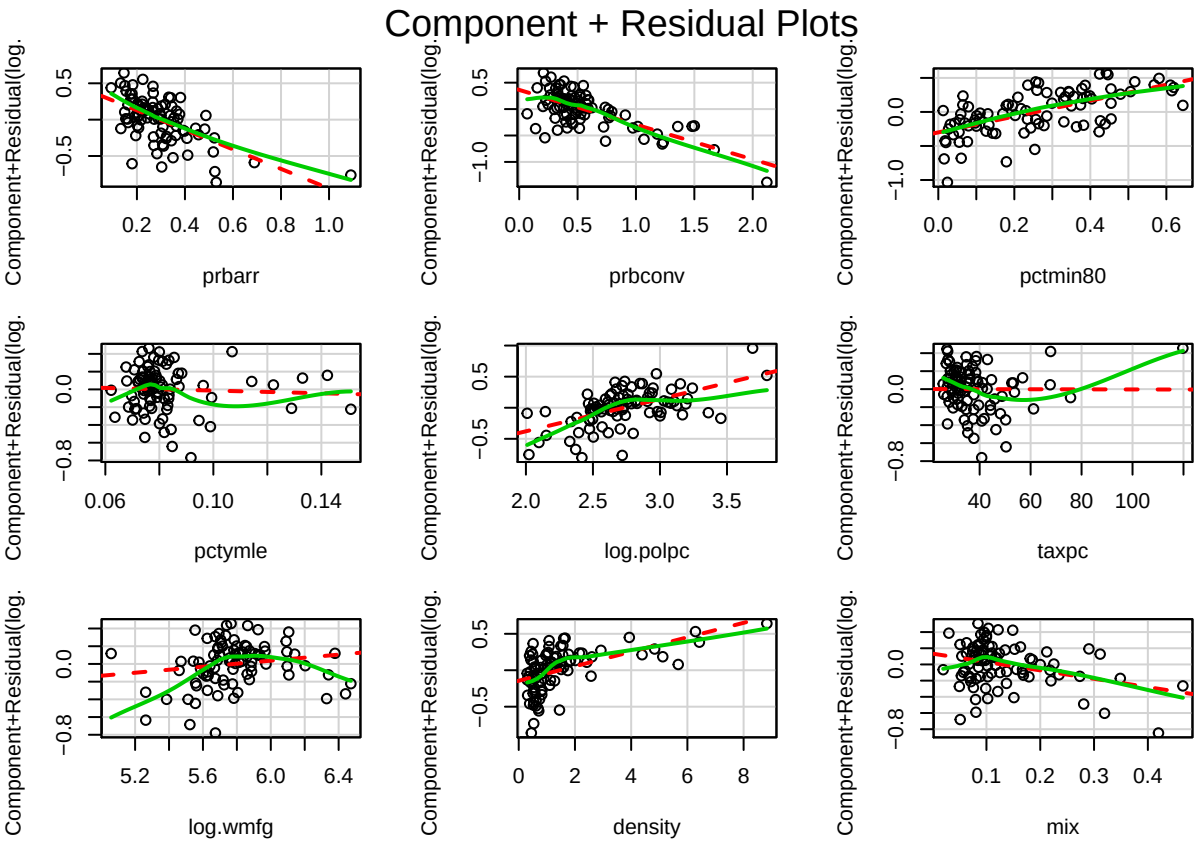
Linearity Analysis

Earlier, we transformed *polpc* from per capita rate to per 10000 people to match the transformed units of crime rate. Here we also take a log of *polpc/wmfg* variables before fitting our models as the effect of increase in police or wages to crime rate should eventually lead to diminishing returns. These variables are subsequently referred to as *log.polpc* and *log.wmfg*, respectively.

```
# Log transform wmfg and polpc
Crime.data$log.polpc <- log(Crime.data$polpc); Crime.data$log.wmfg <- log(Crime.data$wmfg)
```

We have identified nine variables with some degree of correlation to *log.crmrte*, but have not yet established whether the relationship is linear. While we could use scatterplots for this, a Component+Residual plot is a better analysis as it takes into account the other independent variables in the group.

```
lin_check <- lm(log.crmrte ~ prbarr + prbconv + pctmin80 + pctymle + log.polpc + taxpc + log.wmfg + dens
crPlots(lin_check)
```



For most independent variables, the linear relationship is clear. However *pctymle*, *taxpc* and *log.wmfg* appear to have some degree of nonlinearity (although in the case of *taxpc* it appears to be one outlier responsible for most of the curve). The slope of the relationship for *pctymle* and *taxpc* appears quite flat, implying a weak relationship, while *mix* looks surprisingly strong.

Model 1 - only the explanatory variables of key interest

In our multivariate analysis, we observed a greater number of correlations between the deterrence variables and *log.crmrte* than for either the economic or demographic variable sets. Therefore for the first model, we include the prime deterrence variables *prbarr*, *prbconv*, *log.polpc*, and include *density* as a control variable. These are the "tough on crime" variables that showed the most promise in our correlation analysis, and should

give a sense of how effective these variables would be in predicting the efficacy of this policy approach.

$$log(crime\ rate) = \beta_0 + \beta_1 \cdot Pr(arrest) + \beta_2 \cdot Pr(conv) + \beta_3 \cdot log(polpc) + \beta_4 \cdot density + u$$

```
with (Crime.data, mod1 <<- lm(log.crmrte ~ prbarr + prbconv + log.polpc + density ))
```

Next, we test the 6 conditions for the validity of the CLM assumptions to ensure that the BLUE conditions hold.

CLM 1 - Linear Model

The model is specified with linear coefficients. Also we have done linearity checks from predictor to output variables in the previous section. There are no violations to this assumption as per our model specification.

CLM 2 - Random Sampling

The data represents a sampling from most (though not all) of the counties of North Carolina. Although county-level data might introduce clustering errors in other datasets, because all these variables are specific to the county and do not include multiple random draws from within counties, we shouldn't worry about this bias. The data might exhibit specific regional/cultural influences from the state and hence we note that a general conclusion cannot be drawn on the pattern of crimes aross the country. Since the data are for a particular year, we also do not expect auto-correlation bias arising from sampling the same region in different points of time.

CLM 3 - Perfect Multicollinearity

We dont need to explicitly check for Multicollinearity as R would abort while running the model in such a case. However, we can check the extent of collinearity between our independent variables using the Variance Inflation Factors (VIF):

```
vif(mod1)
```

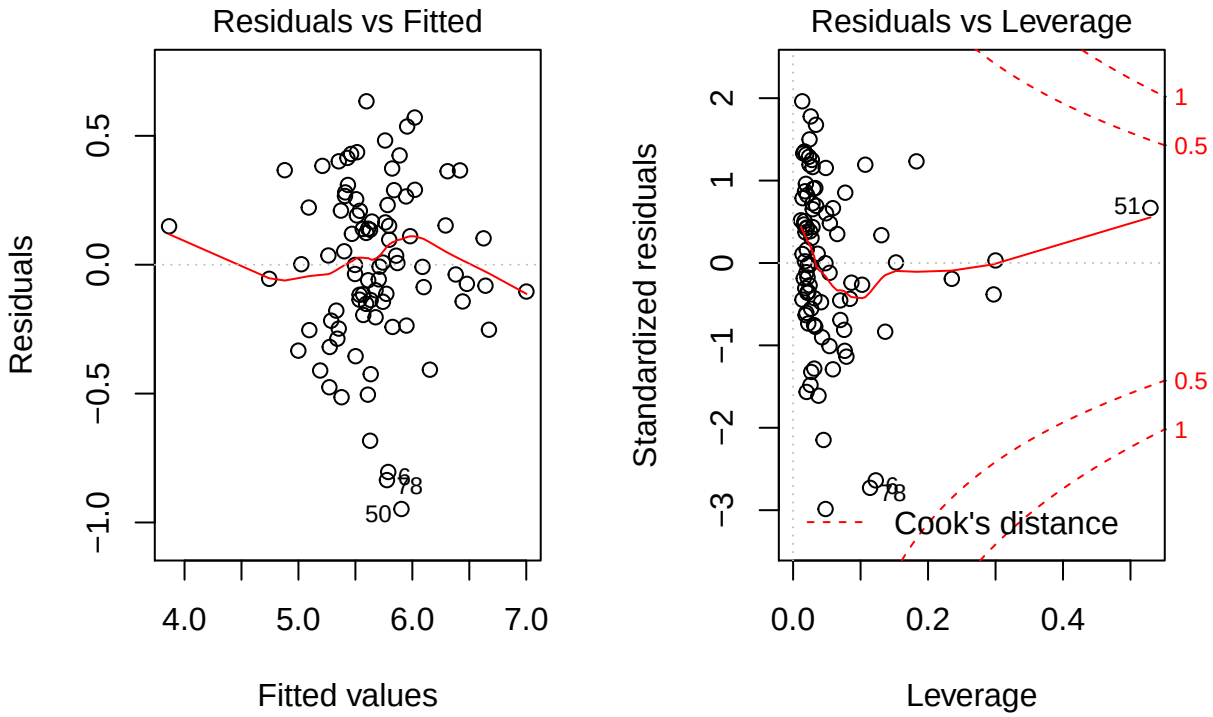
```
##    prbarr   prbconv log.polpc   density
##  1.109335  1.082004  1.251862  1.407563
```

As per the results above, the VIF's for all the independent variables are quite close to 1. This indicates only a minor extent of influence on the variance of the indpendent variables coefficients by their mutual interaction.

CLM 4 - Zero-Conditional Mean

For observing the zero-conditional mean we plot the fitted vs residual plots and the leverage plot of the standardized residuals:

```
par(mfrow=c(1,2))
plot(mod1,which=c(1,5))
```

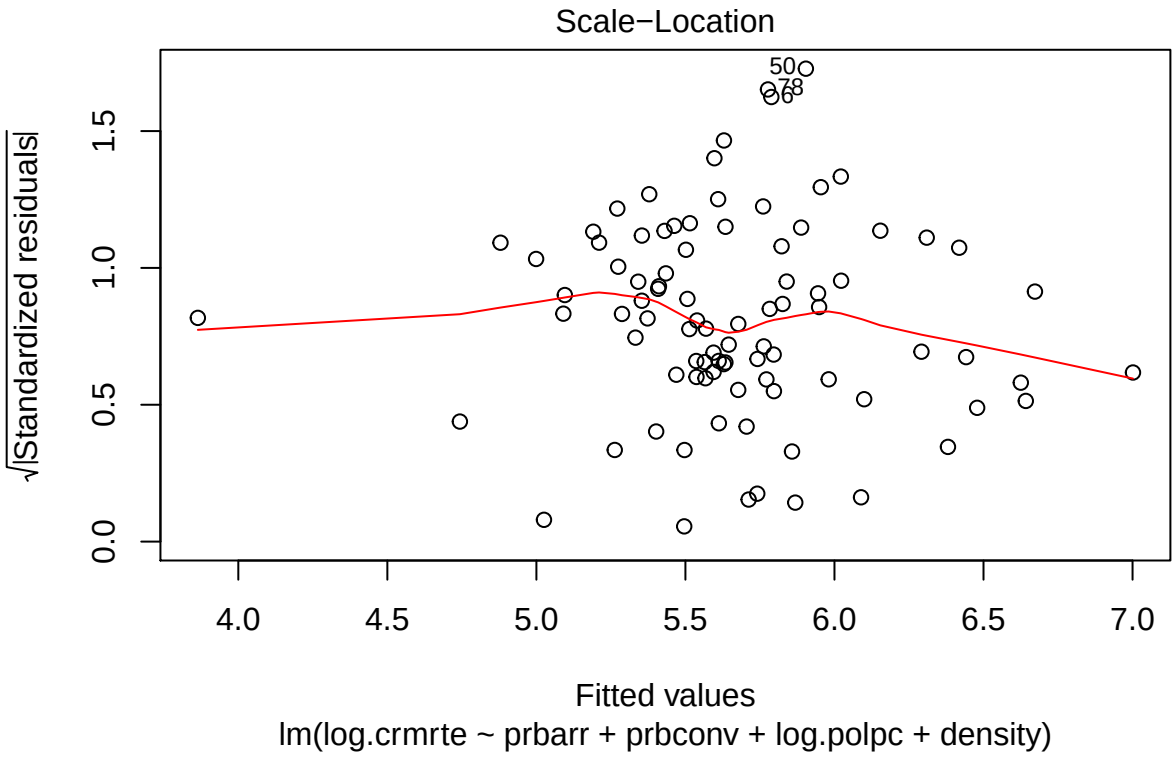


From the residual plots we do not observe any serious deviations from zero conditional mean. There is some curvature at both the extremes, mostly due to a few outlier values. The bulk of the residuals seem to be uniformly scattered around the mean. For the extreme residual points, although they have leverage, they do not violate Cook's distance even at the 0.5 level. The data seems to indicate that our coefficients could potentially be biased. We perform a detail omitted variable analysis in a later section to explore the sources of bias.

CLM 5 - Homoskedasticity

Firstly, we use the fitted values vs standardized residuals plot to check for homoskedasticity.

```
plot(mod1,which=3)
```



It is a bit hard to determine given the relative scarcity of datapoints, but the oblong shape of the distribution suggests that we may have a violation of homoskedaticity. To confirm further we run the Breusch-Pagan test and the non-constant variance test:

```
bptest(mod1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 16.688, df = 4, p-value = 0.002222
```

```
ncvTest(mod1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.01472129    Df = 1     p = 0.9034286
```

The bptest provides a statistically significant result (indicating heteroskedasticity, although this test is sensitive to large sample size), however the ncvTest has a high p-value, suggesting the opposite. So, we see mixed signals for heteroskedasticity. To address this possible violation, we generate **heteroskedasticity robust standard errors** to be conservative in computing our significance tests. We use these standard errors in all our models, simply as good practice.

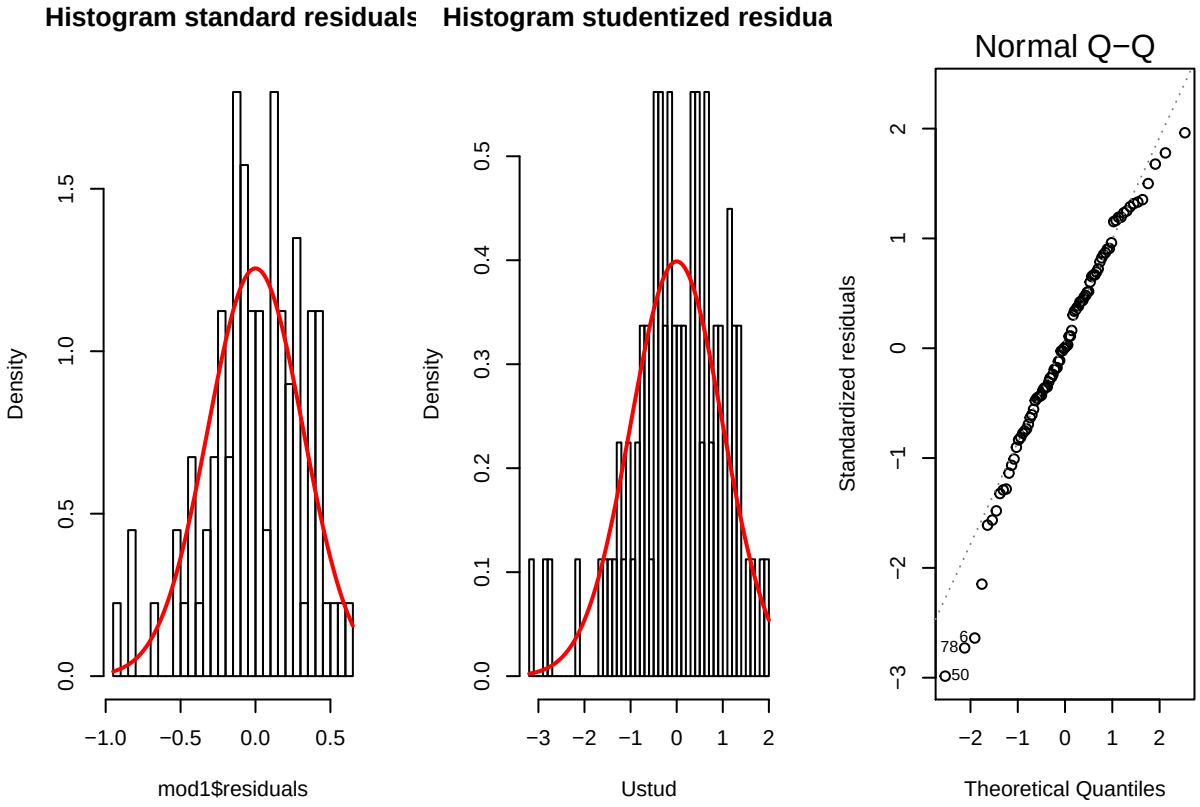
```
se.m1 <- sqrt(diag(vcovHC(mod1)))
```

CLM 6 - Normality of Residuals

We check the normality of errors using both histograms, qq-plots and the Shapiro-Wilk test.

```
# For model1
par(mfrow=c(1,3))
```

```
hist(mod1$residuals, main="Histogram standard residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(mod1$residuals)), col="red", lwd=2, add=TRUE)
Ustud = rstudent(mod1)
hist(Ustud, main="Histogram studentized residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
plot(mod1,which=2)
```




```
shapiro.test(mod1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.97628, p-value = 0.1025
```

The histogram and quantile plot indicates the majority of points follow the normal curve. There is some misalignment in the lower and higher quantiles, however given the large sample size, we are confident this is not a major issue given the Central Limit Theorem.

The Shapiro-Wilk test confirms this by yielding a high p-value (0.10) confirming the NULL hypothesis of normality.

With this check, all the 6 assumptions of the CLM are met to a satisfactory extent given the sample size. Next we evaluate the coefficients generated by the model and examine their statistical and practical significance.

```
stargazer(mod1,report = "vct*", title = "Model 1 Summary", add.lines=list(c("AIC", round(AIC(mod1), 1))),
          se=list(se.m1), star.cutoffs=c(0.05,0.01,0.001),header = FALSE,float=FALSE)
```

	Dependent variable:
	log.crmrte
prbarr	−1.507
	t = −4.750***
prbconv	−0.541
	t = −5.547***
log.polpc	0.472
	t = 3.027**
density	0.112
	t = 4.443***
Constant	4.961
	t = 11.578***
AIC	59.5
Observations	89
R^2	0.668
Adjusted R^2	0.652
Residual Std. Error	0.325 (df = 84)
F Statistic	42.294*** (df = 4; 84)
Note:	*p<0.05; **p<0.01; ***p<0.001

Statistical significance: All the variables above have high statistical significance, as evidenced by the high t stats and p values at the 0.01 or lower levels. **Practical significance:** All other factors remaining equal, increasing *prbarr* by 0.1 is predicted to lead to a 15.07% reduction in crime rate. While this estimate needs to be tempered due to omitted variable biases, it clearly has a high practical significance. While the remaining 3 variables exhibit lower coefficients, it's important to remember the units of each variable in this evaluation. For example, adding just one person per square mile in a county predicts an 11.2% increase (*ceteris paribus*) in crime rate–highlighting that this dataset describes an extremely sparse state. The log-log nature of regressing *log.polpc* against *log.crmrte* means that a 1% increase in a county's police force predicts nearly half a percentage increase in that county's crime rate, although this hearkens back to the feedback loop dynamic we described earlier.

While the Deterrence variables alone do a good job of predicting crime, adjusted R^2 of 0.668 could be improved.

Model 2: Adding Covariates which increase the accuracy of the model

Here we introduce some of the other variables which show some degree of correlation with *log.crmrte* as well as a solid linear relationship, namely *west*, *pctmin80* and *mix*. While economic variables such as *taxpc* and *log.wmfg* are tempting due to their fairly high correlation, the nonlinearity of their relationship to *log.crmrte* weakens CLM assumptions, and as we will see in model 3, they do not ultimately add to the accuracy of the model.

$$log(crime\,rate) = \beta_0 + \beta_1 \cdot Pr(arrest) + \beta_2 \cdot Pr(conv) + \beta_3 \cdot log(polpc) + \beta_4 \cdot density + \beta_5 \cdot pctMinority + \beta_6 \cdot mix + \beta_7 \cdot west + u$$

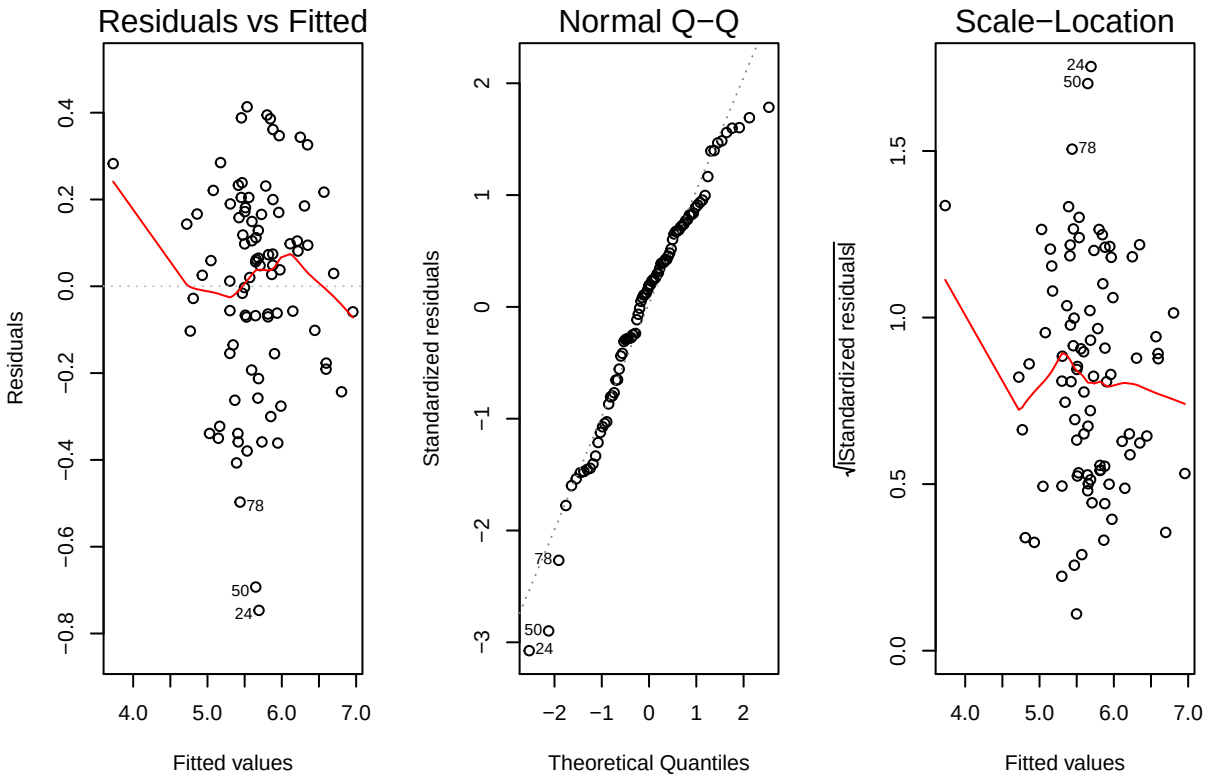
CLM Assumptions for Model 2

For Models 2 and 3, we will provide a less-detailed accounting of the CLM analysis than we provided for Model 1.

```
with (Crime.data, mod2 <<- lm(log.crmrte ~ prbarr + prbconv + log.polpc + density + pctmin80 + mix + wes
se.m2 <- sqrt(diag(vcovHC(mod2)))
vif(mod2)
```

```
##    prbarr   prbconv log.polpc   density  pctmin80       mix      west
##  1.347557  1.233491  1.277379  1.447290  1.923645  1.444992  1.893321
```

```
par(mfrow=c(1,3))
plot(mod2,which=c(1,2,3))
```




```
bptest(mod2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 8.7724, df = 7, p-value = 0.2694
```

```
ncvTest(mod2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.3196898    Df = 1     p = 0.5717941
```

```
shapiro.test(mod2$residuals)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  mod2$residuals
## W = 0.9625, p-value = 0.01138
```

- **CLM 1: Linearity** - Additional variables *pctmin80* and *mix* show linear relationship with dependent variable.
- **CLM 2: Random Sampling** - met with Model 1.
- **CLM 3: Multicollinearity** - *pctmin80* and *west* show modest VIF of nearly 2, but not enough to warrant exclusion.
- **CLM 4: Zero Conditional Mean** - Residuals vs. Fitted values shows a mean close to zero, albeit with some movement by a low outlier.
- **CLM 5: Homoskedasticity** - Both the BP test and ncvTest show heteroskedasticity, which Scale-Location confirms. We use robust Standard Errors.
- **CLM 6: Normality of Residuals** - QQ Plot and low p-value for Shapiro-Wilk reveal some deviation from normality at the highest quartile. However the relatively large sample size should address this.

Model 2

```
stargazer(mod2,report = "vct*", title = "Model 2 Summary", add.lines=list(c("AIC", round(AIC(mod2), 1))
          se=list(se.m2), header = FALSE, float = FALSE)
```

	Dependent variable:
	log.crmrte
prbarr	−1.191
	t = −2.574**
prbconv	−0.626
	t = −4.941***
log.polpc	0.551
	t = 4.723***
density	0.102
	t = 4.744***
pctmin80	0.814
	t = 3.536***
mix	−1.142
	t = −2.171**
west	−0.205
	t = −2.094**
Constant	4.701
	t = 12.274***
AIC	16.8
Observations	89
R^2	0.808
Adjusted R^2	0.792
Residual Std. Error	0.252 (df = 81)
F Statistic	48.753*** (df = 7; 81)
Note:	*p<0.1; **p<0.05; ***p<0.01

Statistical significance: Every variable except *prbarr*, *west*, and *mix* are highly significant ($p < 0.01$). Those three are significant at the ($p<0.05$) level. **Practical significance:** The -1.142 coefficient for *mix*, however, implies that shifting the mixture of face-to-face up 10% would lower the crime rate by 11.42%. (However, although this would mean fewer crimes overall, a higher proportion of them would be violent crime, so this might be ill-advised as a policy goal.) The new demographic variables help to paint a clearer picture - a 10 percentage point increase in minority population predicts an 8% rise in crime rate, while a western county can expect a 20.5% lower crime rate.

With an adjusted R^2 of .792, this model explains almost 80% of the crime rate with only seven independent variables, which is a good balance of accuracy and parsimony. AIC of 16.8 compares to 59.5 in Model 1, which is a clear improvement.

Model 3: Every variable with any sign of relationship

Here we include any variable with any kind of correlation, regardless of strength or linearity (these include *prbpris*, *avgsen*, *log.wmfg*, *central* and *taxpc*), and compare the three models side-by-side.

$$\begin{aligned} log(crime\ rate) = \beta_0 &+ \beta_1 \cdot Pr(arrest) + \beta_2 \cdot Pr(conv) + \beta_3 \cdot Pr(pris) + \beta_4 \cdot avgsen \\ &+ \beta_5 \cdot log(polpc) + \beta_6 \cdot density + \beta_7 \cdot taxpct + \beta_8 \cdot log(wage.mfg) + \\ &+ \beta_9 \cdot pctMinority + \beta_{10} \cdot pctYMale + \beta_{11} \cdot mix + \beta_{12} \cdot west + \beta_{13} \cdot central + u \end{aligned}$$

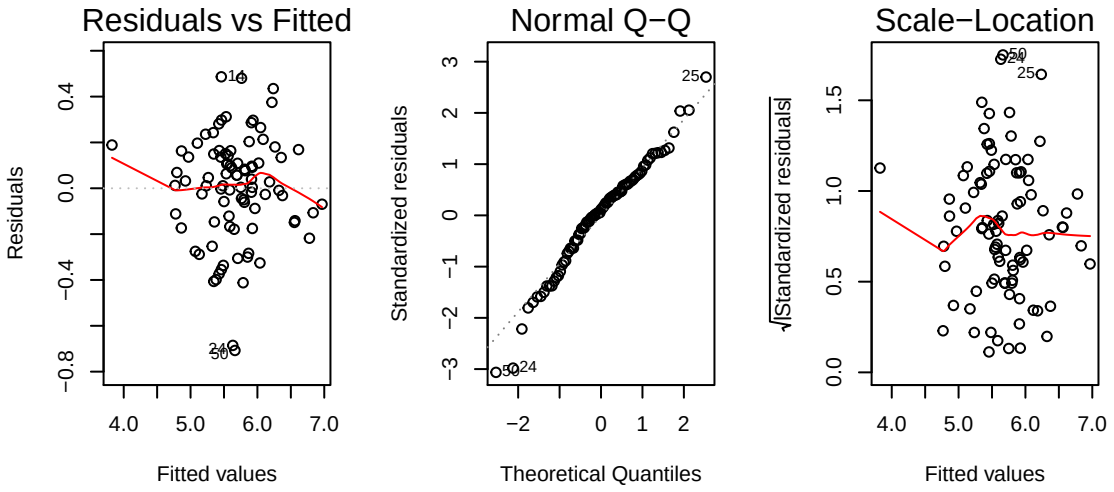
CLM Assumptions for Model 3

For Models 2 and 3, we will provide a less-detailed accounting of the CLM analysis than we provided for Model 1.

```
with (Crime.data, mod3 <<- lm(log.crmrte ~ prbarr + prbconv + prbpris + avgsen + log.polpc +
                                density + taxpc + log.wmfg + pctmin80 + pctymle + mix + west + central))
se.m3 <- sqrt(diag(vcovHC(mod3)))
vif(mod3)
```

```
##    prbarr   prbconv   prbpris    avgsen log.polpc   density     taxpc
##  1.442006  1.299882  1.114941  1.227866  2.108419  1.768511  1.707227
##  log.wmfg  pctmin80   pctymle       mix      west   central
##  1.636592  2.490569  1.339669  1.828331  2.987362  1.847772
```

```
par(mfrow=c(1,3))
plot(mod3,which=c(1,2,3))
```




```
bptest(mod3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 24.672, df = 13, p-value = 0.02548
```

```
ncvTest(mod3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.03345743    Df = 1     p = 0.8548657
```

```
shapiro.test(mod3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod3$residuals
## W = 0.97238, p-value = 0.05443
```

- **CLM 1: Linearity** - Additional variables *prbpris*, *pctymle* and *avgsen* are linear with *log.crmrte*, while *taxpc* and *log.wmfg* show a more parabolic relationship. This is problematic and taking a log transformation of *wmfg*, while aiding interpretation, does little to address linearity. *taxpc* would be much closer to linear if it weren't for one outlier variable (county 55, as discussed in the EDA).
- **CLM 2: Random Sampling** - met with Model 1
- **CLM 3: Multicollinearity** - Several variables (*log.polpc*, *pctmin80* and *west*) show VIF of over 2, which is still "moderate" but not ideal.
- **CLM 4: Zero Conditional Mean** - Residuals vs. Fitted values shows a mean close to zero, albeit with some movement by a low outlier. The additional residuals have brought the mean closer to zero than previous models.
- **CLM 5: Homoskedasticity** - Both the BP test and ncvTest show heteroskedasticity, which Scale-Location plot confirms. We use robust Standard Errors.
- **CLM 6: Normality of Residuals** - The addition of so many more residuals has strengthened their normality, as demonstrated by the QQ plot and higher p-value on Shapiro-Wilk.

Comparing models 1, 2, and 3

```
stargazer(mod1,mod2,mod3,report = "vct*", title = "Comparison of Models",
          add.lines=list(c("AIC", round(AIC(mod1), 1), round(AIC(mod2), 1), round(AIC(mod3), 1))),
          se=list(se.m1, se.m2, se.m3), omit.stat="f", header = FALSE,float = FALSE)
```

	Dependent variable:		
	log.crmrte		
	(1)	(2)	(3)
prbarr	−1.507	−1.191	−1.240
	t = −4.750***	t = −2.574**	t = −3.397***
prbconv	−0.541	−0.626	−0.619
	t = −5.547***	t = −4.941***	t = −5.153***
prbpris			0.258
			t = 0.472
avgsen			−0.011
			t = −0.755
log.polpc	0.472	0.551	0.610
	t = 3.027***	t = 4.723***	t = 3.520***
density	0.112	0.102	0.109
	t = 4.443***	t = 4.744***	t = 3.268***
taxpc			−0.003
			t = −0.452
log.wmfg			0.155
			t = 0.775
pctmin80		0.814	0.545
		t = 3.536***	t = 1.973**
pctymle			−1.286
			t = −0.801
mix		−1.142	−1.036
		t = −2.171**	t = −1.803*
west		−0.205	−0.350
		t = −2.094**	t = −2.728***
central			−0.168
			t = −2.067**
Constant	4.961	4.701	4.015
	t = 11.578***	t = 12.274***	t = 3.524***
AIC	59.5	16.8	20.6
Observations	89	89	89
R^2	0.668	0.808	0.825
Adjusted R^2	0.652	0.792	0.795
Residual Std. Error	0.325 (df = 84)	0.252 (df = 81)	0.250 (df = 75)
Note:			*p<0.1; **p<0.05; ***p<0.01

Statistical significance: Of the new variables added to model 3, only *central* shows statistical significance ($p < 0.05$). The addition of these variables reduces the statistical significance of *pctmin80* and *mix*, suggesting intercorrelation and an independent effect on these variables by one or more of the new ones. *west*, however, becomes slightly more significant, probably due to the addition of *central*, with which it has negative correlation. The other variables remain the same.

Practical significance: *prbarr* and *log.polpc* remain the most practically significant values, with the addition of so many variables actually increasing *log.polpc's* influence. The coefficients of the newly-added economic variables are very low, suggesting a relatively low effect, while *pctymle* has switched signs to negative from its positive correlation with *log.crmrte*, implying a problem of collinearity with another variable (likely *log.polpc*) that was not captured in our CLM analysis.

The additional variables have very slightly improved accuracy, as measured by an adjusted R^2 of .795 compared to .792 for model 2, while AIC is worse at 20.6 vs 16.8. A small difference to be sure, but it shows the addition of several non-significant variables has added complexity with diminished clarity, and weakened the model by undermining key CLM assumptions of linearity and multicollinearity. This model appears overfitted.

Model summary

Model 1 shows that the deterrence variables are a good place to start, but insufficient in describing the crime rate by themselves. Model 2, which incorporates more demographic variables as controls, is a good balance of accuracy and parsimony. Model 3 goes too far, using more variables than is necessary and running the risk of interpreting some of the noise of the residuals as meaningful information. We focus our omitted variable analysis and policy recommendations below on **Model 2**:

$$\begin{aligned} log(crime\ rate) = 4.701 - 1.191 \cdot Pr(arrest) - 0.626 \cdot Pr(conv) + 0.551 \cdot log(polpc) \\ + 0.102 \cdot density + 0.814 \cdot pctMinority - 1.142 \cdot mix - 0.205 \cdot west + u \end{aligned}$$

OMITTED VARIABLE ANALYSIS

There are a number of variables not included in this dataset that might influence the results. For any of the omitted variables, the bias in the measured variable can be represented as:

$$Bias(\widetilde{\beta}_{included}) = \hat{\beta}_{excluded} \cdot \widetilde{\delta}_{interaction}$$

Where $\hat{\beta}_{excluded}$ is the expected coefficient when regressing the output variable (*crmrte*) on the excluded variable and $\widetilde{\delta}_{interaction}$ is the expected coefficient of regressing the excluded variable on the included variable(s). The impact on the significance tests depends on the sign of the $\hat{\beta}_{included}$ coefficients(s) w.r.t the sign of theinteraction term.

It is critical in any analysis to consider what you *don't* know, and how that might affect your findings. Here is our best estimation of those variables, their size and direction of influence.

Deterrence

Mix of property crime vs. violent crime

A popular theory of policing and crime prevention proposed in the early 1980's was "broken windows," which suggested that active policing of even very minor offenses would "nip in the bud" any progression of incipient criminals to more serious crimes. It would therefore be useful to have a sense of the relative severity of crimes committed in each county, to see if there is a correlation between police presence and arrests and severity

of crime. This would be even more useful as time series data, to test if an increase in arrests for small infractions led to a decrease in arrests for violent crime.

positive bias: *polpc*, positive bias: *prbarr*, positive bias: *mix* Because property crime is far more prevalent than violent crime, one would expect counties with higher share of property crime to predict more crime in general. If we represent $ratiocrime = propcrime/totalcrime$, in terms of coefficients we expect $\hat{\beta}_{ratiocrime}$ to be positive. For the included variables, we estimate most impact on *polpc* and *prbarr*. For *polpc*, the $\widetilde{\delta}_{ratiocrime,polpc}$ is expected to be positive. Hence, with the omitted variable, $\hat{\beta}_{polpc}$ has a positive bias away from zero leading us to overestimate it's significance. For *prbarr*, the $\widetilde{\delta}_{ratiocrime,prbarr}$ is expected to be negative. Hence, with the omitted variable, $\hat{\beta}_{prbarr}$ also has a negative bias away from zero leading to an overestimation as well. In the absence of *ratiocrime*, *mix* is a decent proxy - given that face-to-face crimes are more likely to be violent than property crimes. Therefore we would expect these variables to be highly correlated, but the existence of face-to-face property crimes (ie, muggings) weakens the strength of *mix* as a proxy. Given the positive coefficient for $\hat{\beta}_{ratiocrime}$ and it's strong correlation with *mix* implying a positive $\widetilde{\delta}_{ratiocrime,mix}$, and resultant positive bias of *mix* towards zero it is likely that we are underestimating the effect of *mix* in our regression.

Close rate on cases

This would be a measure of both the police efficiency and of the judicial system - are they arresting the right people and closing cases? Theoretically a more efficient and accurate police force could have a stronger effect than a simply larger force. Also a higher rate of sentencing arrested criminals would instill greater respect in the law–as long as there was no evidence of systematically punishing innocent people.

positive bias: *polpc*, negative bias: *prbconv* If we model this omitted effect by the variable $closeRatePerMonth = casesClosedPerMonth/totalCases$, then we expect $\hat{\beta}_{closeRatePerMonth}$ to be negative, as a more effective system of law should predict lower crime. For the included variables, we estimate most impact on *polpc* and *prbconv*. For *polpc*, the $\widetilde{\delta}_{closeRatePerMonth,polpc}$ is expected to be negative, as a more efficient police would imply need for lesser police presence. Therefore with the omitted variable, *polpc* has a positive bias away from zero leading to overestimation. For *prbconv* we expect the $\widetilde{\delta}_{closeRatePerMonth,prbconv}$ to be positive. Hence *prbconv* has a negative bias towards zero and underestimating it's impact with this variable omitted. Although these biases may be weak given the similarity to variables in our model, we are likely overestimating the effect of *polpc* while underestimating *prbconv*.

Economics

Employment

Under the model that criminals weigh the potential gains from crime against the potential loss from getting caught, low employment should indicate a higher crime rate, as more people would be desperate to make ends meet and therefore more willing to break the law to do it. Therefore $\hat{\beta}_{employment}$ should be negative. However, prior research suggests that by itself, unemployment is not an effective predictor of crime, so we cannot expect the effect to be a strong one. **negative bias: *wmfg*, *wcon*, and *wtrd*** Low employment would hit the lower-paying jobs harder, so we would expect a positive correlation with sectors such as *wcon*, *wmfg*, and *wtrd* - as employment increases, these jobs should pay better. Therefore $\widetilde{\delta}_{employment,wmfg}$ should be positive. Combined with the negative $\hat{\beta}_{employment}$, this indicates a negative bias towards zero, suggesting we are underestimating these wage variables.

Income inequality as measured by the *Gini* index

While wages are a good measure of income, they do not provide information on the wealth **inequality**, which perhaps, is a better indicator of crime. The Gini index measures the level of wealth inequality with values ranging from 0-1. A Gini value of 0 indicates perfect distribution of wealth among people, and a value of 1 indicates perfect inequality - i.e. concentration of all the wealth with a single individual. Research has shown that a high Gini coefficient has a strong correlation with crime, both homicide and robbery.

positive bias: *pctmin80*, negative bias: *wmfg* Higher levels of income inequality (higher Gini) would generate higher levels of socio-economic imbalance, prompting an increase in *crmrte*. Therefore, we expect $\hat{\beta}_{gini}$ to be positive. Also, using conventional wisdom, we expect higher Gini rates as the *pctmin80* gets higher, leading to a positive $\widetilde{\delta}_{gini,pctmin80}$. This omission would then lead us to overestimate the impact of *pctmin80*. We also expect a negative correlation with the Gini index with the wage vars. Higher average wages should typically lead to less inequality in the income distribution. Hence, w.r.t our reference wage variable *wmfg*, $\widetilde{\delta}_{gini,wmfg}$ is expected to be negative. This would lead to a negative bias towards zero in its estimation and possibly lower p-values for *wmfg* in our inference procedure.

Demographics

Minority race as share of arrests

Are police disproportionately targeting and arresting minorities? Racial profiling, while controversial, is another philosophy of policing. In that theory, if populations at-risk of committing crimes (ie, minorities) believe that they are more likely to be caught and punished, they will be less likely to follow through. If the profiling theory is correct, then the more minority people you arrest, the fewer minorities will commit crimes. So in that case the direction should be negative. There is a counterargument, however, that the long-term effect of profiling is to alienate and engender hopelessness in a community, in which case one would expect the theory to backfire and crime to go up. There is also the possibility that heavily targeting one group effectively emboldens a different group, so net crime remains unchanged. Regardless of policy, since *pctmin80* is positively correlated with crime rate, we would expect $\hat{\beta}_{pctminarr}$ to be positive as well–in the absence of a specific policy or philosophy of policing, the sample of arrests should roughly mirror the overall population.

positive bias: *pctmin80*, negative bias: *mix* If the police are disproportionately arresting minorities–as is likely the case in North Caroline in 1987–then $\widetilde{\delta}_{pctminarr,pctmin80}$ will be positive and we are overestimating the effect of *pctmin80*–assigning to the population some of the influence that should really be borne by the policing. On the other hand, $\widetilde{\delta}_{pctminarr,mix}$ is likely negative, as disproportionately targeting a minority population means more arrests for small crimes, and hence a lower mix of face-to-face crime. By that logic we are underestimating the contribution of *mix*.

After-school program access

This was not a crime-related philosophy at the time, but recent research has shown that access to after-school programs from 3-6pm is an effective method to mitigate crime. A variable such as "average hours of afterschool available to youth 12-18", *avgAfterSchool* would help examine this variable further.

negative bias: *pctymle*, negative bias: *taxpc* If this theory is correct, then the more access a given county has to after school programs, the fewer crimes should be expected giving us a negative $\hat{\beta}_{avgAfterSchool}$. Access to after-school programs might be targeted more towards counties with higher % of young males causing a positive $\widetilde{\delta}_{avgAfterSchool,pctymle}$. Hence, the absence of this variable (negative bias away from zero) might cause us to overestimate the impact of *pctymle* on *crmrte*. Afterschool programs require funding, and are therefore more likely to be found in wealthy areas–or at least more heavily taxed areas–so we can expect a positive $\widetilde{\delta}_{avgAfterSchool,taxpc}$, causing us to overestimate any contribution from *taxpc* due to the negative bias away from zero.

CONCLUSION AND POLICY RECOMMENDATION

Judging by the data we have, the best approach for a campaign looking to lower crime is to **advocate for stronger policing, especially in urban areas.** The reasoning being that there is simply more crime in the cities, while there are not significantly more police per capita. As a result, more crimes go unpunished, and the perception of the potential risk for committing a crime is lowered. Similarly, steps to improve the *perception* of punishment appear to have a deterrent effect, in particular a higher probability of arrest or conviction would appear to predict less crime–although it may be sufficient simply to have more *visible* arrests and convictions.

To be more precise, if we consider *prbarr* as the percentage of reported crimes that result in an arrest, then an increase in that arrest rate of 10% would result in a 12.4% reduction in crime rate, all else being equal. Similarly, convicting 10% more of those arrested would diminish crime by over 6%. These are the most effective policy levers at our disposal for addressing crime, although it is important to note that there could be feedback effects not captured by our model, and simply assigning quotas that are 10% higher to cops and prosecutors without consideration of unforeseen effects might backfire.

In determining where best to target these policies, it is clear that crime is a bigger problem in denser areas–ie, cities. For every additional person per square mile, crime goes up nearly 11%. While denser areas should be the primary focus of a "tough on crime" policy, the data supports other ways to focus targeting. For example, western counties appear to have about 20% less crime than elsewhere, though we don't have the data to say where, geographically-speaking, the most crimes ARE. There appears to be a correlation between face-to-face crimes (violent crimes and hold-ups) and counties with a higher share of minority populations, so while focusing efforts on minority communities might not move the needle much on overall crime rate, it might help reduce some of the more visible, scarier crimes. More research should be done to see what sort of crimes are more prevalent in higher-diversity neighborhoods, and what should be done to address this–it doesn't appear that these areas are under-policed, so rather than simply bulking up the police forces here, an approach based on outreach and education might be more appropriate.

It does not appear that economic measures such as wages or taxes have much effect on crime, so while they may feature in other campaign policies, they are not a focus on crime policy. There may be other economic measures, such as unemployment or wealth distribution, that could exert a more measurable effect.

As described in our Omitted Variable Analysis, there are several hidden factors that could be influencing our determinations–in particular those causing us to overestimate the importance of strong policing and the rate of arrests or convictions. We recommend more study in these areas, if possible, to confirm our recommendation of a "tough on crime" approach.