W203 Lab 3: Reducing Crime by Regression Analysis

Thomas Drage, Venkatesh Nagapudi, Miguel Jaime

November 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.

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 of them.

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.We noticed six rows with no data, all fields were "NA" data. 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), ]
```

```
##    county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 92     NA   NA      NA     NA               NA     NA    NA       NA    NA
## 93     NA   NA      NA     NA               NA     NA    NA       NA    NA
## 94     NA   NA      NA     NA               NA     NA    NA       NA    NA
## 95     NA   NA      NA     NA               NA     NA    NA       NA    NA
## 96     NA   NA      NA     NA               NA     NA    NA       NA    NA
## 97     NA   NA      NA     NA        `      NA     NA    NA       NA    NA
##    west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta
## 92   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 93   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 94   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 95   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 96   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 97   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
##    wloc mix pctymle
## 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. 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. Probability values are greater than one in some cases for probability of arrest, and probability of conviciton. 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 dividied by the number of offenses, it is plausible that a given offense was committed by more than one individuak. 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 varable 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 numbers while leaving other overstimations 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)
```

```
# TODO: remove if edit looks good to rest of the team.

# There are 11 such values, which we removed as they indicate faulty data.

#sum(crime_data_corr$prbarr > 1)
#sum(crime_data_corr$prbconv_fix > 1)
#sum(crime_data_corr$prbpris > 1)
#good_prob_cond =
#   !((crime_data_corr$prbarr > 1) |
#   (crime_data_corr$prbconv_fix > 1) |
#   (crime_data_corr$prbpris > 1))
#crime_data_corr2 = subset (crime_data_corr, good_prob_cond)
#str(crime_data_corr2)
```

4. There is a duplicate entry for county #193. We verified that all the data was the same, including the county, 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. 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)
```

After cleansing we have 89 records, which we store as our master dataset.

```
crime_data_clean = crime_data_corr3
```

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.

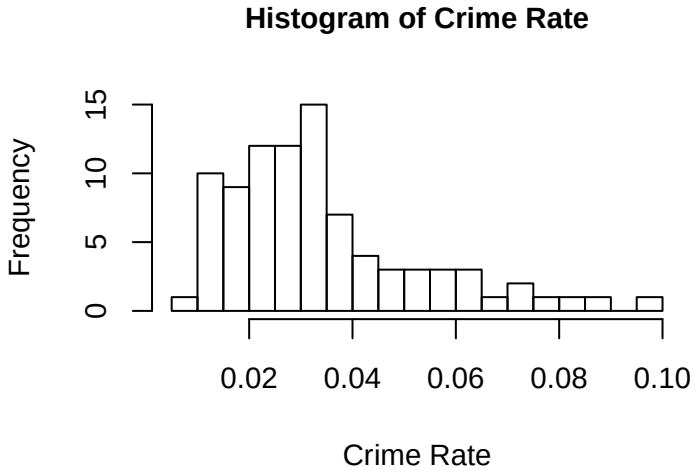
```
summary(crime_data_clean$crmrte)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.005533 0.021573 0.030018 0.033729 0.040857 0.098966
```

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

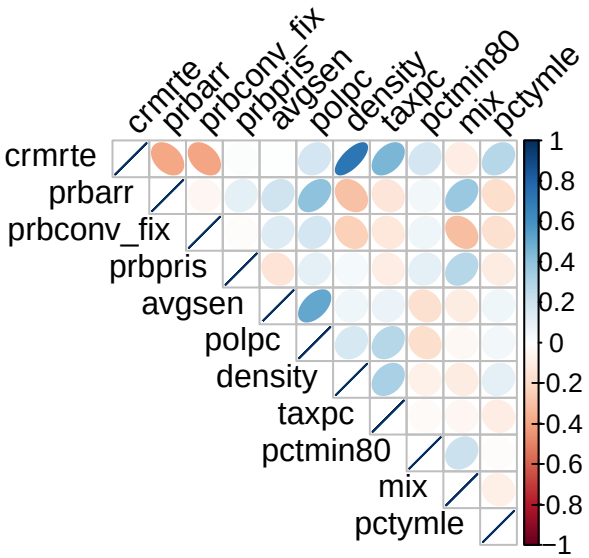
Even though the number of observations (89) is large enough for modelling without concern for the skew noted in the variable. In the data transformation section we will determine if a transformation is needed for separate reasons.

```
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 indepedent 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.
```



We note the cases with particularly high correlation with crmrte and examine them below as candidate regression variables.

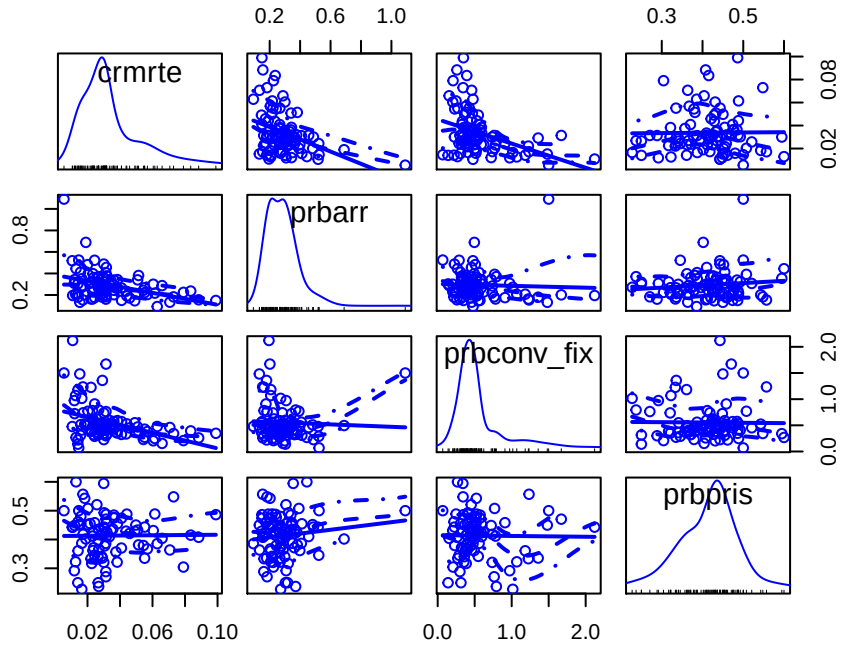
Independent Variables - Judicial

1. Probablity 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 crmrte is negatively correlated with prbarr and prbconv_fix, which is intuitive. There is perhaps a positive correlation to prbpris, the probability of prison sentencing, 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.

Anlyzing the average sentence (“avgsen”)“:

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

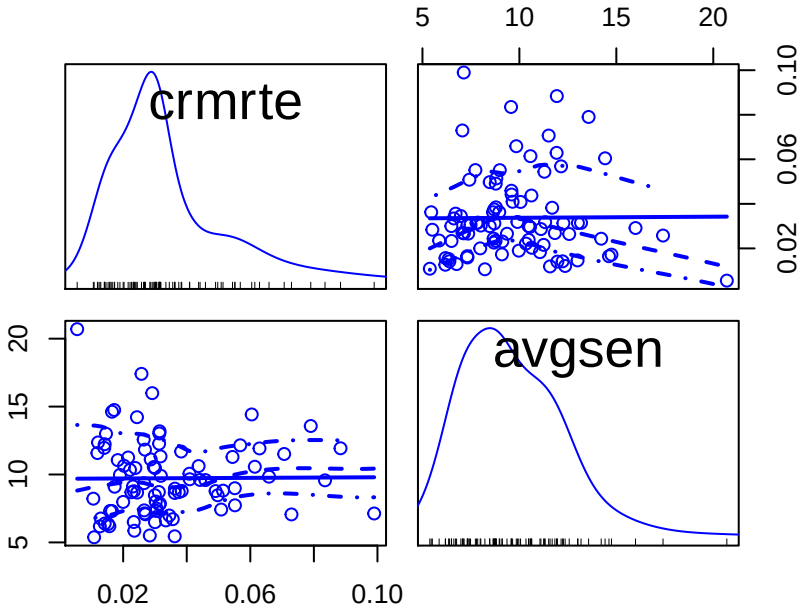
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.380   7.420   9.120   9.723  11.510  20.700
```

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

```
## [1] 0.007258477
```

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.

```
scatterplotMatrix(~ crmrte + avgsen, data=crime_data_clean)
```



Independent Variables - Demographic

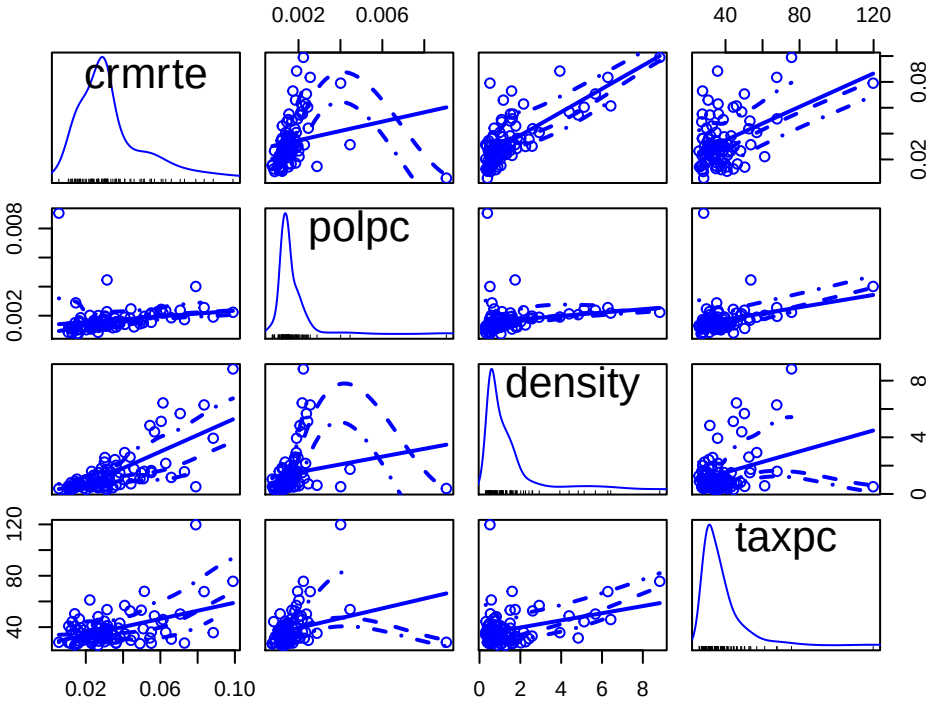
1. Police per capita ("polpc")
2. Density ("density")
3. Tax revenue per capita ("taxpc")
4. Percentage of Young males ("pctymle")
5. Percentage of minorities ("pctmin80")

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

We examined the effect of police staffing, population density and tax revenue:

```
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.

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  0.5479  0.9962  1.4518  1.5703  8.8277
```

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

```
## [1] 0.7253618
```

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, more wealthy. 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 higher security measures (such as alarms, gated communities, etc.)

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

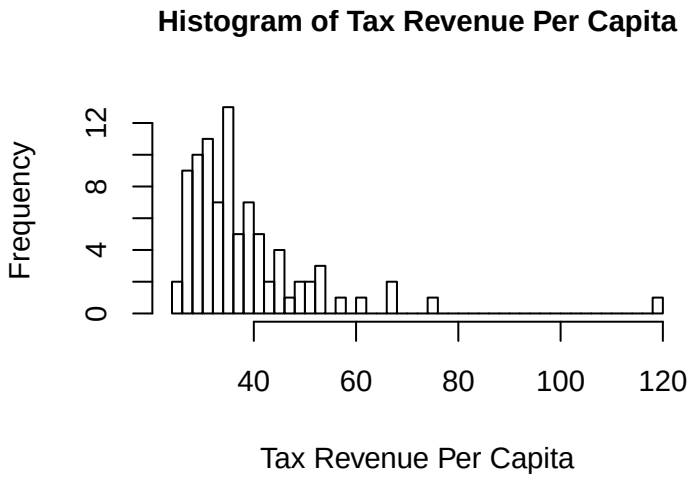
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.70   34.87   38.17   41.07  119.76
```

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

```
## [1] 0.4510738
```

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.

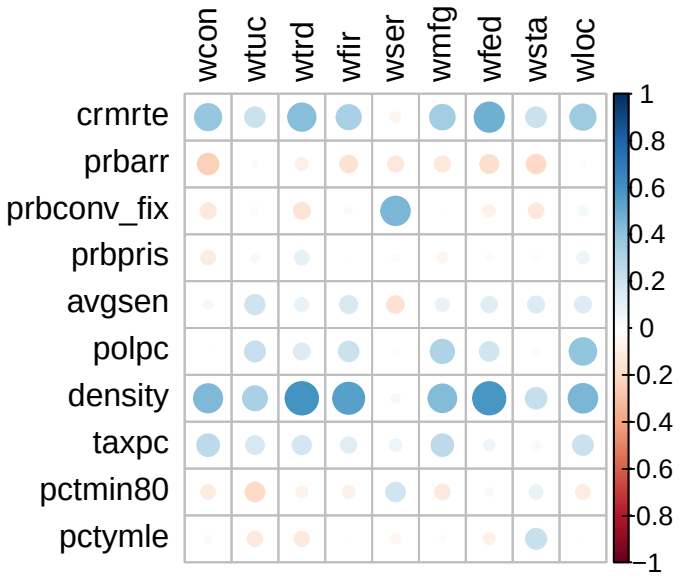
```
hist(crime_data_clean$taxpc, breaks = 50,
     main = 'Histogram of Tax Revenue Per Capita',
     xlab = 'Tax Revenue Per Capita', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



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:

```
corr_w = cor(crime_data_clean[, 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 postively 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 pooer 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-neglibigble 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 avaiable in this dataset.

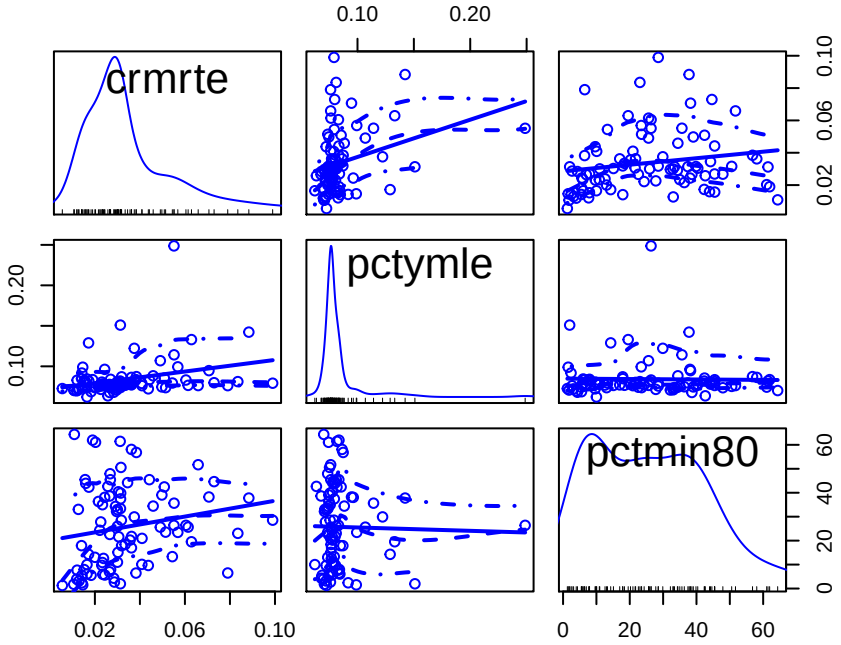
An unfortunate correlation is that of the presence of minorities with the service wage - a high proportion

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 0.07431 0.07771 0.08413 0.08354 0.24871
```

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

```
## [1] 0.2876448
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.284  10.005  24.312  25.716  38.223  64.348
```

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

```
## [1] 0.1825394
```

The correlation is weak in both cases.

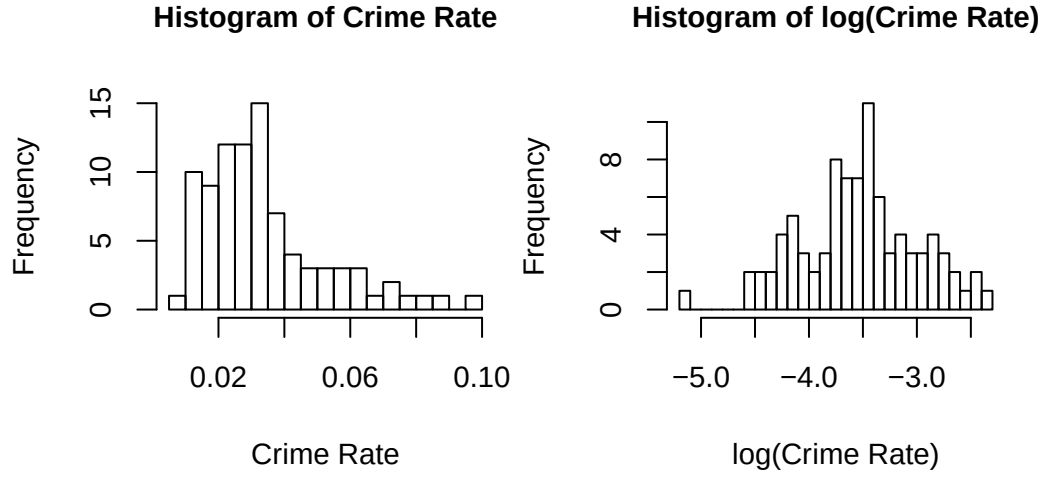
3. Data Transformation

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 log() transformation, as shown above to the variable which addresses both issues well.

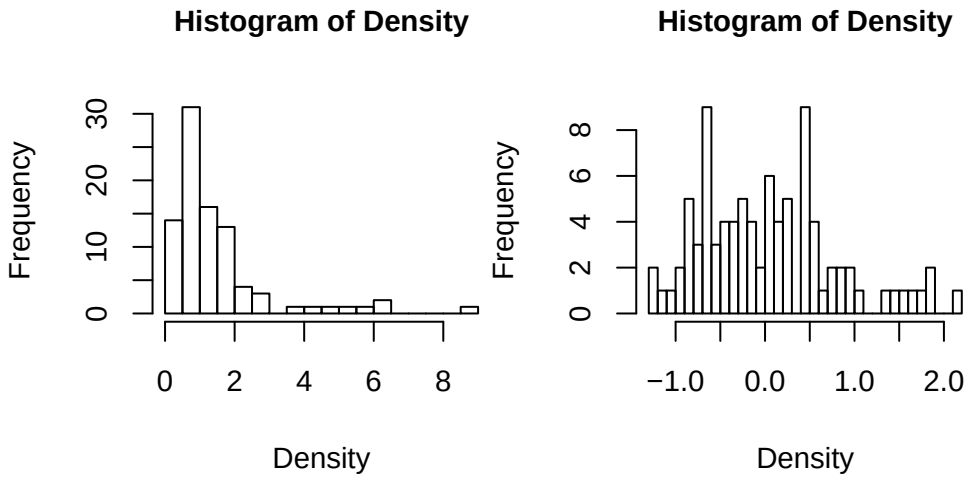
This transformation will change our interpretation, since the model coefficients will represent percentage changes for crime rate. Given the intended usage in reducing this rate and small values of the variable in its original units, this change will make the results easier to interpret.

Density

Density is right-skewed. Variable becomes more normal if we apply a log transformation.

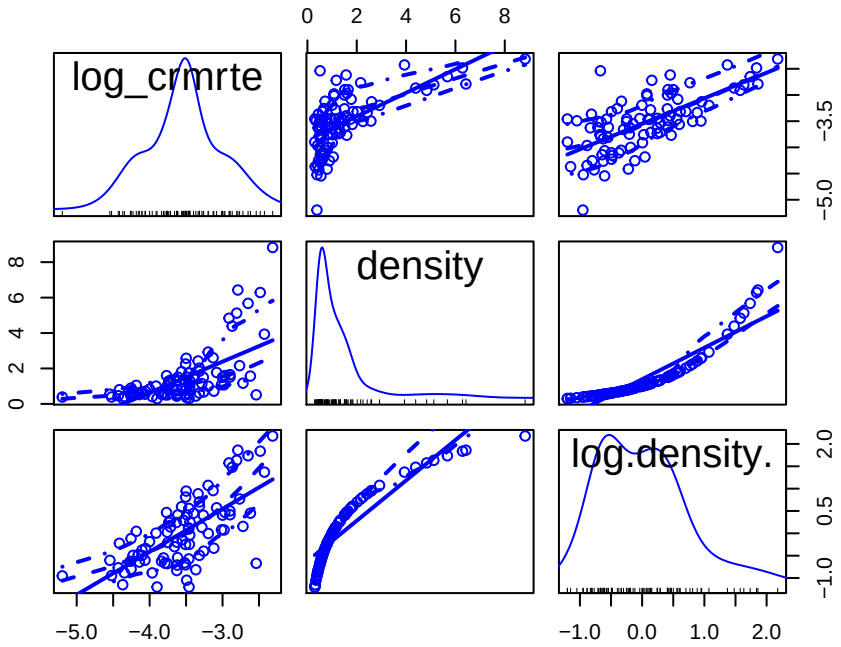
```
hist(crime_data_clean$density, breaks = 30,
     main = 'Histogram of Density',
     xlab = 'Density', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)

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)
```



The variable has high correlation with our target variable, which increases slightly with the log transformation.

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




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

```
## [1] 0.6282693
```

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

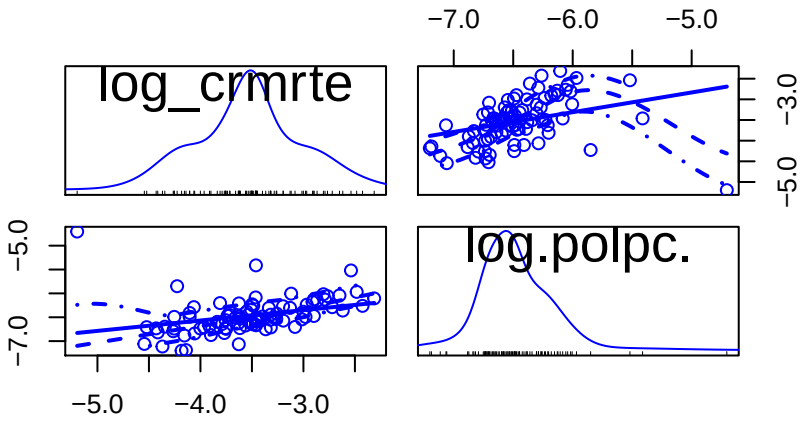
```
## [1] 0.6843266
```

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

Polpc

We can use a log transformed version of polpc potentially in our models even though it likely a lagging indicator.

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




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

```
## [1] 0.03327426
```

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

```
## [1] 0.3225051
```

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

4. Regression Modelling

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$prbarr
##                     -2.557993                     -1.929079
## crime_data_clean$prbconv_fix
##                     -0.742948
```

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.

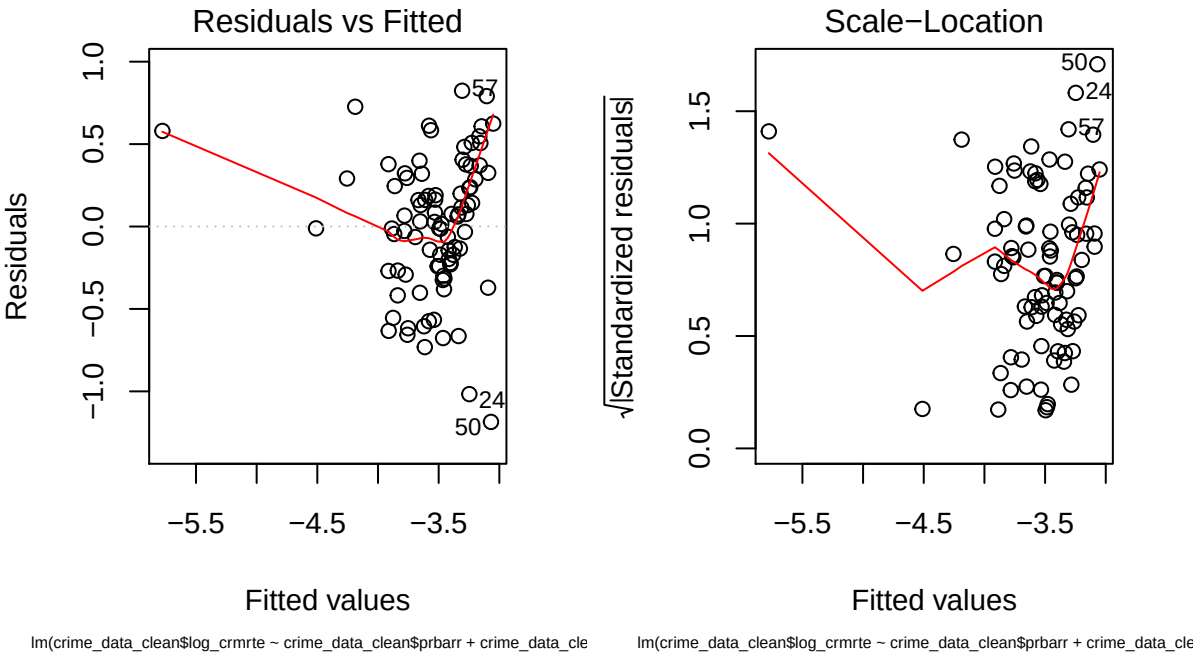
However, this is not the most complete model, and the R^2 value is relatively poor and reveals scope for a more sophisticated model:

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

```
## [1] 0.4440375
```

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)
```

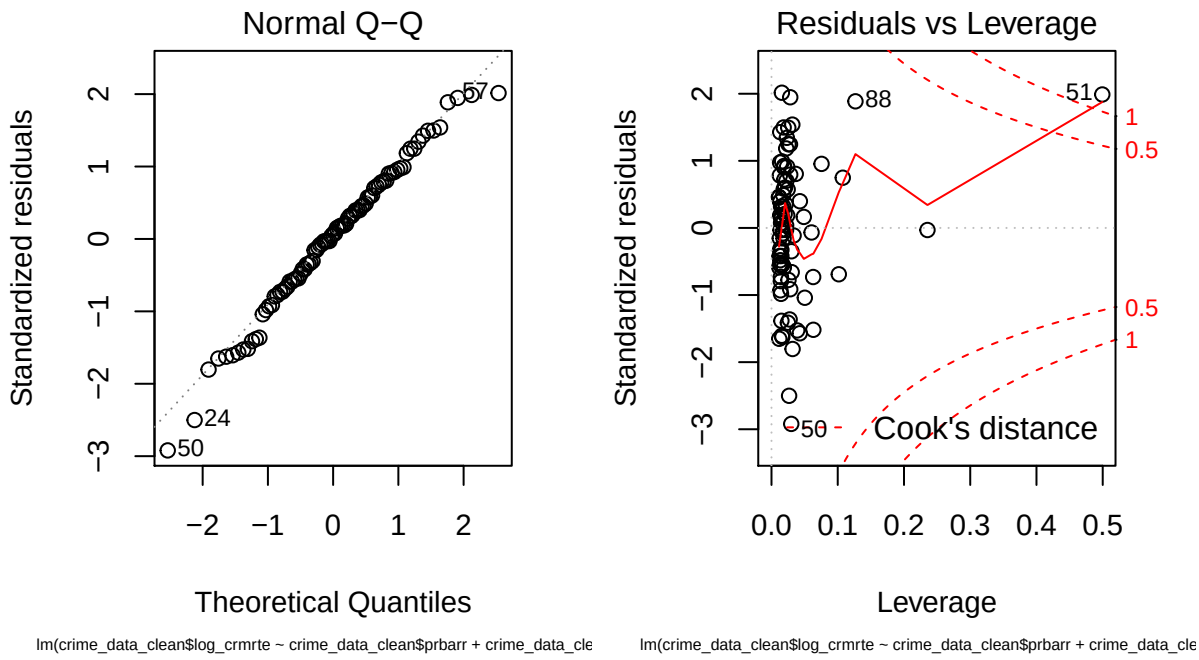


- *MLR1 Linearity*: The model is linear in the parameters given.
- *MLR2 Random Sampling*: The sampling process is not clear but assumed to be a random selection of counties.
- *MLR3 Colinearity*: Inspection of scatterplots above did not reveal any perfect colinearity amongst the chosen varaibles.
- *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 indepedents.
- *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 = 1.1494, df = 2, p-value = 0.5629
```

```
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.

Finally, there is one point (51) with Cook's distance > 1 which we need to examine.

```
crime_data_clean[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 log_crmrte log_density log_polpc
## 51         1.5  -5.196989  -0.9524121 -4.704512
```

Since 51 is an outlier due to high prbconv and prbarr both being greater than 1, we will remove this from the model.

```
crime_data_clean2 = crime_data_clean[-c(51),]
nrow(crime_data_clean2)
```

```
## [1] 88
```

Redoing model1 to remove row 51:

```
model1 = lm(crime_data_clean2$log_crmrte ~
              crime_data_clean2$prbarr +
              crime_data_clean2$prbconv_fix
              )
model1$coefficients
```

```
##                   (Intercept)      crime_data_clean2$prbarr
##                     -2.341403                     -2.508591
## crime_data_clean2$prbconv_fix
##                     -0.851451
```

Model 3 - using judicial and demographic system variables

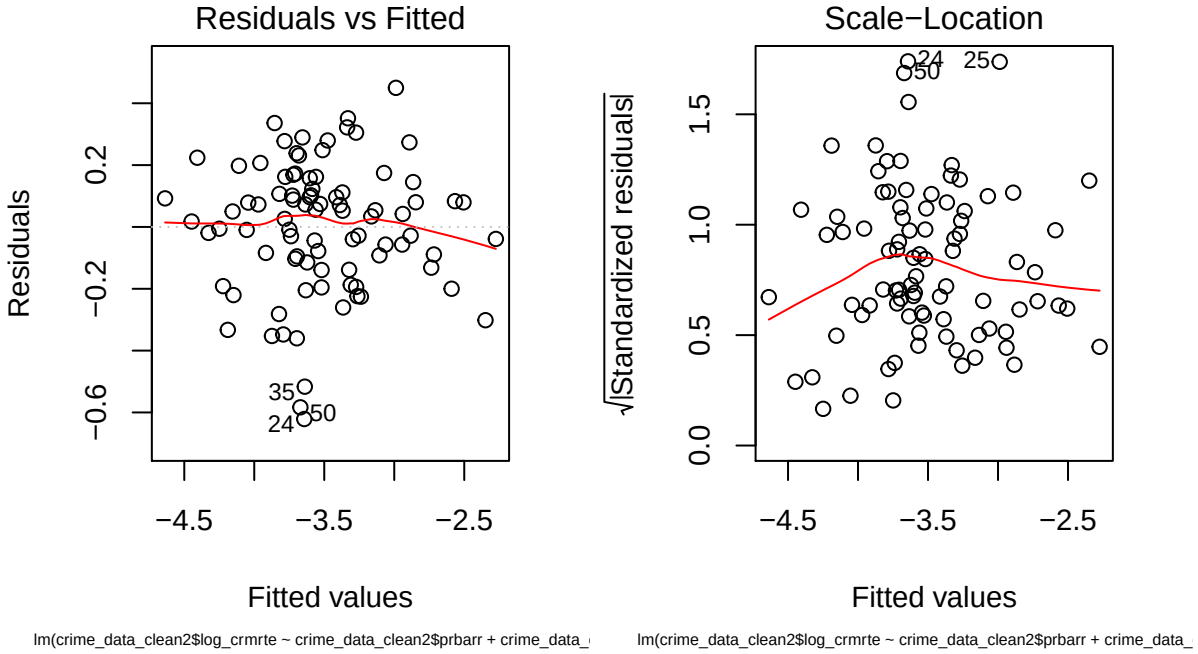
Model 3 is a more elaborate model that takes into account both judicial and demographic system variables to come up with a better causal explanation of crime rate. In this model, we included all meaningful

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_clean2$log_crmrte ~
              crime_data_clean2$prbarr +
              crime_data_clean2$prbconv_fix +
              crime_data_clean2$prbpris +
              crime_data_clean2$avgsen +
              #crime_data_clean2$polpc +
              crime_data_clean2$log_polpc +
              crime_data_clean2$log_density +
              crime_data_clean2$taxpc +
              crime_data_clean2$pctmin80 +
              crime_data_clean2$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)
```

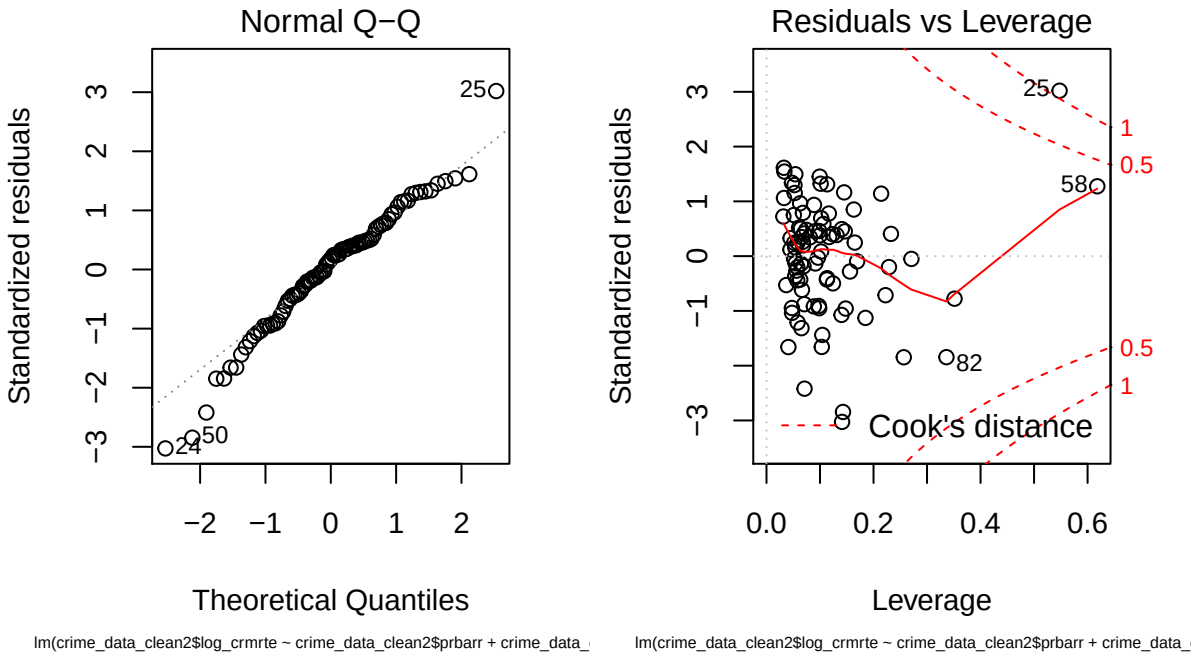


- *MLR1-3*: As per Model 1 above.
- *MLR4 Zero Conditional Mean*: The inclusion of more explanatory variables improves the mean residual, making it closer to zero.
- *MLR5 Homoskedasticity*: This model appears to have improved the scale location plot and the Breusch-Pagan test does not reject homoskedasticity.

```
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: Sun, Dec 09, 2018 - 21:57:56

	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 significanlty 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 standardised 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_clean2$log_crmrte ~ crime_data_clean2$prbarr +
```

```
##     crime_data_clean2$prbconv_fix + crime_data_clean2$prbpris +
##     crime_data_clean2$avgsen + crime_data_clean2$log_polpc +
##     crime_data_clean2$log_density + crime_data_clean2$taxpc +
##     crime_data_clean2$pctmin80 + crime_data_clean2$pctymle)
##
## Standardized Coefficients::
##                   (Intercept)      crime_data_clean2$prbarr
##                    0.00000000                   -0.33135723
## crime_data_clean2$prbconv_fix     crime_data_clean2$prbpris
##                   -0.40156299                   -0.06428477
##      crime_data_clean2$avgsen   crime_data_clean2$log_polpc
##                   -0.05023913                    0.26909498
## crime_data_clean2$log_density       crime_data_clean2$taxpc
##                    0.41363685                    0.07881216
##    crime_data_clean2$pctmin80     crime_data_clean2$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 the density, polpc and pctmin80 show particularly strong contribution to the model. The prbconv_fix variable seems to of lower significance in the Model 3. This is because it correlates quite a bit with polpc. If we remove "polpc" from the model3, prbconv_fix becomes significant. We can see this here:

```
#everything in model3 except polpc
model4 = lm(crime_data_clean2$log_crmrte ~
              crime_data_clean2$prbarr +
              crime_data_clean2$prbconv_fix +
              crime_data_clean2$prbpris +
              crime_data_clean2$avgsen +
              #crime_data_clean$polpc +
              #crime_data_clean$log_polpc +
              crime_data_clean2$log_density +
              crime_data_clean2$taxpc +
              crime_data_clean2$pctmin80 +
              crime_data_clean2$pctymle)
lm.beta(model4)
```

```
##
## Call:
## lm(formula = crime_data_clean2$log_crmrte ~ crime_data_clean2$prbarr +
##     crime_data_clean2$prbconv_fix + crime_data_clean2$prbpris +
##     crime_data_clean2$avgsen + crime_data_clean2$log_density +
##     crime_data_clean2$taxpc + crime_data_clean2$pctmin80 + crime_data_clean2$pctymle)
##
## Standardized Coefficients::
##                   (Intercept)      crime_data_clean2$prbarr
##                    0.00000000                   -0.32874532
## crime_data_clean2$prbconv_fix     crime_data_clean2$prbpris
##                   -0.40543986                   -0.02977654
##      crime_data_clean2$avgsen crime_data_clean2$log_density
##                    0.01541994                    0.50732410
##       crime_data_clean2$taxpc    crime_data_clean2$pctmin80
##                    0.20308191                    0.42527334
##     crime_data_clean2$pctymle
##                    0.07655393
```

We can see here that prbconv_fix > 0.1 in magnitude and is a good candidate for our parsimonious model.

We therefore select the following for our second model, but we will go through one more iteration after we look at the joint significance of these variables.

1. prbarr
2. density
3. pctmin80
4. prbconv_fix
5. taxpc

Note that we elected to remove the police per capita variable ("polpc") from this model as we believe it is likely an effect rather than a cause, and heavily correlated to other variables in the regression.

Creating Model 2 out of these variables:

```
model2_ver1 = lm(crime_data_clean2$log_crmrte ~
              crime_data_clean2$prbarr +
              crime_data_clean2$log_density +
              crime_data_clean2$prbconv_fix +
              crime_data_clean2$taxpc +
              crime_data_clean2$pctmin80)
coeftest(model2_ver1, vcov = vcovHC)
```

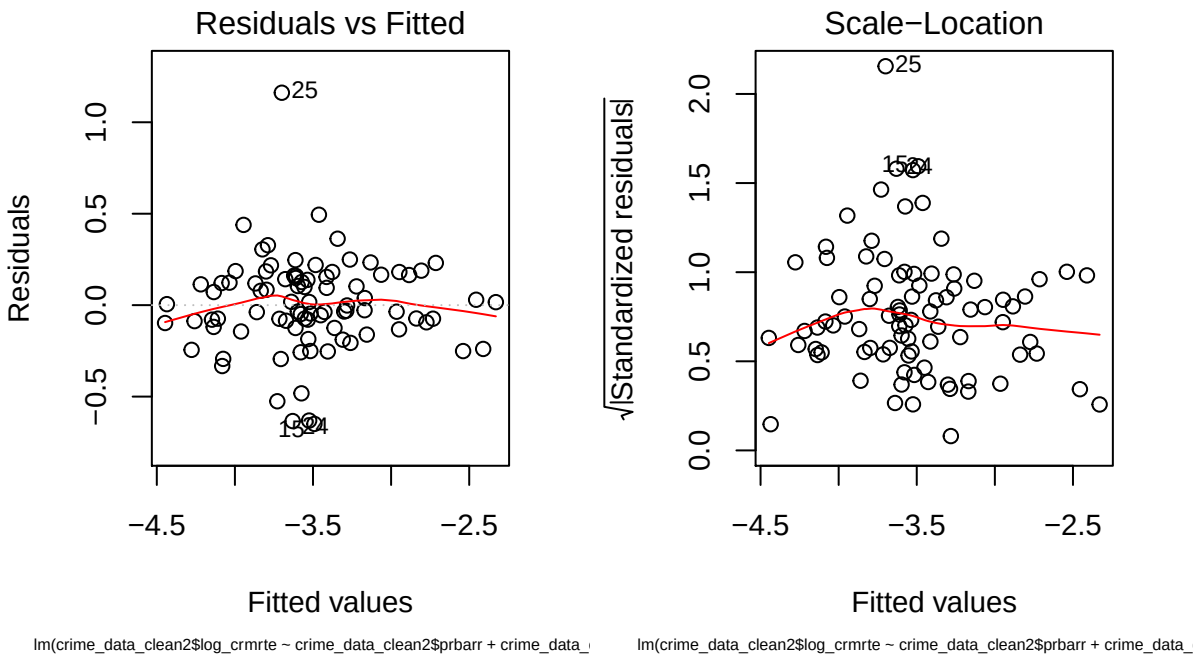
```
##
## t test of coefficients:
##
##                                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                   -3.3338864  0.2201317 -15.1450 < 2.2e-16 ***
## crime_data_clean2$prbarr      -1.6990307  0.3118676  -5.4479 5.219e-07 ***
## crime_data_clean2$log_density  0.3496533  0.0673339   5.1928 1.479e-06 ***
## crime_data_clean2$prbconv_fix -0.6374485  0.1195927  -5.3302 8.464e-07 ***
## crime_data_clean2$taxpc        0.0076726  0.0065951   1.1634     0.248
## crime_data_clean2$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 taxpc is not showing up as being signficant, as a result we will remove it from the final model and go with the other 4 variables to get to a more parsimonious model.

```
model2 = lm(crime_data_clean2$log_crmrte ~
              crime_data_clean2$prbarr +
              crime_data_clean2$log_density +
              crime_data_clean2$prbconv_fix +
              crime_data_clean2$pctmin80)
```

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)
```

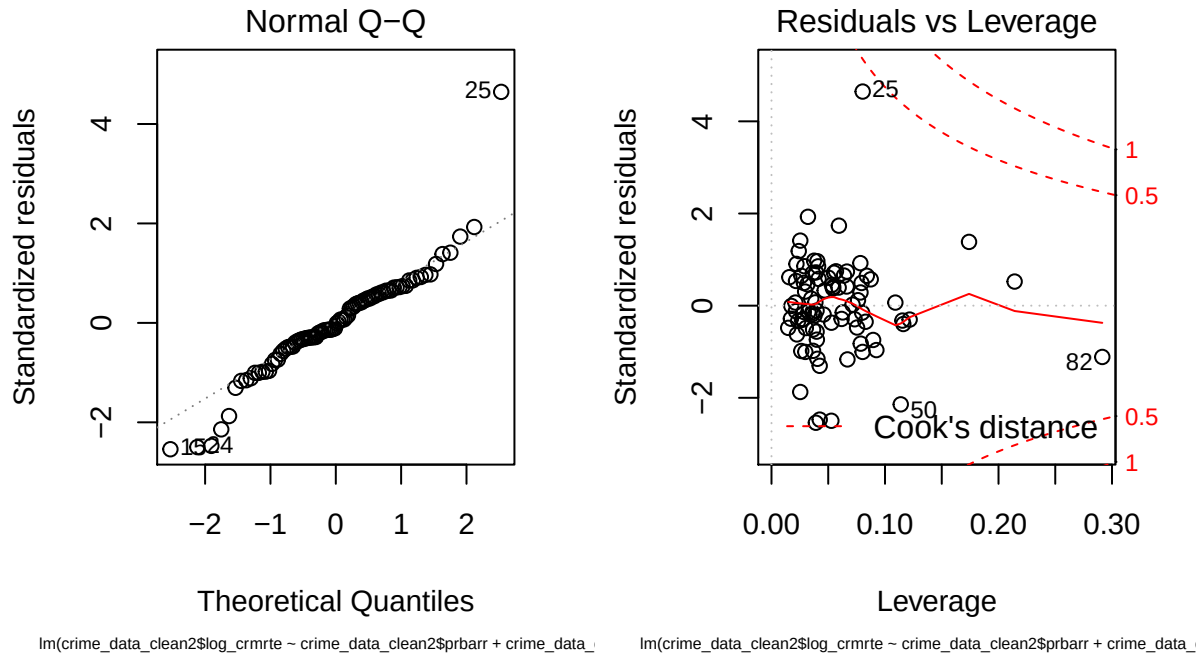


- *MLR1-3*: As per Model 1 above.
- *MLR4 Zero Conditional Mean*: Further improvement and no violation of zero conditional mean in this model.
- *MLR5 Homoskedasticity*: This model appears to have futher improved the scale location plot and the Breusch-Pagan test does not reject homoskedasticity.

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

We can now compare all three models:

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

stargazer(model1, model2, model3,
          star.cutoffs =c(0.05,0.01, 0.001),
          se = robust_errors,
          type = 'latex',
          column.labels = c('Model 1', 'Model 2', '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: Sun, Dec 09, 2018 - 21:57:57

	Dependent variable:		
	log_crmrte		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
prbarr	-2.509^{***}	-1.805^{***}	-1.616^{***}
	(0.516)	(0.367)	(0.341)
log_density		0.365^{***}	0.281^{***}
		(0.060)	(0.068)
taxpc			0.003
			(0.006)
prbconv_fix	-0.851^{***}	-0.673^{***}	-0.609^{***}
	(0.166)	(0.132)	(0.128)
prbpris			−0.439
			(0.324)
avgsen			−0.010
			(0.011)
log_polpc			0.439^{**}
			(0.142)
pctmin80		0.013^{***}	0.013^{***}
		(0.002)	(0.002)
pctymle			0.638
			(1.813)
Constant	-2.341^{***}	-2.994^{***}	−0.121
	(0.209)	(0.202)	(1.027)
Observations	88	88	88
R^2	0.406	0.760	0.837
Adjusted R^2	0.392	0.748	0.818
Residual Std. Error	0.405 (df = 85)	0.261 (df = 83)	0.221 (df = 78)
F Statistic	29.092^{***} (df = 2; 85)	65.599^{***} (df = 4; 83)	44.486^{***} (df = 9; 78)
Note:			*p<0.05; **p<0.01; ***p<0.001

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

```
##        df       AIC
## model1  4 95.526674
## model2  6 19.941403
## model3 11 -4.187382
```

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.

5. Discussion - Model Specification & Omitted Variables

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

Category	Description
Demographics	There is very little information on demographics other than pctmin80 which is based on dated information about minorities; the nature and proportion of particular minorities is omitted here. * If, say, a particular cultural group was more prone to crime we would expect greater crime rate. It could be useful to get additional information on the type of people comprising the county population, and inclusion of variables expressing religious make-up could, for example show less crime in regions densly populated by religions valuing non-violence.
Education Level	Typically we would expect higher the education level to lead to lower the crime rate. This is likely related to educated decision making but also to employment outcomes which promote stability.
Wages	The more affluent neighborhoods will tend to have lesser crime. We thought this would be reflectedby tax revenues per capita, but it does not appear to be the case. This may be due to the nature of the taxation system itself in creating a representative variable.
Employment	Gainful employment of citizens is well known to decrease crime rate however, we have no information about the employment rates in each county. The only possibility is to compare the wage rates in each county which could express greater employment when competition has driven rates up. Limited employment may result in the presence of higher-density (social) housing and therefore postitively bias the density coefficient.
Commercial Sectors	High crime might be associated with regions which have a lot of bars or similar entertainment venues but less crime in rural residential areas. Such an omitted variable (say indication number of nightlife venues) would potentially bias our density dependance by increasing the co-efficient or even increase the dependence of our model probability of arrest due to police presence in entertainment areas.
Age Distribution	We are provided with a variable indicating the percent of the population who are young males, but it would be equally useful to be provided with data for children and geriatrics - both parties whose presence would decrease the crime rate. Both of these groups typically increase density, but decrease crime and we would expect their presence to bias the density variable negatively.
Detailed Crime Info	Not all crime is equal, and different types of crime (violent crime, property crime, etc.) might be explained by different factors. Such omitted variables bias the probability of arrest, conviction and average sentence variables. Sentence length, for example, would be dependant on a measure of the proportion of non-violent crime and we may find sentence length is also negatively biased by this omission.

6. Conclusion

Based on our analysis, the probability of arrest and conviction help drive down crime rates. Increasing the probability of arrest by one unit is correlated with a 186% decrease in crime rate. Increasing the probability of conviction by one unit is correlated with a 71% decrease in crime rate.

Density is positively correlated with crime rate. An increase of one unit in density is correlated with a 35% increase in the crime rate.

Based on these results, our policy recommendations would be to:

1. Increase awareness of the effectiveness of the judicial system in counties that are effective at bringing perpetrators to justice; and increase resources, training, and oversight in those that are not.
2. Further investigation of the link between population density and crime rate. Are there economic factors at play? Demographics? Policing techniques in urban settings?