W203 Lab 3: Reducing Crime by Regression Analysis

Thomas Drage, Venkatesh Nagapudi, Miguel Jaime

December 2018

1. Introduction

This statistical investigation aims to understand the determinants of crime to suggest policies to the local government. The study is based upon development of causal models for crime rate, based on county level demographic and judicial data for 1987. We identified factors which modify the rate and extended this to the development of policy proposals for the incoming administration.

Contents

2. Review of Source Data

```
rm(list = ls())
crime_data = read.csv("crime_v2.csv")
objects(crime_data)
```

```
##  [1] "avgsen"   "central"  "county"   "crmrte"   "density"  "mix"
##  [7] "pctmin80" "pctymle"  "polpc"    "prbarr"   "prbconv"  "prbpris"
## [13] "taxpc"    "urban"    "wcon"     "west"     "wfed"     "wfir"
## [19] "wloc"     "wmfg"     "wser"     "wsta"     "wtrd"     "wtuc"
## [25] "year"
```

Overview of type and number of observations:

```
str(crime_data)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

There are 97 records in total.

Data Cleansing

Initially, we examined the data and removed values which were measurement or recording errors and ensured the formatting of the dataset was consistent and able to be processed.

1. *Empty Rows*: We noticed six rows with no data, all fields were "NA". Proceeded to remove these rows, since they are most likely an import error, and contain no data that could be analyzed.

```
crime_data[!complete.cases(crime_data), 1:3]
```

```
##    county year crmrte
## 92     NA   NA     NA
## 93     NA   NA     NA
## 94     NA   NA     NA
## 95     NA   NA     NA
```

```
## 96     NA   NA     NA
## 97     NA   NA     NA
```

```
crime_data_corr = na.omit(crime_data)
```

2. *Erroneous Import as Factor*: Due to the presence of a random back-tick character in the now removed "NA" rows at the end of the dataset, the variable "prbconv" was interpreted as a factor of levels - we can convert it back to numeric data with no loss.

```
crime_data_corr$prbconv_fix = as.numeric(as.character(crime_data_corr$prbconv))
summary(crime_data_corr$prbconv_fix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

3. *Analyzing Probability Value*: Probability values are greater than one in some cases for probability of arrest, and probability of conviction. Probability of prison time does not exhibit this behavior. Having an event with probability greater than one does not make sense: there cannot be a probability value higher than "certain to occur".

If we assume the probability of arrest is the number of arrests or number of convictions divided by the number of offenses, it is plausible that a given offense was committed by more than one individual. In these cases, there could be more than one arrest or conviction for a single offense. The variable, then, overestimates the probability of being arrested or convicted for a given offense. This issue could be present on all observations, not just on the ones where the justice system secured enough arrests or convictions to have the variable in question be greater than one.

Removing the variables would remove from our analysis certain counties would not fix the potential overestimation, it would simply remove from our analysis counties that seem to have better-than-average arrest or conviction rates. In light of this, we decided to include the rows in our analysis. We elected not top code these as one either, since we would be artificially lowering their values while leaving other overestimations intact.

```
crime_data_corr[crime_data_corr$prbarr > 1, "prbarr"]
```

```
## [1] 1.09091
```

```
crime_data_corr[crime_data_corr$prbconv_fix > 1, "prbconv_fix"]
```

```
##  [1] 1.48148 1.22561 1.23438 1.50000 1.35814 1.06897 1.01538 2.12121
##  [9] 1.67052 1.18293
```

```
crime_data_corr[crime_data_corr$prbpris > 1, "prbpris"]
```

```
## numeric(0)
```

4. *Duplicates*: There is a duplicate entry for county #193. We verified that all the data was identical, including the county number, and once confirmed we removed the observation from the dataset.

```
crime_data_corr[crime_data_corr$county == 193, 1:6]
```

```
##    county year    crmrte   prbarr     prbconv  prbpris
## 88    193   87 0.0235277 0.266055 0.588859022 0.423423
## 89    193   87 0.0235277 0.266055 0.588859022 0.423423
```

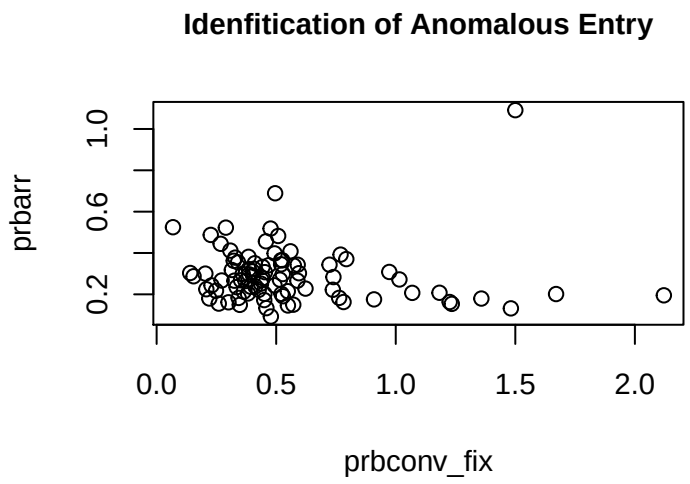
```
crime_data_corr2 = crime_data_corr[!duplicated(crime_data_corr), ]
```

5. *Density Values*: There is a density value of 0.0002 - this is approximately one person in an area the size of Alabama and presumably a measurement error. Therefore, we removed this record from the dataset.

```
good_density = (crime_data_corr2$density > 0.001)
crime_data_corr3 = subset(crime_data_corr2, good_density)
```

6. *Manual Entries*: During EDA, whilst investigating the relationship of probability of arrest ("prbarr") and probability of conviction ("prbconv_fix") in detail we noted an extreme outlier:

```
plot(prbarr ~ prbconv_fix, data = crime_data_corr3,
     cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9,
     main = "Idenfitication of Anomalous Entry")
```



The value at $prbconv = 1.5$ and $prbarr = 1.1$ is most likely a manual recording to cope with missing data; we note that all other probability values are specified to six significant figures, however, for this record probability of arrest ("prbarr") and probability of going to prison ("prbpris") were 1.5 and 0.5 respectively (see below). We also found that keeping this value resulted in a leverage point with high Cook's distance in our models. On the basis that this data is a poor manually entered estimate we chose to remove this row from the dataset.

```
crime_data_corr3[51,]
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
##    prbconv_fix
## 51         1.5
```

```
crime_data_corr4 = crime_data_corr3[-c(51),]
```

After cleansing we have 88 records, which we store as our master dataset:

```
crime_data_clean = crime_data_corr4
```

3. Identification of Key Variables

Dependent Variable

Crime rate ("crmrte") is the key dependent variable in this study and represents the number of crimes committed per person in each county.

Summarizing the variable we note a small range of fractional values, centred on a mean of approximately 3.5 crimes per hundred people in the year period.

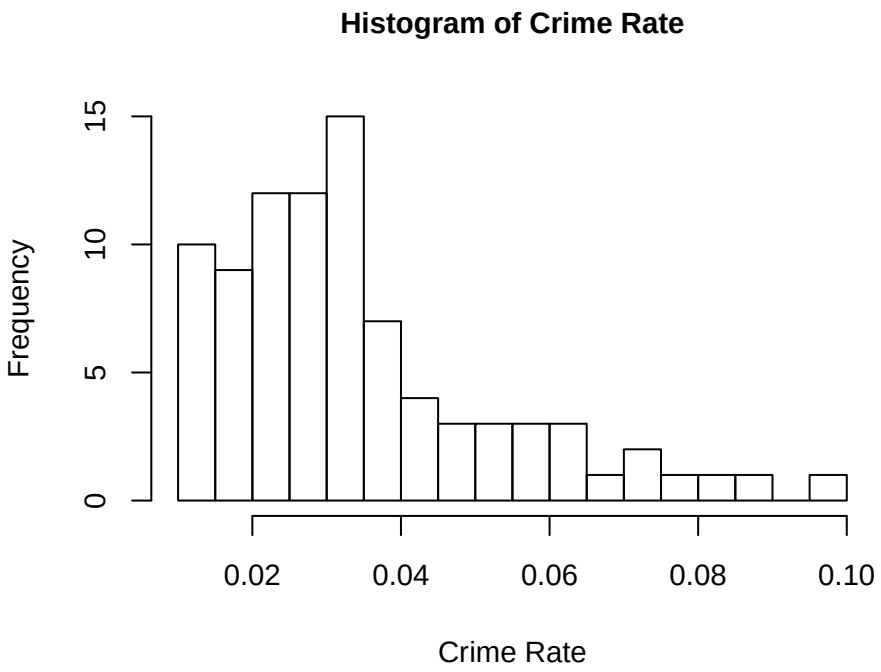
```
summary(crime_data_clean$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02201 0.03002 0.03405 0.04088 0.09897
```

The distribution of crime rate is right-skewed in this dataset.

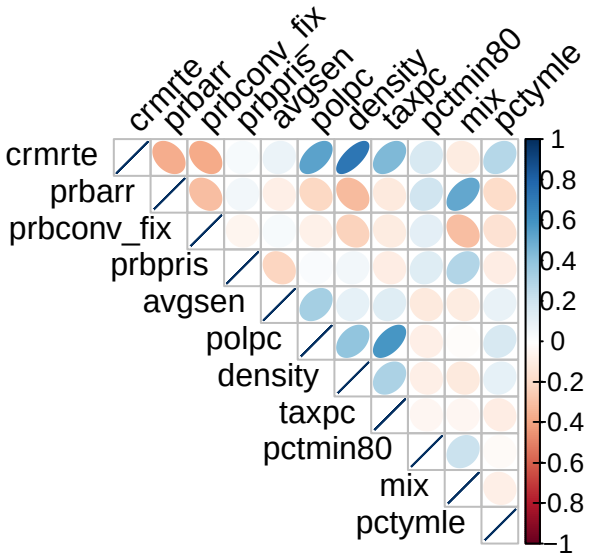
The number of observations (88) is large enough for modelling without concern for the skew noted in the variable. In the data transformation section we will determine if a transformation is needed for separate reasons.

```
hist(crime_data_clean$crmrte, breaks = 30,
     main = 'Histogram of Crime Rate',
     xlab = 'Crime Rate', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



We can take an immediate view of this dependents correlation with our set of available independent variables, firstly excluding the wage variables:

```
corrplot(cor(crime_data_clean[, c(3,4,26,6,7,8,9,10,14,24,25)]),
         method="ellipse", type="upper", tl.col="black", tl.srt=45)
```



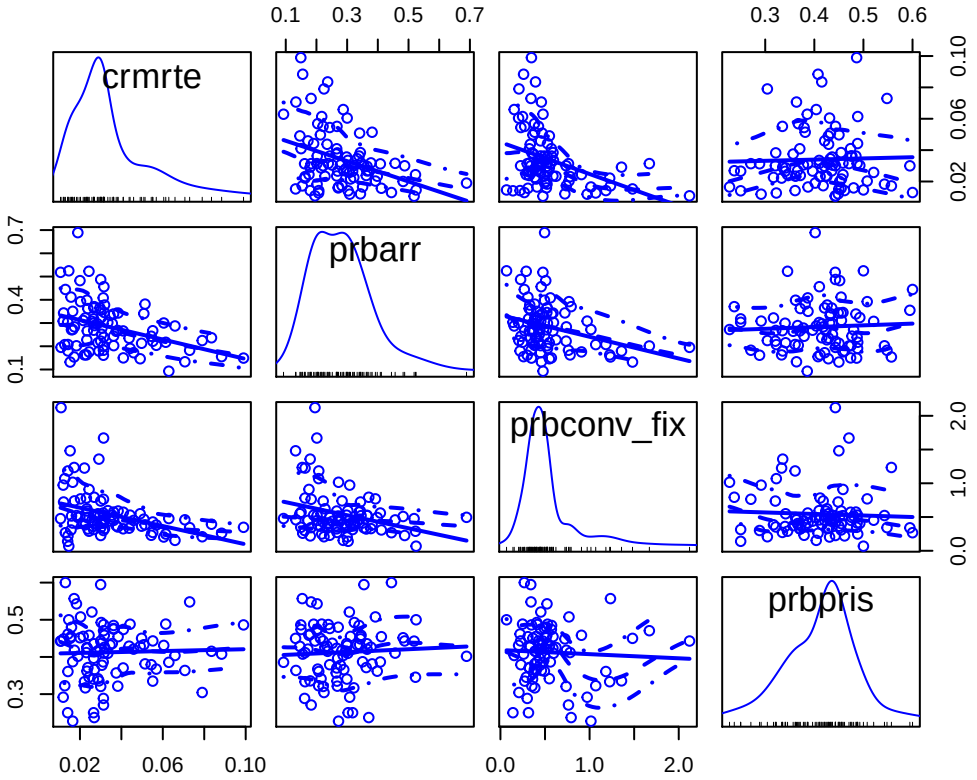
We note the cases with particularly high correlation with crmrte and examine them below as candidate regression variables. Note that due to the sample size and desire to maintain applicability of the Central Limit Theorem as well as limited definition of the differences, sizes or boundaries of these geographic regions, we have not partitioned our dataset using the "central"/"west"/"urban" variables.

Independent Variables - Judicial

1. Probability of Arrest ("prbarr")
2. Probability of Conviction ("prbconv")
3. Probability of Going to Prison ("prbpris")
4. Average Sentence ("avgsen")

It is likely that crime rate will be lower when the probability of getting arrested, convicted or going to prison is higher due to the deterrent effect. These variables are expected to have causal relationships with crime rate ("crmrte") and should reveal correlation, which we examine through a scatterplot matrix:

```
scatterplotMatrix(~ crmrte + prbarr +  prbconv_fix + prbpris, data=crime_data_clean)
```



The crime rate ("crmte") is negatively correlated with the probability of arrest ("prbarr") and probability of conviction ("prbconv_fix"), which is intuitive. There is perhaps a positive correlation to the probability of prison sentencing ("prbpris"), which is not intuitive, but the direction of the correlation is not clear from the dataset, therefore we excluded this from our key variable set.

In the correlation matrix above we noted little correlation for the average sentence ("avgsen")" with crime rate:

```
summary(crime_data_clean$avgsen)
```

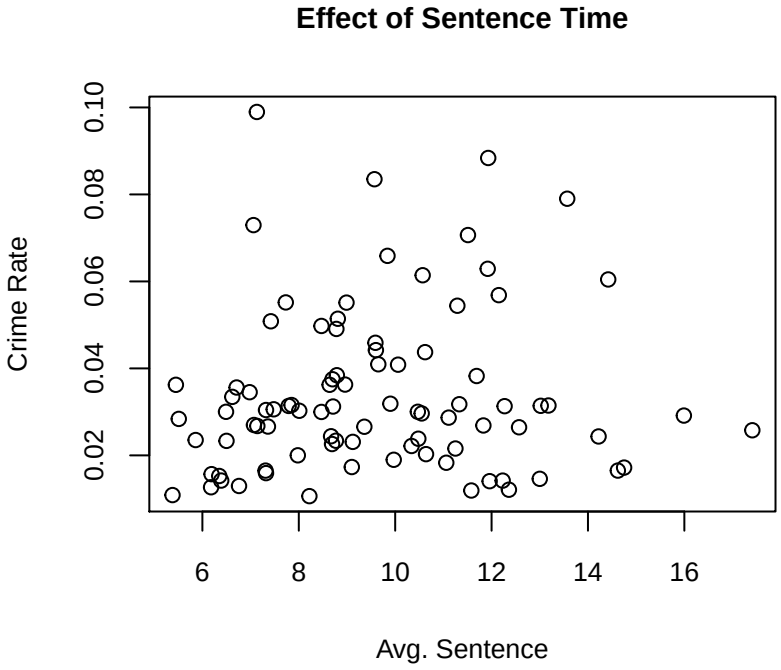
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.380   7.405   9.110   9.598  11.375  17.410
```

```
cor(crime_data_clean$crmrte,crime_data_clean$avgsen )
```

```
## [1] 0.08220147
```

There is a small correlation, but it is unclear as to whether there will be a causal relationship and which way it would be directed. For this reason we discount this variable from our analysis.

```
plot(crime_data_clean$avgsen, crime_data_clean$crmrte,
     cex.main = 0.9, cex.lab = 0.8, cex.axis = 0.8,
     ylab="Crime Rate", xlab="Avg. Sentence", main="Effect of Sentence Time")
```



Independent Variables - Demographic

1. Police per capita ("polpc")
2. Density ("density")
3. Tax revenue per capita ("taxpc")
4. Percentage of Young males ("pctymle")
5. Percentage of minorities ("pctmin80")

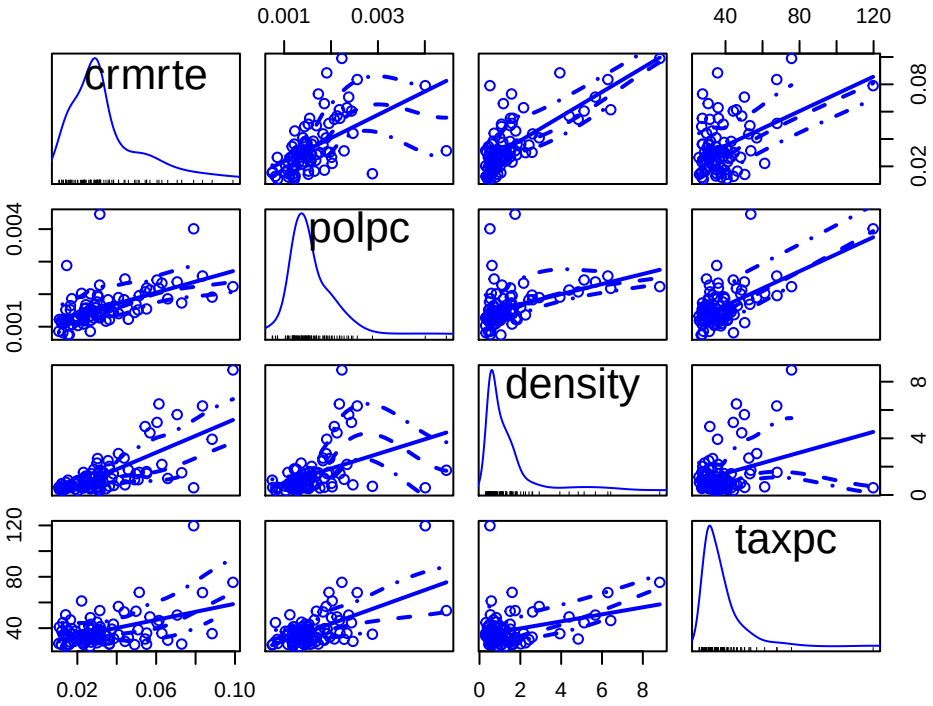
The second set of independent variables are demographic factors which may lead to changes in crime rate, typically in relation to the affluence of the county. Given that the data is collected at county level, these represent an average and any one county may contain a mix of areas (urban/suburban, wealthy/low-income) with corresponding variations in demographics and crimes, which are not captured in this dataset.

Policing / Density / Tax Revenue

Next, we examined the effect of police staffing, population density and tax revenue which we consider to be potentially related variables which describe the nature of the county.

Once again we investigate via scatterplot matrix:

```
scatterplotMatrix(~ crmrte + polpc +  density + taxpc, data=crime_data_clean)
```



Crime Rate is positively correlated to police per capita. We consider police staffing a lagging indicator: where crime rate is high, more police officers are deployed. There does not appear to be a logical causality where deployment of police leads to greater crime, but this is one of the strongest correlations revealed.

Looking at population density, there is a positive correlation between crime and density. This is not unexpected given high density housing is often associated with lower incomes and, in some cases, social issues. The density distribution is not normal, and might need to be transformed.

```
summary(crime_data_clean$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3006  0.5599  1.0008  1.4639  1.5762  8.8277
```

```
cor(crime_data_clean$crmrte,crime_data_clean$density)
```

```
## [1] 0.7247111
```

Tax revenue per capita ("taxpc") can be considered a proxy for the income level of a county. We assume that the higher the tax paid the more likely that the people are, on average, wealthier. Wealthier counties might be a more attractive target for property crime; though these effects might be tempered by a higher opportunity cost for committing crime, and higher likelihood of having security measures (such as alarms, gated communities, etc.)

```
summary(crime_data_clean$taxpc)
```

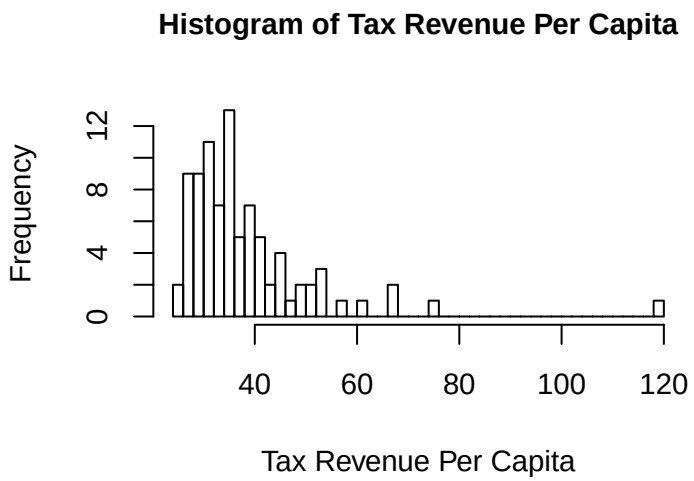
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.81   34.92   38.28   41.08  119.76
```

```
cor(crime_data_clean$crmrte, crime_data_clean$taxpc)
```

```
## [1] 0.4452837
```

We see a positive correlation between "taxpc" and crime rate. The distribution of "taxpc" is not optimal and we may need to examine outliers closely if this is to be used in modelling.

```r
hist(crime_data_clean$taxpc, breaks = 50,
     main = 'Histogram of Tax Revenue Per Capita',
     xlab = 'Tax Revenue Per Capita', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



Wage Data

A number of variables are provided with average wages in various sectors in each county. We can first examine these to see which might be correlated with crime rate or our other key variables.

We initially noted an anomaly with the service industry wage ("wser") variable, in that the correlation had the opposite sign. However, it was found that a single value greatly exceeded the normal range of values (see below). The figure is likely a result of decimal place recording error as no other remarkable features are found in this county and a service industry wage this high is abnormal, for the purpose of investigating wage data, this row was temporarily removed.

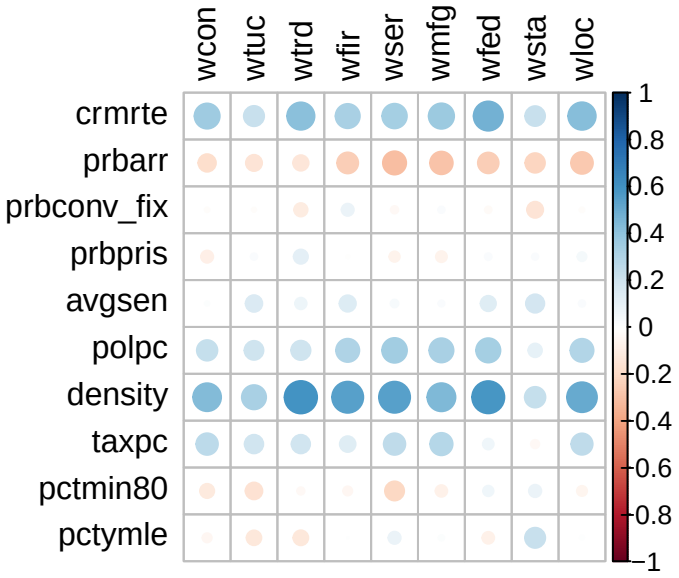
```r
summary(crime_data_clean$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   230.0   253.4   276.5   279.3  2177.1
```

```r
crime_data_clean3 = subset(crime_data_clean, crime_data_clean$wser < 1000)
```

We then examine a correlation matrix:

```r
corr_w = cor(crime_data_clean3[, c(3,4,26,6,7,8,9,10,14,25,15,16,17,18,19,20,21,22,23)])
corrplot(corr_w[seq(1,10),seq(11,19)], tl.col = "black")
```



The first observation is that all of the wage factors are correlated positively with crime rate. This is quite interesting as one might have assumed that areas where people are paid less would be poorer and would be prone to greater crime. This is apparently untrue, most likely because the comparative average wage in each sector is more of a function of the competitiveness of the economy in the county, evidenced by the strong correlation with density. E.g. a person in a particular industry may make more when employed in a city, which for other social reasons has a higher crime rate than a rural area. Incidentally, this data may also not capture the nature of poverty because it appears to be the average wage of the *employed* in this industry and gives no indication of the proportion of unemployed and hence poor or criminally employed people in the county.

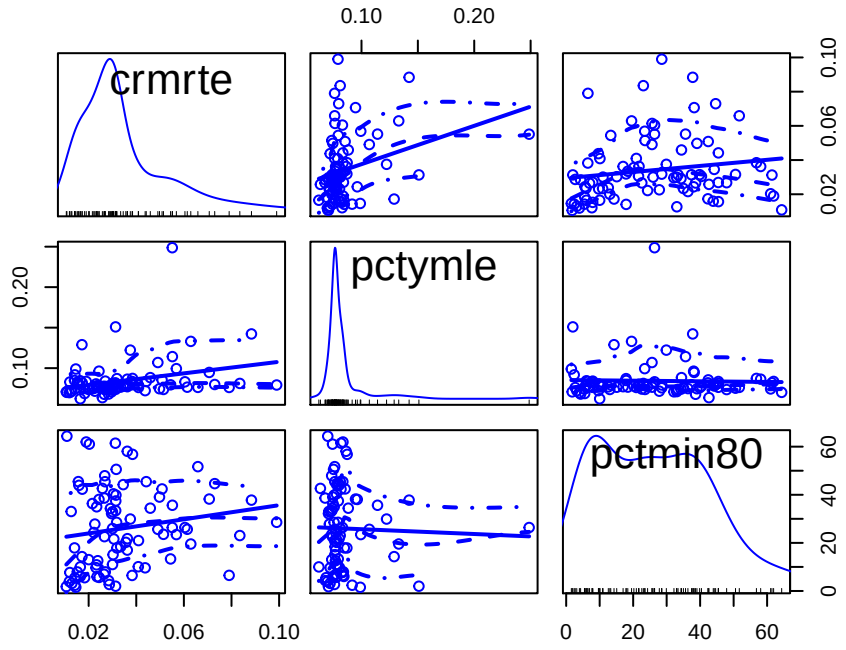
Another possibility is that crime is logged based on the county where it is committed, not the county where the offender reside. This would further assume that a non-negligible number of criminals would live in a county with a lower average wage, and go on to commit crime in a nearby, better-off county where victims would be more likely to possess valuables worth stealing. This effect would manifest primarily in property crimes, and the type of crime is unfortunately not available in this dataset.

An unfortunate correlation is that of the presence of minorities with the service wage - a high proportion of minority residents appears to push down the service wage, possibly due to competition for such jobs. However, a higher service wage does potentially decrease the probability of arrest. As this effect, while useful, appears to be confounded by the density effect, we do not choose to include such variables in our regression.

Minorities and Young Males

Here we examine the relationship between the proportion of young males ("pctymle") and the percentage of minority population ("pctmin80") with crime rate:

```
scatterplotMatrix(~ crmrte + pctymle +  pctmin80, data=crime_data_clean)
```



The crime rate is higher in places with a higher percentage of young males. The crime rate is also higher when the percentage of minority population is higher. Both variables seem to have non-ideal distributions.

Looking at the correlation between the variables:

```
summary(crime_data_clean$pctymle)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06216 0.07449 0.07779 0.08426 0.08355 0.24871
```

```
cor(crime_data_clean$crmrte,crime_data_clean$pctymle)
```

```
## [1] 0.2832397
```

```
summary(crime_data_clean$pctmin80)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.541  10.064  24.970  25.994  38.326  64.348
```

```
cor(crime_data_clean$crmrte,crime_data_clean$pctmin80)
```

```
## [1] 0.1619668
```

The correlation is weak in both cases.

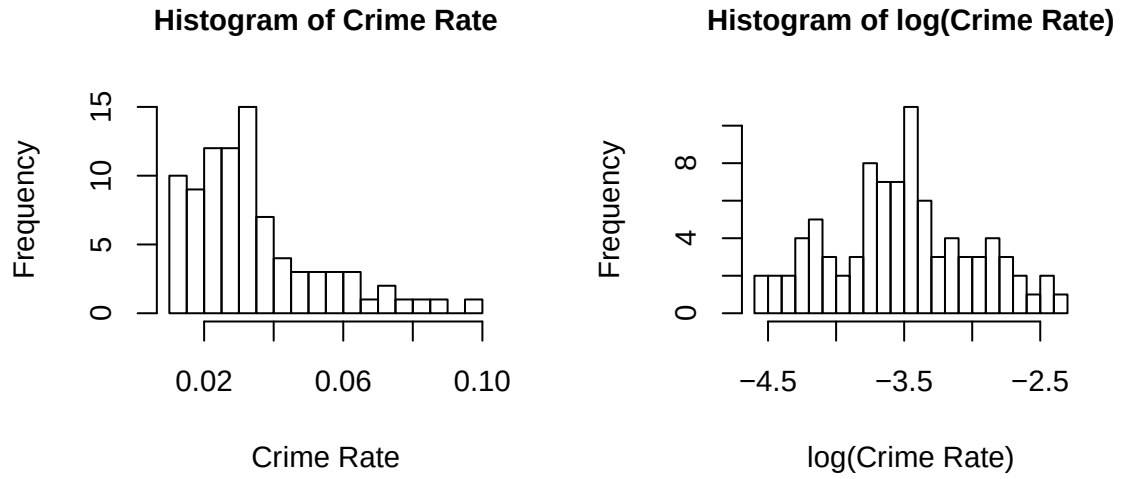
4. Data Transformation

Crime Rate

As discussed in section 2, our main variable of interest, crime rate, is measured in a way that results in small variations between values and a skewed distribution. The histogram below shows this distribution.

```
hist(crime_data_clean$crmrte, breaks = 30,
     main = 'Histogram of Crime Rate',
     xlab = 'Crime Rate', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)

crime_data_clean['log_crmrte'] = log(crime_data_clean$crmrte)
hist(crime_data_clean$log_crmrte, breaks = 30,
     main = 'Histogram of log(Crime Rate)',
     xlab = 'log(Crime Rate)', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



We applied a logarithmic transformation, as shown above to the variable which addresses both issues well.

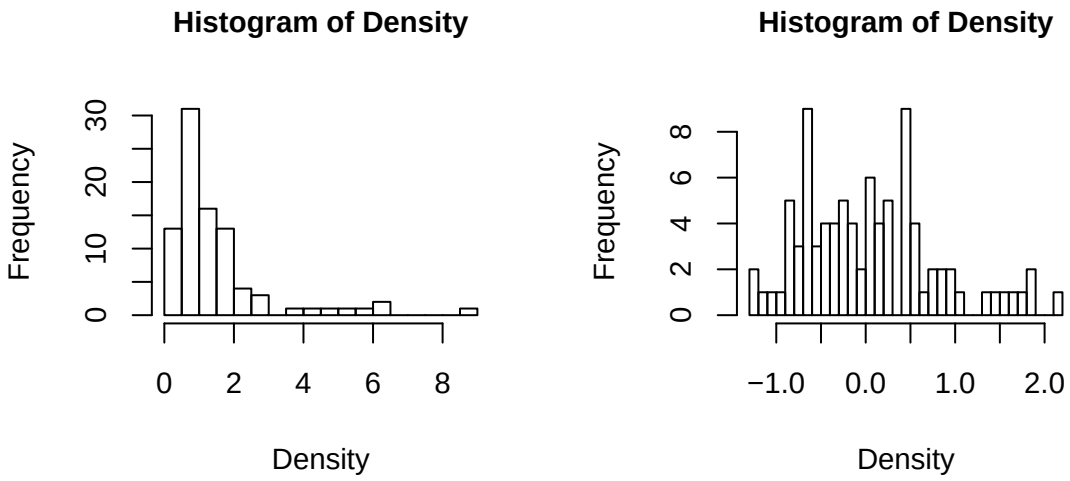
This transformation will change our interpretation, since the model coefficients will represent percentage changes for crime rate. Given the intended usage in reducing this rate and small values of the variable in its original units, this change will make the results easier to interpret.

Density

Density is right-skewed, however, the variable becomes more normal if we apply a log transformation, as shown below.

```
hist(crime_data_clean$density, breaks = 30,
     main = 'Histogram of Density',
     xlab = 'Density', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)

hist(log(crime_data_clean$density), breaks = 30,
     main = 'Histogram of Density',
     xlab = 'Density', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



Previously we noted that this variable has high correlation with our target variable, which increases slightly with the log transformation. The effect of removing this non-linearity is quite visible in the plot below.

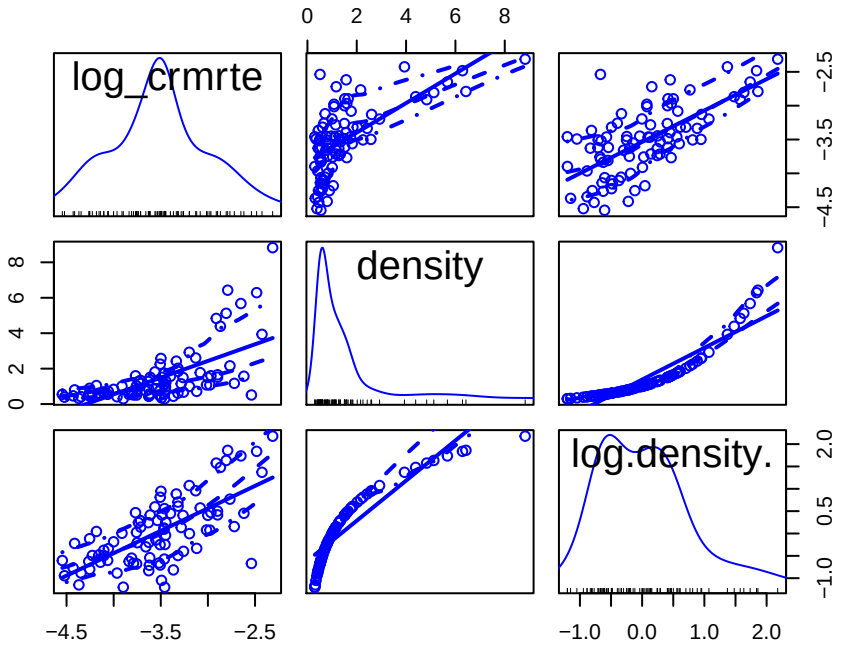
```
cor(crime_data_clean$log_crmrte,crime_data_clean$density)
```

```
## [1] 0.6405616
```

```
cor(crime_data_clean$log_crmrte,log(crime_data_clean$density))
```

```
## [1] 0.6828645
```

```
scatterplotMatrix(~ log_crmrte + density + log(density), data = crime_data_clean)
```



For this reason, we establish a transformed variable for use in our analysis:

```
crime_data_clean['log_density'] = log(crime_data_clean$density)
```

Police per Capita

The normality and correlation of the police per capita ("polpc") variable also benefit from logarithmic transformation, with an improvement in normality and correlation. The plot below shows an improvement in the linearity with the logarithm of crime rate.

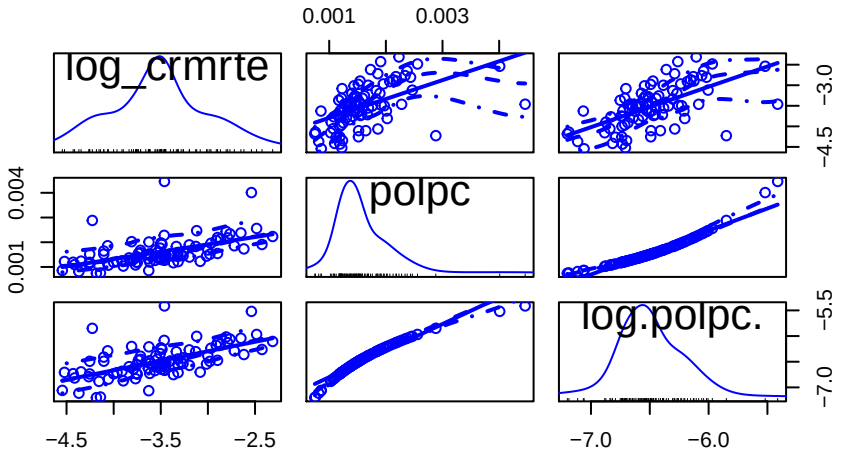
```
cor(crime_data_clean$log_crmrte,crime_data_clean$polpc)
```

```
## [1] 0.5225159
```

```
cor(crime_data_clean$log_crmrte,log(crime_data_clean$polpc))
```

```
## [1] 0.6032183
```

```
scatterplotMatrix(~ log_crmrte + polpc + log(polpc), data = crime_data_clean)
```



We can define a new transformed variable, however we note that this quantity is still believe to be a lagging indicator and potentially unsuitable for causal modelling.

```
crime_data_clean['log_polpc'] = log(crime_data_clean$polpc)
```

5. Regression Modeling

Model 1 - minimal using the Judicial system variables only

```
model1 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$prbconv_fix
              )
model1$coefficients
```

```
##                   (Intercept)      crime_data_clean$prbarr
##                     -2.341403                    -2.508591
## crime_data_clean$prbconv_fix
##                     -0.851451
```

Our hypothesis underlying this simple model is that the crime rate is correlated with the efficiency of the justice system, all other demographic factors being approximately equal as justice deters and controls the proliferation of criminal activity. The negative coefficients above support this with the probability of arrest being a stronger contributor than the probability of conviction.

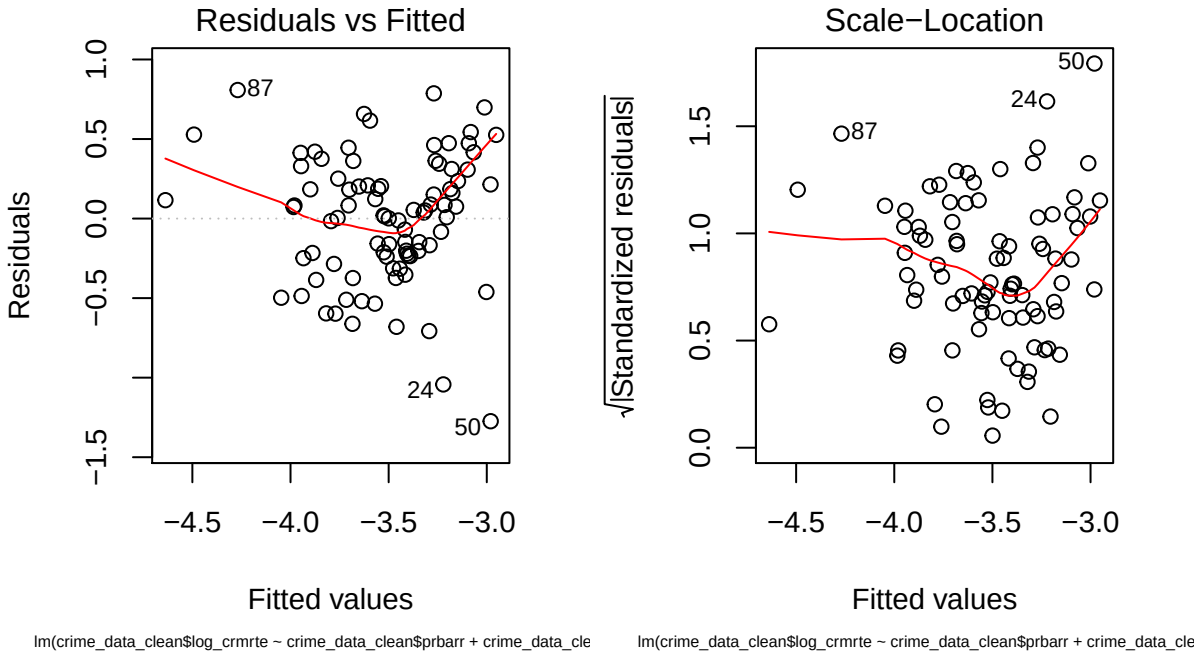
However, this is not the most complete model, and the R^2 value is relatively poor and reveals scope for a more sophisticated model:

```
summary(model1)$r.square
```

```
## [1] 0.4063575
```

We can then plot diagnostics for Model 1 to evaluate OLS assumptions:

```
plot(model1, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 1)
plot(model1, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 3)
```



We do note two issues:

- *MLR4 Zero Conditional Mean*: The fitted vs. residuals plot above shows a violation of zero-conditional mean for this model. This suggests some non-linearity with our chosen independent

variables.

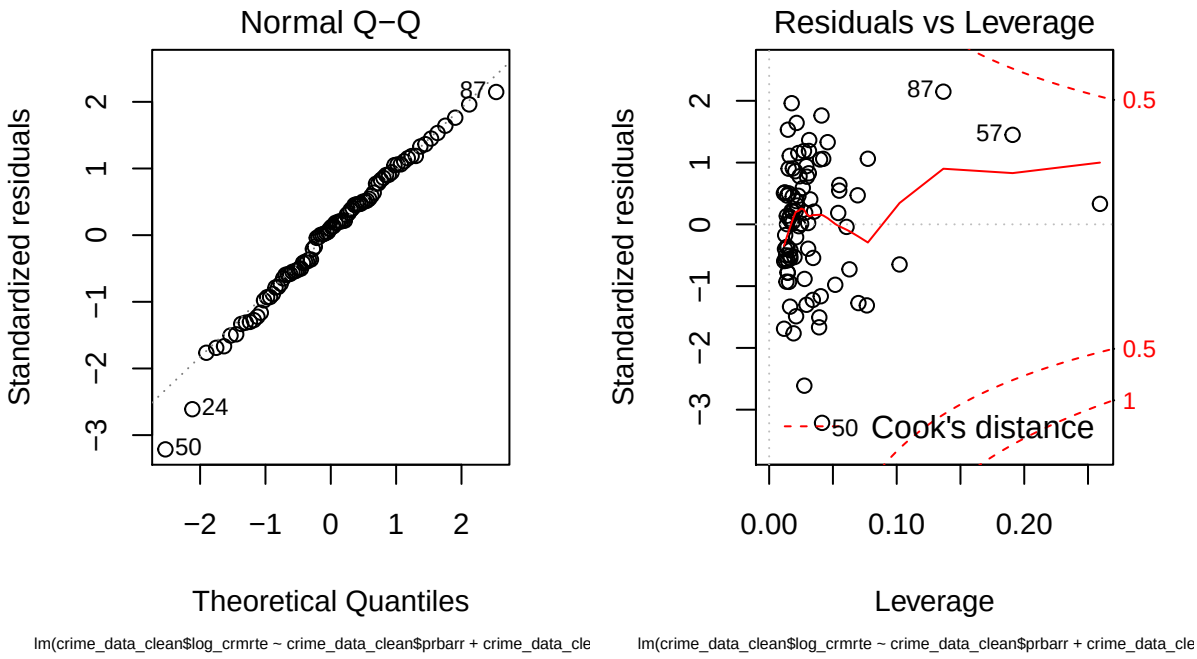
- *MLR5 Homeskedasticity*: This model is heteroskedastic. We confirm this using the Breusch-Pagan test and note marginal confirmation. For this reason, we will use robust standard errors going forward.

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 1.2058, df = 2, p-value = 0.5472
```

We check the other assumptions anyway in order to gain insight into issues with this variable selection:

```
plot(model1, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 2)
plot(model1, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 5)
```



- *MLR6 Normality*: The Q-Q plot indicates a good normality of the residuals.

There are no points with Cook's distance > 1 indicating that we removed all significant outliers earlier on in the analysis.

```
model1 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$prbconv_fix
              )
model1$coefficients
```

```
##                   (Intercept)      crime_data_clean$prbarr
##                     -2.341403                    -2.508591
## crime_data_clean$prbconv_fix
##                     -0.851451
```

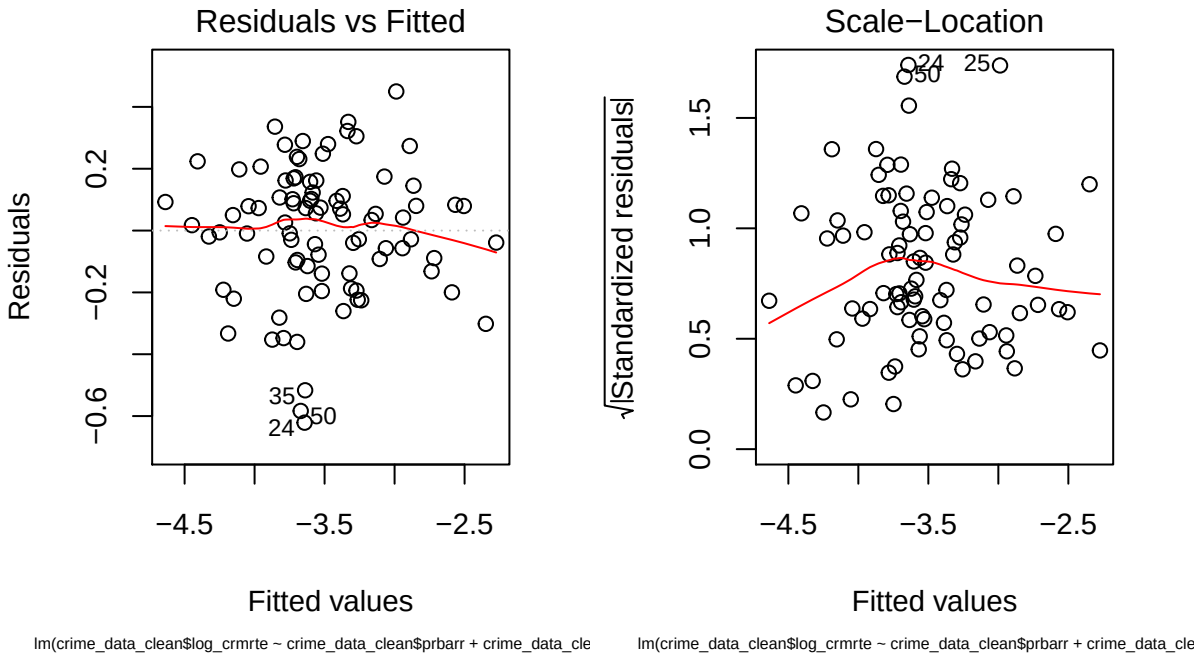
Model 3 - using judicial and demographic system variables

Model 3 is a more elaborate model that takes into account both judicial and demographic system variables to come up with a better causal explanation of crime rate. In this model, we included all meaningful variables. We decided to leave out wage-related variables since we did not find them to be relevant to our analysis.

```
model3 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$prbconv_fix +
              crime_data_clean$prbpris +
              crime_data_clean$avgsen +
              crime_data_clean$log_polpc +
              crime_data_clean$log_density +
              crime_data_clean$taxpc +
              crime_data_clean$pctmin80 +
              crime_data_clean$pctymle)
```

We can then plot Model 3 to evaluate OLS assumptions:

```
plot(model3, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 1)
plot(model3, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 3)
```



With this model, we note significant improvement:

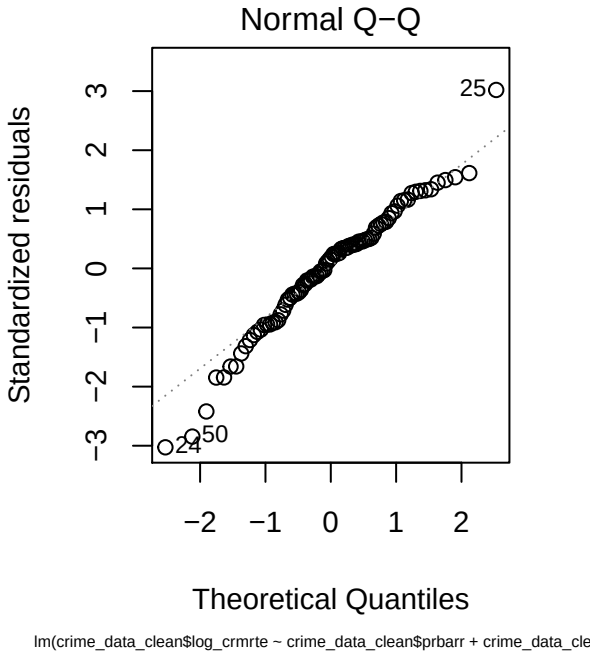
- *MLR4 Zero Conditional Mean*: The inclusion of more explanatory variables improves the mean residual, making it closer to zero.
- *MLR5 Homoskedasticity*: This model appears to have improved the scale location plot and the Breusch-Pagan test does not reject homoskedasticity.

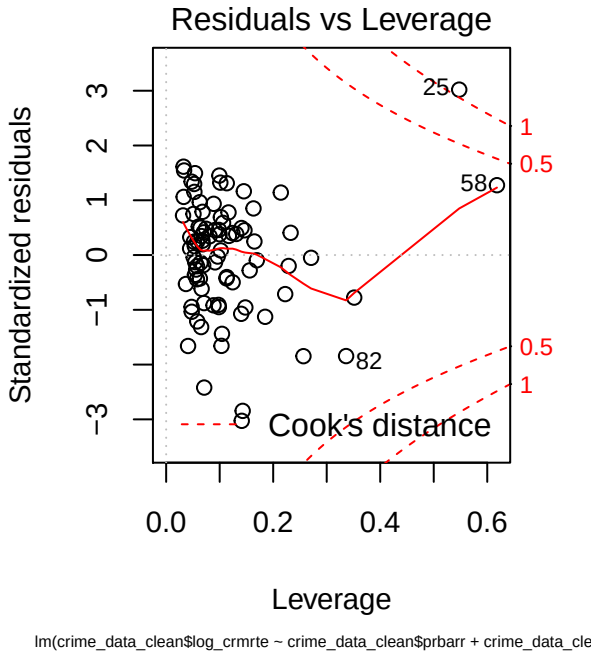
```
bptest(model3)
```

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

```
## BP = 20.592, df = 9, p-value = 0.01459
```

```
plot(model3, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 2)
plot(model3, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 5)
```





- *MLR6 Normality*: The Q-Q plot shows some deviations from normality which we confirm with a Shapiro-Wilks test. This suggests non-linearity in one or more of our model variables, most prominently in lower quartiles.

```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.97438, p-value = 0.07828
```

However, we note there are a couple of high influence outliers which may be negatively affecting the regression.

We can then display Model 3 using heteroskedastic robust standard errors:

```
# Using robust errors to compensate for heteroskedasticity
robust_se <- function(model) {
  cov <- vcovHC(model)
  sqrt(diag(cov))
}

robust_errors <- list(robust_se(model3))

stargazer(model1, model3,
          star.cutoffs =c(0.05,0.01, 0.001),
          se = robust_errors,
          type = 'latex',
          column.labels = c('Model 1','Model 3'),
          font.size = 'small',
```

```
          float = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Dec 11, 2018 - 7:51:08 AM

	Dependent variable:	
	log_crmrte	
	Model 1	Model 3
	(1)	(2)
prbarr	-2.509^{***}	-1.616^{***}
	(0.341)	(0.275)
prbconv_fix	-0.851^{***}	-0.609^{***}
	(0.128)	(0.082)
prbpris		−0.439
		(0.331)
avgsen		−0.010
		(0.010)
log_polpc		0.439^{***}
		(0.109)
log_density		0.281^{***}
		(0.040)
taxpc		0.003
		(0.002)
pctmin80		0.013^{***}
		(0.001)
pctymle		0.638
		(1.112)
Constant	-2.341^{*}	−0.121
	(1.027)	(0.874)
Observations	88	88
R^2	0.406	0.837
Adjusted R^2	0.392	0.818
Residual Std. Error	0.405 (df = 85)	0.221 (df = 78)
F Statistic	29.092^{***} (df = 2; 85)	44.486^{***} (df = 9; 78)
Note:		*p<0.05; **p<0.01; ***p<0.001

We can also evaluate the Akaike Information Criterion for both models to check goodness of fit relative to parsimony:

```
AIC(model1, model3)
```

```
##        df       AIC
## model1  4 95.526674
## model3 11 -4.187382
```

As we can see, Model 3 has significantly improved on the AIC, R2 and Residual SE, but there are some coefficients which are not statistically significant ("prbconv_fix", "polpc"). There is likely a more optimized model that has fewer coefficients that we can derive out of the Model 1 and Model 3 experiments above.

By looking at the standardized co-efficients, we can evaluate compare the effect of changes of each variable

on the crime rate:

```
lm.beta(model3)
```

```
##
## Call:
## lm(formula = crime_data_clean$log_crmrte ~ crime_data_clean$prbarr +
##     crime_data_clean$prbconv_fix + crime_data_clean$prbpris +
##     crime_data_clean$avgsen + crime_data_clean$log_polpc + crime_data_clean$log_density +
##     crime_data_clean$taxpc + crime_data_clean$pctmin80 + crime_data_clean$pctymle)
##
## Standardized Coefficients::
##                   (Intercept)      crime_data_clean$prbarr
##                    0.00000000                  -0.33135723
## crime_data_clean$prbconv_fix     crime_data_clean$prbpris
##                   -0.40156299                  -0.06428477
##      crime_data_clean$avgsen   crime_data_clean$log_polpc
##                   -0.05023913                   0.26909498
## crime_data_clean$log_density       crime_data_clean$taxpc
##                    0.41363685                   0.07881216
##    crime_data_clean$pctmin80     crime_data_clean$pctymle
##                    0.41902434                   0.02908422
```

Based on the above we may consider removing those with lower (e.g. <0.1) gain in addition to those which are not statistically significant.

Model 2 - with optimized Judicial and Demographic system variables

From the above models, it is clear that some of the variables added to the model such as density, police per capita ("polpc") and proportion of minorities ("pctmin80") show particularly strong contribution to the model. Whilst still statistically significant and one of our strongest gain variables, the "prbconv_fix" coefficient has decreased in magnitude in Model 3. This is likely due to correlation with density and complex relationships which are not revealed with the available variables (see omitted variable discussion below).

We previously noted that we do not believe police per capita ("polpc") to indicate causality, merely correlation and so we attempt the remove of this from the model. Examining standardized coefficients below we note that the removal of police per capita ("polpc") has increased the effect size of variables such as tax per capita ("taxpc") and proportion of young males ("pctymle") as well as density. We therefore consider this variable to have confounded these individual effects being a product of crime and affluence in the community. For this reason, as we seek a causal model for crime rate, we chose not to include this variable in our next model.

```
#everything in model3 except polpc
model3_no_polpc = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$prbconv_fix +
              crime_data_clean$prbpris +
              crime_data_clean$avgsen +
              #crime_data_clean$polpc +
              #crime_data_clean$log_polpc +
              crime_data_clean$log_density +
              crime_data_clean$taxpc +
              crime_data_clean$pctmin80 +
              crime_data_clean$pctymle)
lm.beta(model3_no_polpc)
```

```
##
## Call:
## lm(formula = crime_data_clean$log_crmrte ~ crime_data_clean$prbarr +
```

```
##     crime_data_clean$prbconv_fix + crime_data_clean$prbpris +
##     crime_data_clean$avgsen + crime_data_clean$log_density +
##     crime_data_clean$taxpc + crime_data_clean$pctmin80 + crime_data_clean$pctymle)
##
## Standardized Coefficients::
##                   (Intercept)      crime_data_clean$prbarr
##                    0.00000000                  -0.32874532
## crime_data_clean$prbconv_fix     crime_data_clean$prbpris
##                   -0.40543986                  -0.02977654
##      crime_data_clean$avgsen crime_data_clean$log_density
##                    0.01541994                   0.50732410
##       crime_data_clean$taxpc    crime_data_clean$pctmin80
##                    0.20308191                   0.42527334
##     crime_data_clean$pctymle
##                    0.07655393
```

Further to this, we have chosen to remove average sentence length ("avgsen"), probability of imprisonment ("prbpris") and proportion of young males ("pctymle") due to the high variance associated with the coefficient and low effect size. We therefore select the following for our second model, subject to further statistical testing:

1. prbarr
2. log(density)
3. pctmin80
4. prbconv_fix
5. taxpc

Creating Model 2 out of these variables:

```
model2_ver1 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$log_density +
              crime_data_clean$prbconv_fix +
              crime_data_clean$taxpc +
              crime_data_clean$pctmin80)
coeftest(model2_ver1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                  -3.3338864  0.2201317 -15.1450 < 2.2e-16 ***
## crime_data_clean$prbarr      -1.6990307  0.3118676  -5.4479 5.219e-07 ***
## crime_data_clean$log_density  0.3496533  0.0673339   5.1928 1.479e-06 ***
## crime_data_clean$prbconv_fix -0.6374485  0.1195927  -5.3302 8.464e-07 ***
## crime_data_clean$taxpc        0.0076726  0.0065951   1.1634     0.248
## crime_data_clean$pctmin80     0.0129687  0.0017813   7.2807 1.822e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

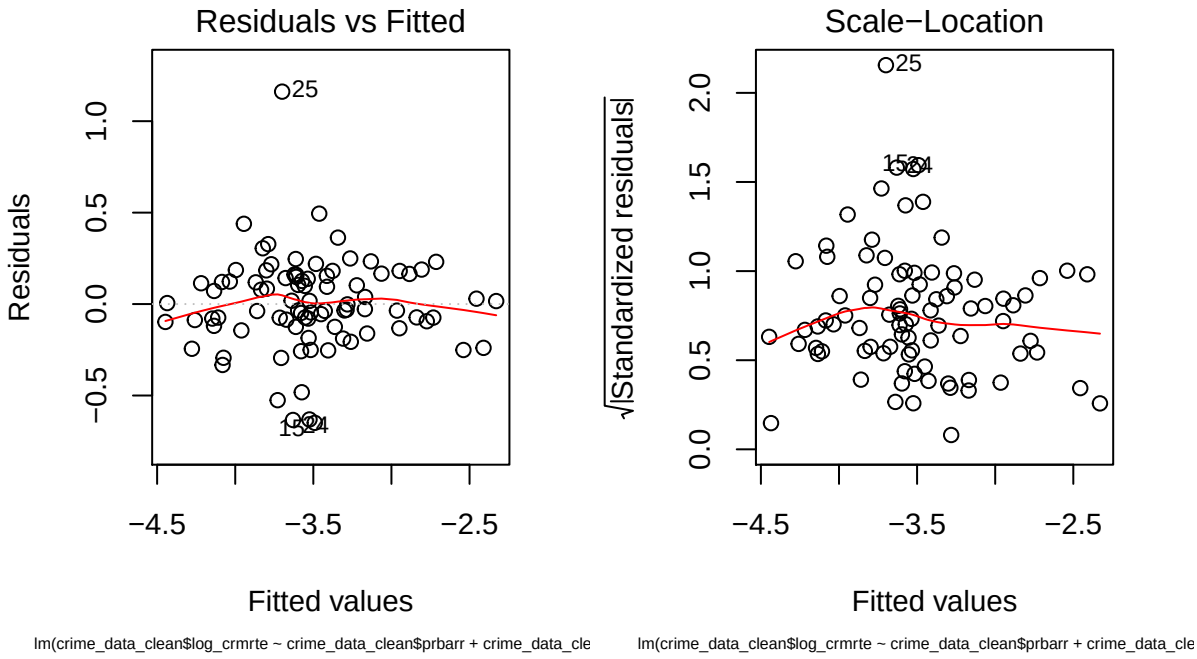
We can see that t-testing indicates that tax per capita ("taxpc") is not statistically significant, as a result we will remove it from the final model and retain the other four variables to get to a more parsimonious model.

```
model2 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$log_density +
              crime_data_clean$prbconv_fix +
              crime_data_clean$pctmin80)
```

Detailed Verification of OLS Assumptions

We can then plot diagnostics for Model 2 to evaluate OLS assumptions:

```
plot(model2, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 1)
plot(model2, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 3)
```

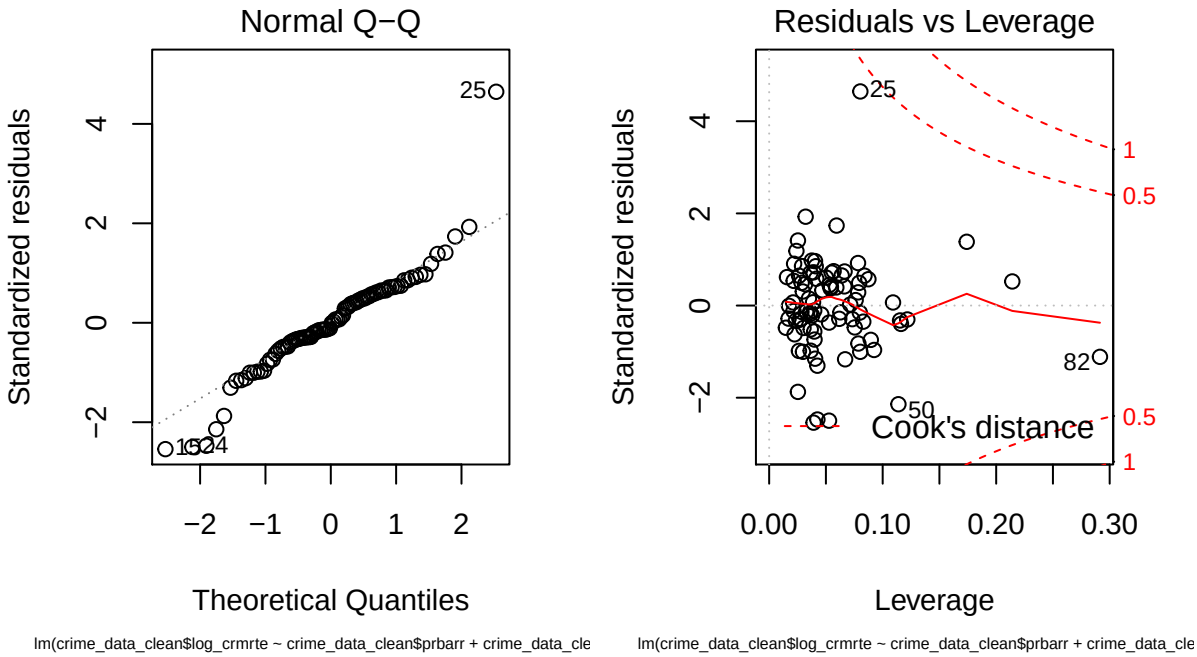


- *MLR1 Linearity*: The model is linear in the parameters given.
- *MLR2 Random Sampling*: The sampling process is not clear, but given that North Carolina currently has 100 counties, we assume most, if not all, counties are represented in the dataset.
- *MLR3 Colinearity*: Inspection of scatterplots above did not reveal any perfect co-linearity amongst the chosen variables.
- *MLR4 Zero Conditional Mean*: Further improvement and no violation of zero conditional mean in this model.
- *MLR5 Homoskedasticity*: This model appears to have further improved the scale location plot and the Breusch-Pagan test fails to reject homoskedasticity.

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 9.1483, df = 4, p-value = 0.0575
```

```
plot(model2, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 2)
plot(model2, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 5)
```



- *MLR6 Normality*: The Q-Q plot is more normal than Model 3. So the coefficients are more robust. We see that the p-values are all very significant unlike Model 3's p-values showing that the coefficients are more consistent. The Shapiro-Wilks test does not reject normality:

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.92235, p-value = 5.693e-05
```

Additionally, there are no outliers with high influence in this model specification.

Comparison of Models

We can now compare all three models. We used the "model3_no_polpc" so that we can do an "apples to apples"" comparison with Model 2 by comparing selection of causal variables to indicate the robustness of our final model.

```
robust_errors <- list(robust_se(model1),
                      robust_se(model2),
                      robust_se(model3))

stargazer(model1, model2, model3_no_polpc,
          star.cutoffs =c(0.05,0.01, 0.001),
          se = robust_errors,
          type = 'latex',
          column.labels = c('Model 1', 'Model 2', 'Model 3 No polpc'),
          font.size = 'small',
          float = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Dec 11, 2018 - 7:51:08 AM

	Dependent variable:		
	log_crmrte		
	Model 1	Model 2	Model 3 No polpc
	(1)	(2)	(3)
prbarr	-2.509^{***}	-1.805^{***}	-1.603^{***}
	(0.516)	(0.367)	(0.341)
log_density		0.365^{***}	0.345^{***}
		(0.060)	(0.068)
taxpc			0.008
			(0.006)
prbconv_fix	-0.851^{***}	-0.673^{***}	-0.615^{***}
	(0.166)	(0.132)	(0.128)
prbpris			-0.203
			(0.324)
avgsen			0.003
			(0.011)
pctmin80		0.013^{***}	0.013^{***}
		(0.002)	(0.002)
pctymle			1.680
			(1.813)
Constant	-2.341^{***}	-2.994^{***}	-3.473^{***}
	(0.209)	(0.202)	(1.027)
Observations	88	88	88
R^2	0.406	0.760	0.803
Adjusted R^2	0.392	0.748	0.783
Residual Std. Error	0.405 (df = 85)	0.261 (df = 83)	0.242 (df = 79)
F Statistic	29.092^{***} (df = 2; 85)	65.599^{***} (df = 4; 83)	40.334^{***} (df = 8; 79)
Note:			*p<0.05; **p<0.01; ***p<0.001

Additionally, we may compare the models in terms of Aikaike's Information Criterion:

```
AIC(model1, model2, model3_no_polpc)
```

```
##                df      AIC
## model1          4 95.52667
## model2          6 19.94140
## model3_no_polpc 10 10.31194
```

Let's also compare the BIC in these 3 cases:

```
BIC(model1, model2, model3_no_polpc)
```

```
##                df       BIC
## model1          4 105.43602
## model2          6  34.80542
## model3_no_polpc 10  35.08531
```

Model 2 shows significant improvement over Model1 with a better AIC, much better R^2 and lower Residual SE. We note that the F-test supports all three models, but our chosen model has a greater statistic, suggesting strong joint significance of our model. Additionally, the BIC (another goodness of fit measure which penalises addition of variables) for Model 2 is better than Model 3 as well which makes a very strong case for Model 2 as the superior model.

6. Discussion

Model Specification & Omitted Variables

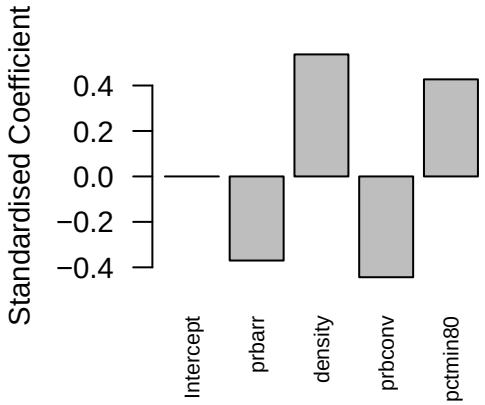
It is likely that crime rate will be heavily influenced by the following omitted variables.

Variable Category	
Demographics	The dataset contains little information about demographics. The "pctmin80" variable, for example, is based on dated information about minorities, and it does not break down into proportions for particular minorities. Updated, more granular information could help model crime rate. Other demographics variables, such as religious make up, could also be relevant to our analysis due to their potential impact on criminal activity.
Education Level	We expect higher education levels to be correlated with lower crime rates, and it could lead to interesting policy recommendations, but that information is unfortunately not available in the dataset.
Employment	We assume employment rate to be negatively correlated with the crime rate, but the dataset does not contain employment-related information. We considered using wage information as a proxy, under the reasoning that lower employment would result in upward pressure on wages. Scarcity of labor is, however, only one of many factors affecting wages, which means the proxy could be significantly biased. The absence of employment data might bias our models by attributing high crime rates to other factors, when lack of job prospects could be the actual cause.
Land Use	Crime rates might be correlated with land use planning practices. We expect, for example, areas with with vibrant nightlife and rural ares to experience different rates and types of crimes. This omitted variable would potentially bias our density dependence by increasing the coefficient or even increase the dependence of our model probability of arrest due to police presence in certain urban areas.
Age Distribution	The only variable provided was the percent of the population who are young males. It would be useful to have similar information about other age groups as well. We expect groups like young children and seniors to have a negative effect on crime rate. Both of these groups typically increase density, but decrease crime and we would expect their presence to bias the density variable negatively.
Crime Information	Different types of crime (violent crime, property crime, etc.) might be correlated with different variables. Omitting this variable could bias the probability of arrest, conviction and average sentence. Sentence length, for example, would be dependent on a measure of the proportion of non-violent crime and we may find sentence length is also negatively biased by this omission.
Wages	We presuppose higher wages affect the risk equation for committing crimes, but increasing the cost (if caught), and lowering the relative value of the rewards. The dataset does include certain wage information, though we found their effects to run counter to our expectations, and it had high correlation with density. Perhaps more information regarding wages (such as quartiles) and employment data could help understand this effect better.

Practical Significance of Causal Model

To investigate the practical significance of our model, we first examine the relative effects of our coefficients. Two variables, the density and proportion of minorities ("pctmin80") increase the crime rate whilst greater probability of arrest ("prbarr") and probability of conviction ("prbconv") decrease crime rate. In this parsimonious model there is not a wide distinction in the size the effect however density is the largest.

```
barplot(lm.beta(model2)$standardized.coefficients,
        las = 2, names.arg = c("Intercept","prbarr", "density", "prbconv", "pctmin80"),
        cex.main = 0.9, cex.lab = 0.9, cex.names = 0.7, cex.axis = 0.9,
        ylab = "Standardised Coefficient")
```



We note that a 1% increase in density is associated with a staggering 0.365% increase in crime rate. Given that the standard deviation of the density is some 104% of the mean, the density correlation appears to affect a substantial proportion of the 55% standard deviation we see in the crime rate. Wide density variation is due to the concentration of population in cities and it is evident that inner cities require the most attention in terms of crime reduction.

```
100*sd(crime_data_clean$density)/mean(crime_data_clean$density)
```

```
## [1] 104.2971
```

```
100*sd(crime_data_clean$crmrte)/mean(crime_data_clean$crmrte)
```

```
## [1] 55.04219
```

Based on our analysis, the probabilities of arrest and conviction help drive down crime rates. Increasing the probability of arrest by one unit is correlated with a 181% decrease in crime rate. Increasing the probability of conviction by one unit is correlated with only 61% decrease in crime rate, so it appears that arrests are a particularly significant deterrent. We might consider these variables an indicator of judicial efficiency and note that large improvements are possible with modest improvements in arrest count and that these might be easily achieved considering that standard deviation is some 10% already.

```
sd(crime_data_clean$prbarr)
```

```
## [1] 0.10648
```

The association with percentage of minorities gives approximately percent-for-percent increase in crime rate. Note that the nature of this variable should not influence a desire to change the value of the measure itself but gives a basis to examine how this correlation can be reduced to the point of insignificance in terms of effect size.

7. Conclusions

Based on these results of our regression, we have a better understanding of causal factors that lead to an increase or decrease in the crime rate. Based on our analysis, we make the following recommendations:

1. Fear of punishment

Our analysis shows that the higher the probabilities of arrest and conviction, the lower the crime rate. The fear of punishment is an effective tool to deter would-be criminals. As such, we recommend that policy makers increase awareness of the effectiveness of the judicial system to reduce crime. Bringing perpetrators to justice with an effective police force could significantly reduce crime rates.

2. Increased police presence in densely populated areas

Our analysis shows that in densely populated areas, the crime rate tends to be higher. There could be several factors at play here including demographics, socio-economic indicators, the size of the police force, earned wages, and others. In particular, policy makers should ensure that there is a well-staffed police force in densely populated areas to deter crime more effectively.

Another useful longer-term approach which is supported by our causal modelling is in the design or development of city areas. High density inner city ghettos might be avoided in favour of a greater distribution of the population.

3. Better education and employment

While the regression data does not include the effects of better education and employment on crime rate, there are some indications that the presence of more minorities and young males results in higher crime rates. Policy makers can focus on better education amongst youth and minorities so that the resulting higher employment rates lead to lower crime rates among that group. This would have the added benefit of increased tax revenues, which would provide another tangible benefit to the community.

8. Acknowledgements

We wish to thank Neha Kumar, Sid Jakkamreddy and Brian Musisi for their valuable feedback on our preliminary report. We also thank them for the opportunity to learn from their analysis. In particular, we understood the brevity and effectiveness of using correlation plots to describe dependencies between variables, and incorporated similar plots in our report as well.