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. "NA" data is removed in some cases(there were 6 empty rows at the end of the dataset) .

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

2. ue 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 > 1 in some cases.

```
sum(crime_data_corr$prbarr > 1)
```

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

```
sum(crime_data_corr$prbconv_fix > 1)
```

```
## [1] 10
```

```
sum(crime_data_corr$prbpris > 1)
```

```
## [1] 0
```

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

```
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)
```

```
## 'data.frame':    81 obs. of  26 variables:
##  $ county      : int  1 5 7 9 11 13 15 17 21 23 ...
##  $ year        : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte      : num  0.0356 0.013 0.0268 0.0106 0.0146 ...
##  $ prbarr      : num  0.298 0.444 0.365 0.518 0.525 ...
##  $ prbconv     : Factor w/ 92 levels "","`","0.068376102",..: 63 13 62 52 3 59 78 42 23 37 ...
##  $ prbpris     : num  0.436 0.6 0.435 0.443 0.5 ...
##  $ avgsen      : num  6.71 6.76 7.14 8.22 13 ...
##  $ polpc       : num  0.00183 0.00123 0.00153 0.00086 0.00288 ...
##  $ density     : num  2.423 0.413 0.492 0.547 0.611 ...
##  $ taxpc       : num  31 34.8 42.9 28.1 35.2 ...
##  $ west        : int  0 1 0 1 1 0 0 0 1 1 ...
##  $ central     : int  1 0 1 0 0 0 0 0 0 0 ...
##  $ urban       : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ pctmin80    : num  20.22 3.16 47.92 1.8 1.54 ...
##  $ wcon        : num  281 227 375 292 250 ...
##  $ wtuc        : num  409 372 398 377 401 ...
##  $ wtrd        : num  221 229 191 207 188 ...
##  $ wfir        : num  453 306 281 289 259 ...
```

```
##  $ wser       : num  274 210 257 215 237 ...
##  $ wmfg       : num  335 238 282 291 259 ...
##  $ wfed       : num  478 359 412 377 391 ...
##  $ wsta       : num  292 332 328 367 326 ...
##  $ wloc       : num  312 281 299 343 275 ...
##  $ mix        : num  0.0802 0.4651 0.2736 0.0601 0.3195 ...
##  $ pctymle    : num  0.0779 0.0721 0.0735 0.0707 0.0989 ...
##  $ prbconv_fix: num  0.5276 0.2679 0.5254 0.4766 0.0684 ...
```

4. There is a duplicate entry for county #193, which we removed from the dataset.

```
crime_data_corr2[crime_data_corr2$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_corr3 = crime_data_corr2[!duplicated(crime_data_corr2), ]
```

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_corr3$density > 0.001)
crime_data_corr4 = subset(crime_data_corr3, good_density)
```

After cleansing we have 79 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.

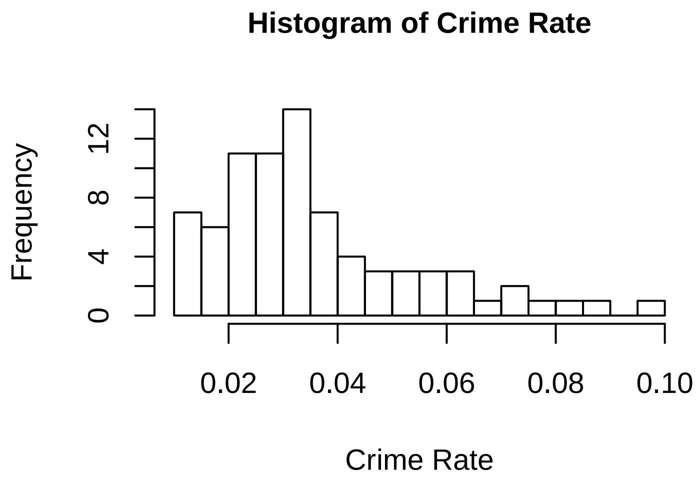
Summarizing the variable we note a small range of fractional values, centred on a mean of approximately 3.5 crimes per hundred people in the year period.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02345 0.03059 0.03578 0.04397 0.09897
```

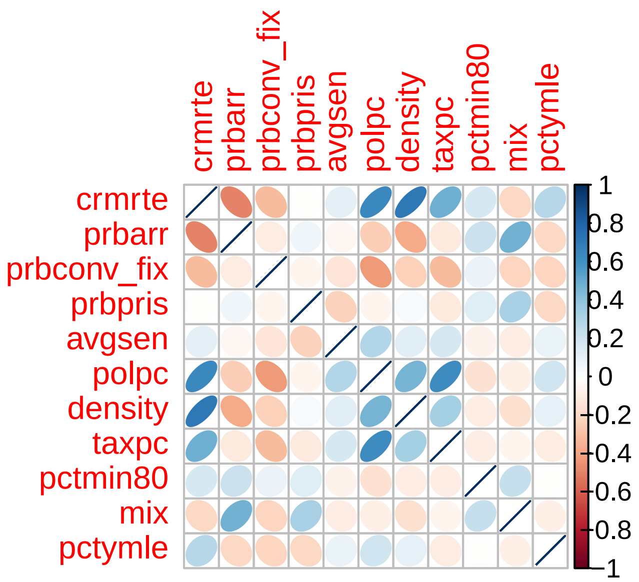
The distribution of crime rate is right-skewed in this dataset but sufficient data is available for modelling.

```
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")
```



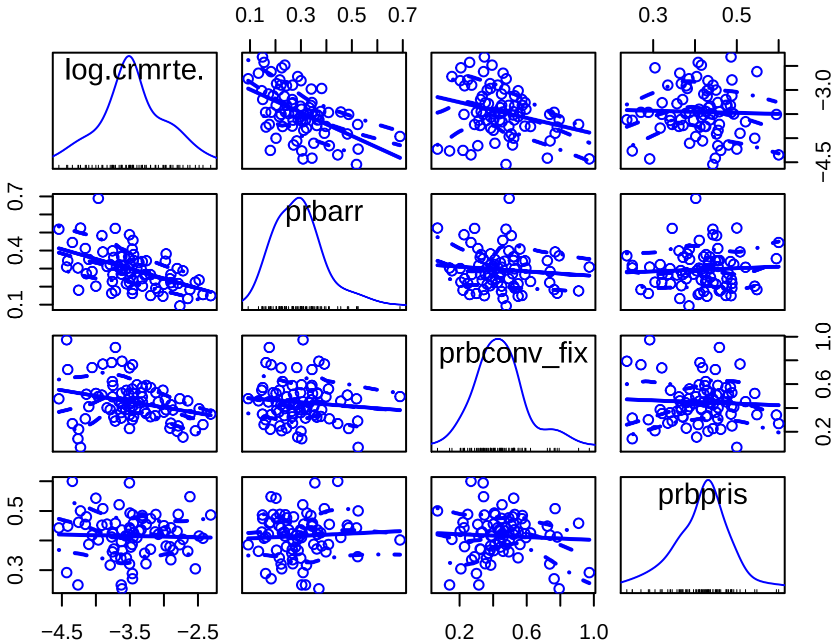
We note the cases with particularly high correlation with crmrte and examine them below.

Independent Variables - Judicial

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

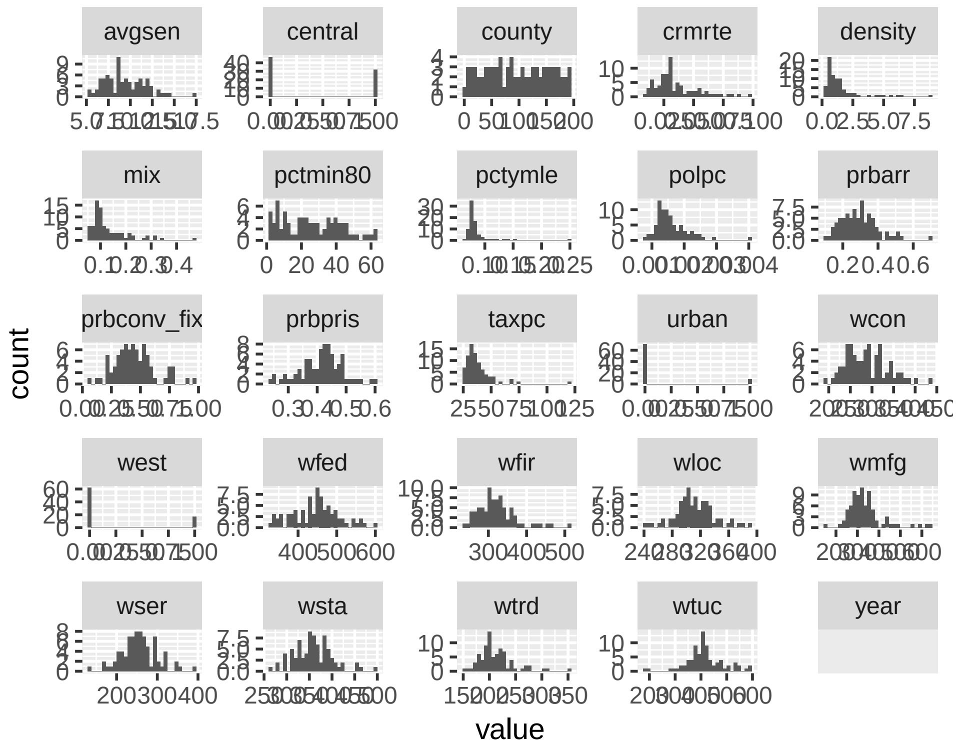
It is likely that crime rate will be lower when the probability of getting arrested, convicted or going to prison is higher due to the deterrent effect. These variables are expected to have causal relationships with crime rate ("crmrte") and should reveal correlation, which we examine through a scatterplot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

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




```r
crime_data_clean %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) + facet_wrap(~key, scales = "free") + geom_histogram(bins = 30)
```

```
## Warning: Computation failed in `stat_bin()`:
## `binwidth` must be positive
```



The log(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)
```

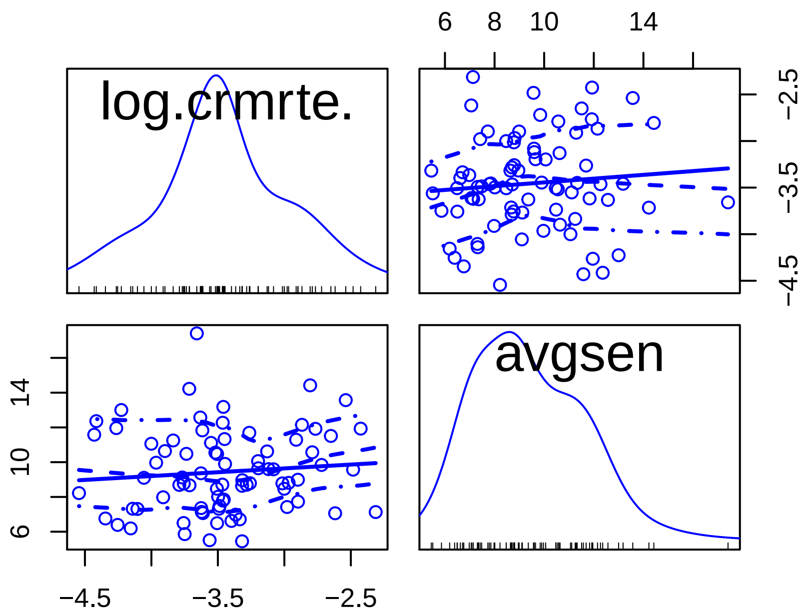
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.450   7.450   8.990   9.441  11.180  17.410
```

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

```
## [1] 0.1195381
```

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



Independent Variables - Demographic

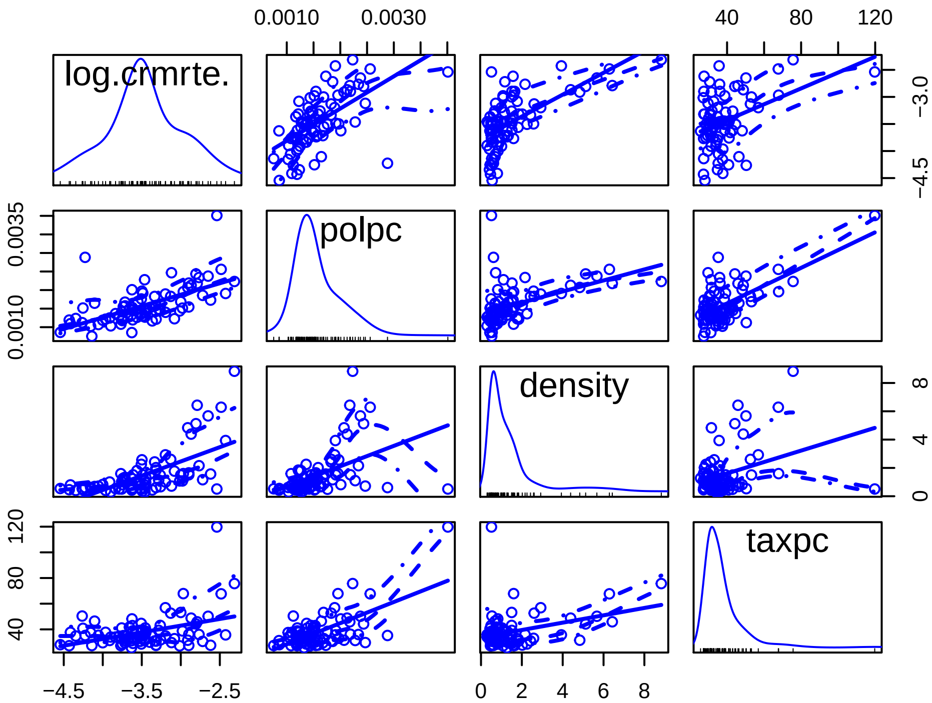
1. Police per capita ("polpc")
2. Density ("density")
3. Tax revenue per capita ("taxpc")
4. Percentage of Young males ("pctymle")
5. Percentage of minorities ("pctmin80")

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



Crime Rate is positively correlated to police per capita. We considered 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  0.5727  1.0262  1.5363  1.5962  8.8277
```

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

```
## [1] 0.7197649
```

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)
```

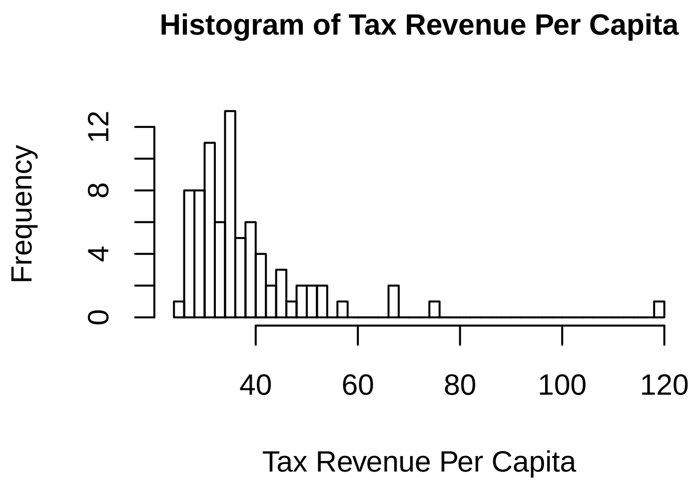
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.92   34.87   38.17   40.94  119.76
```

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

```
## [1] 0.4807509
```

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.

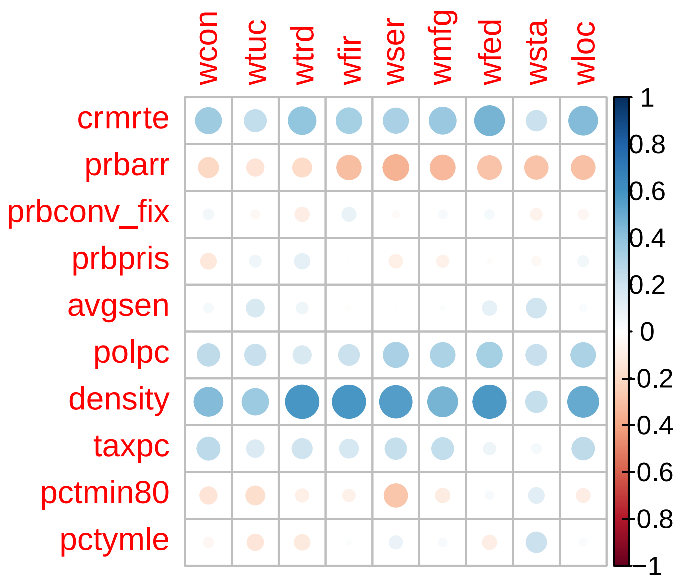
```r
hist(crime_data_clean$taxpc, breaks = 50,
     main = 'Histogram of Tax Revenue Per Capita',
     xlab = 'Tax Revenue Per Capita', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



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:

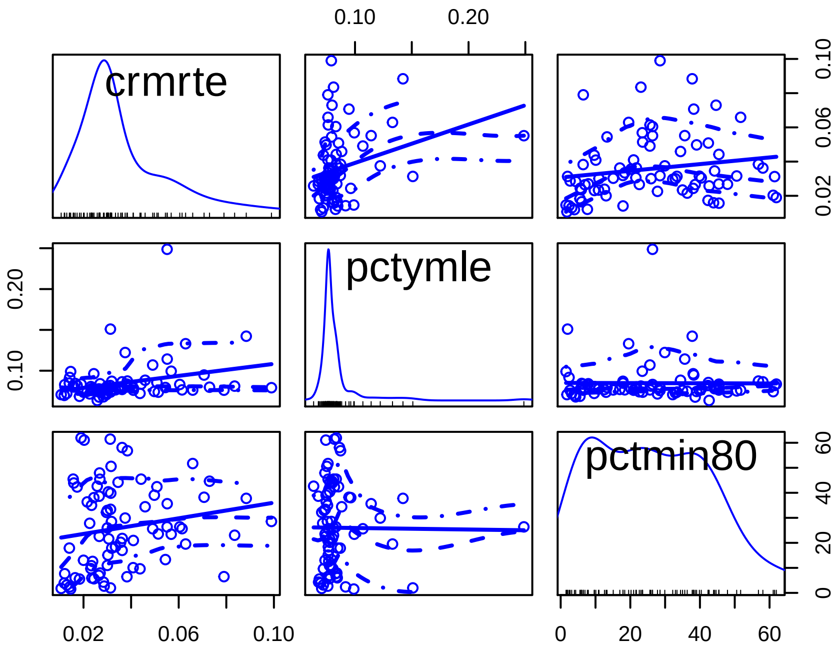
```r
corr_w = cor(crime_data_clean[, c(3,4,26,6,7,8,9,10,14,25,15,16,17,18,19,20,21,22,23)])
corrplot(corr_w[seq(1,10),seq(11,19)])
```



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:

```r
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.
## 0.06356 0.07546 0.07795 0.08475 0.08377 0.24871
```

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

```
## [1] 0.2875495
```

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   1.541  10.477   25.629   26.030   38.842   61.942
```

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

```
## [1] 0.1747599
```

The correlation is weak in both cases.

3. Data Transformation

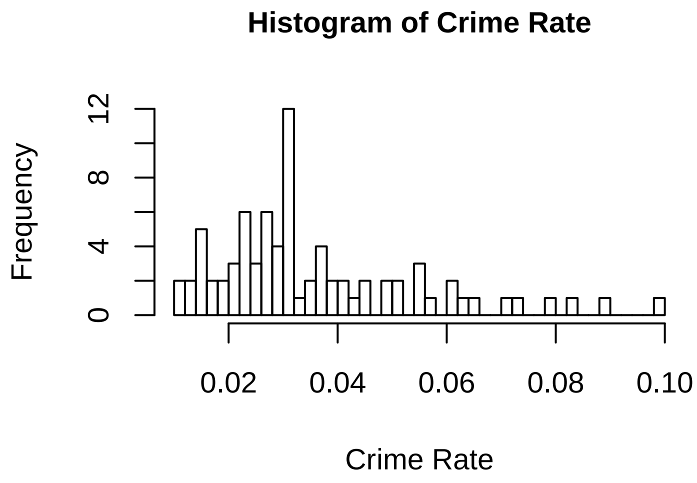
Crime Rate

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:

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.01062 0.02345 0.03059 0.03578 0.04397 0.09897
```

```r
hist(crime_data_clean$crmrte, breaks = 50,
     main = 'Histogram of Crime Rate',
     xlab = 'Crime Rate', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



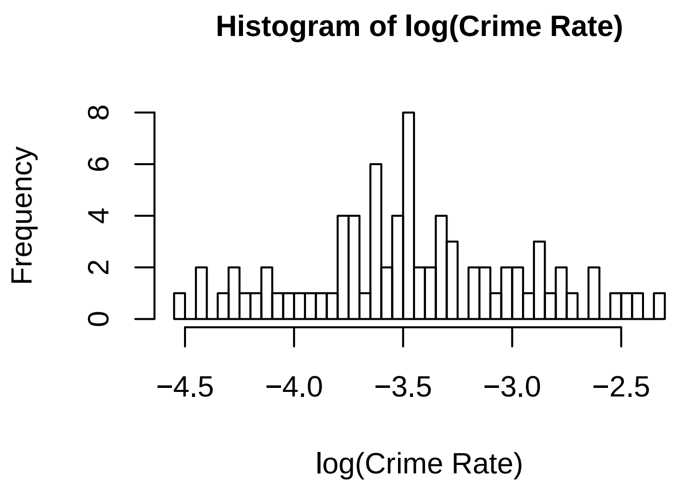
We applied a log() transformation to the variable which addresses both issues.

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

```r
crime_data_clean['log_crmrte'] = log(crime_data_clean$crmrte)
summary(crime_data_clean$log_crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.545  -3.753  -3.487  -3.456  -3.124  -2.313
```

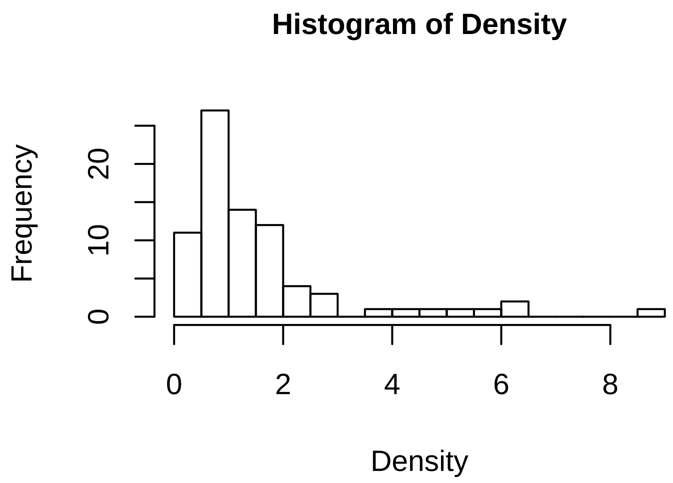
```r
hist(crime_data_clean$log_crmrte, breaks = 50,
     main = 'Histogram of log(Crime Rate)',
     xlab = 'log(Crime Rate)', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



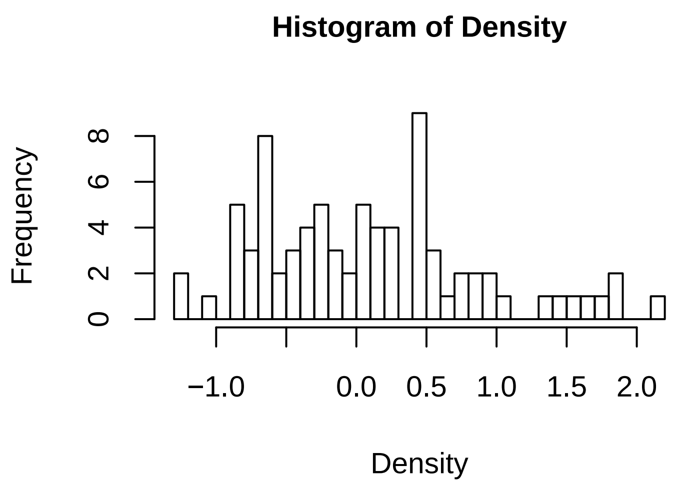
Density

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)
```

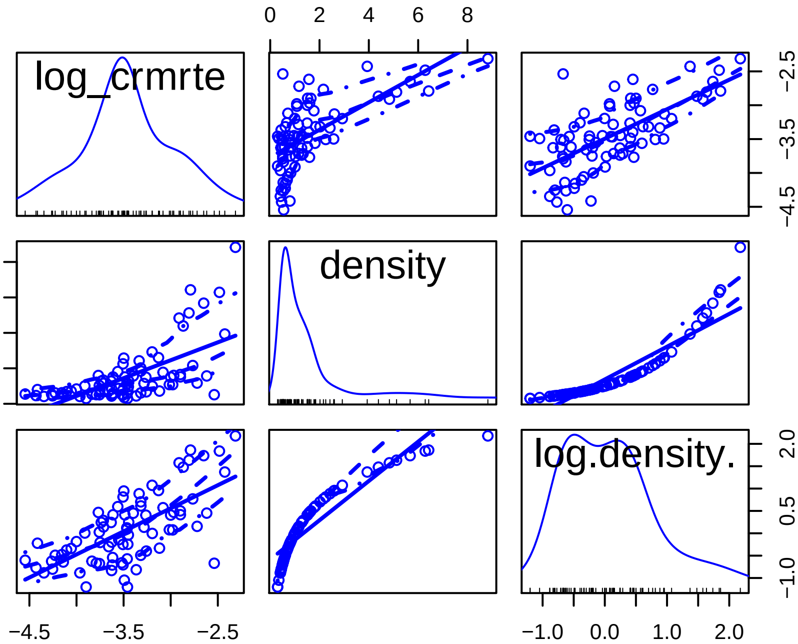



```
hist(log(crime_data_clean$density), breaks = 30,
     main = 'Histogram of Density',
     xlab = 'Density', cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9)
```



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] 0.6387235
```

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

```
## [1] 0.6702706
```

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

4. Regression Modelling

Model 1 - minimal using the Judicial system variables only

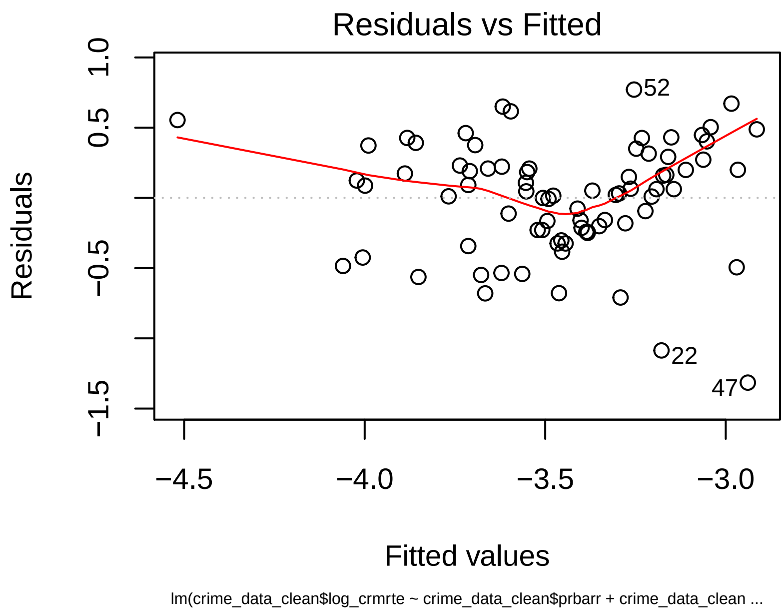
```
model1 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$prbconv_fix
              )
model1
```

Call: lm(formula = crime_data_clean$log_c rmrte\ crime_d ata_c lean$prbarr + crime_data_clean$prbconv_fix)

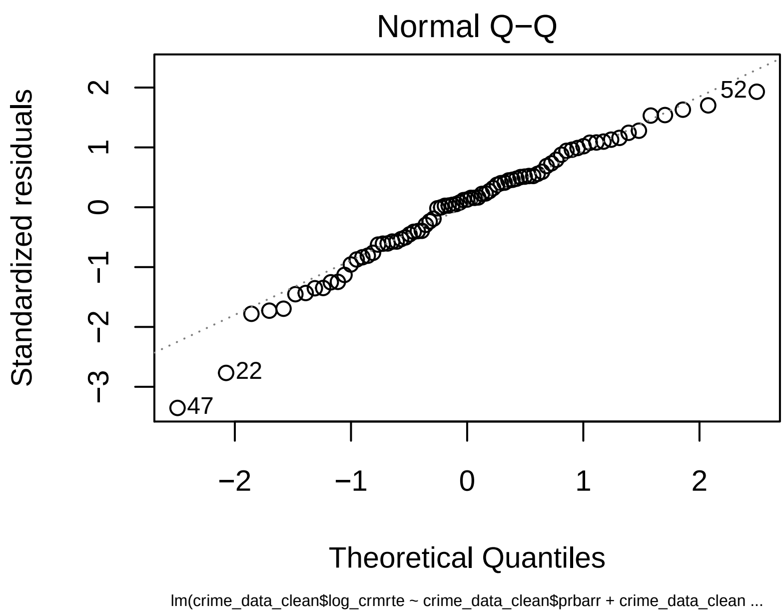
Coefficients: (Intercept) crime_data_clean$prbarr\ -2.260\ -2.574 crime_d ata_c lean$prbconv_fix
-0.978

We can then plot Model1 to evaluate OLS assumptions:

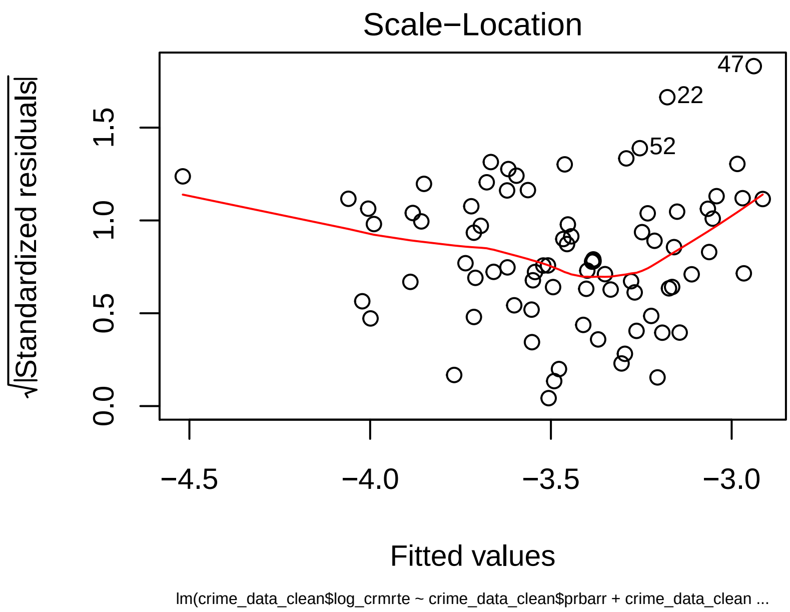
```
plot(model1, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 1)
```



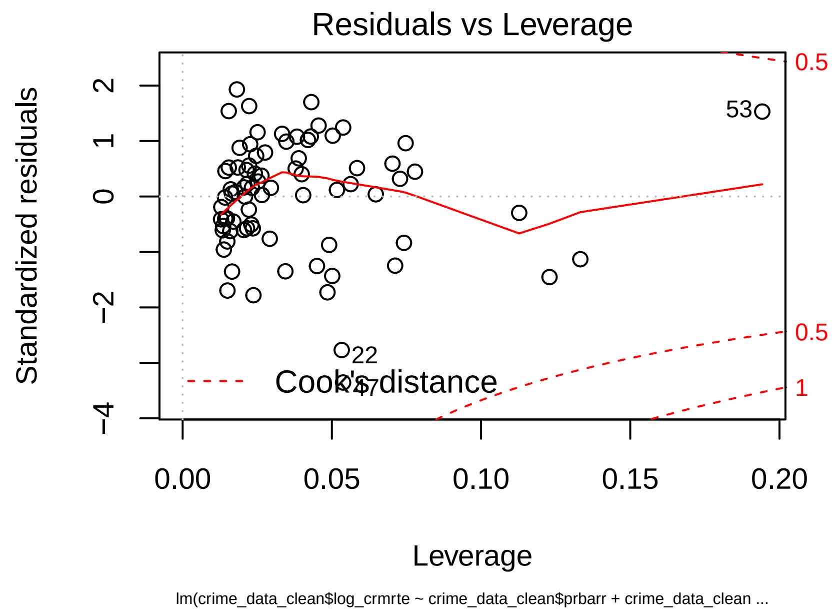

```r
plot(model1, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 2)
```




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




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



Zero conditional mean is violated. The Q-Q plot indicates a good amount of normality. This model is heteroskedastic as shown on the scalel-location plot.There are no points with Cook's distance > 1 which means that there are no significant outliers.

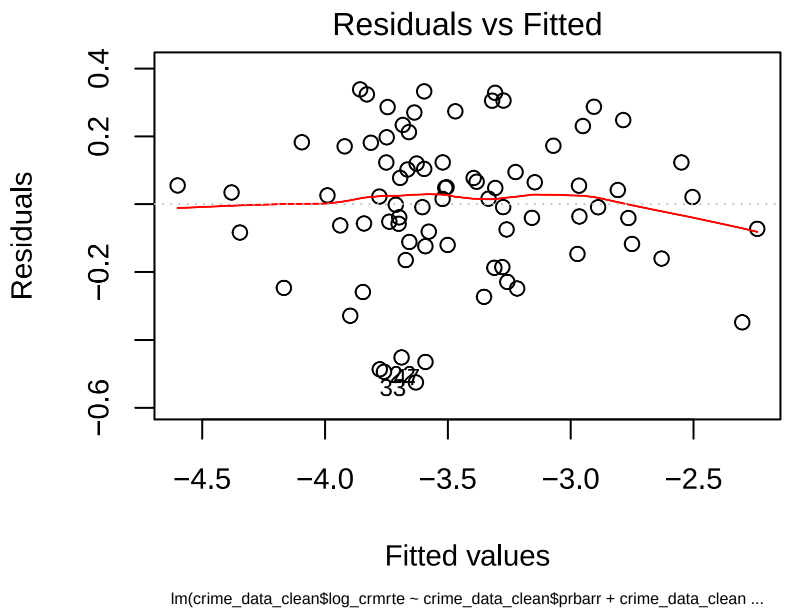
Model 3 - using judicial and demographic system variables

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.

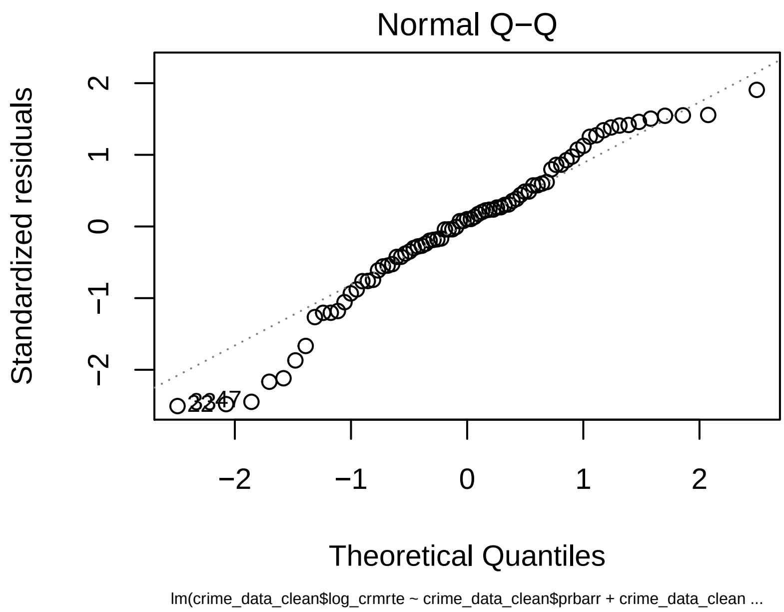
```
model3 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$prbconv_fix +
              crime_data_clean$prbpris +
              crime_data_clean$avgsen +
              crime_data_clean$polpc +
              crime_data_clean$log_density +
    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

We can then plot Model3 to evaluate OLS assumptions:

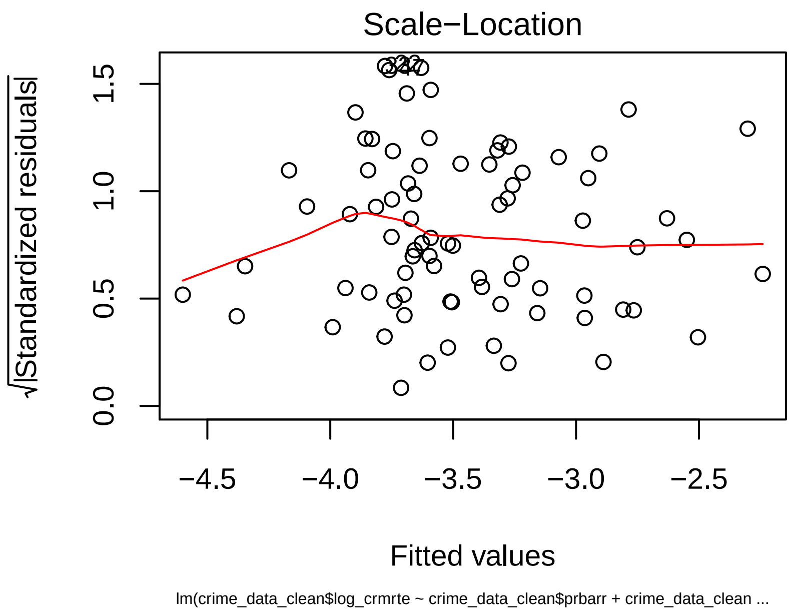
```
plot(model3, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 1)
```



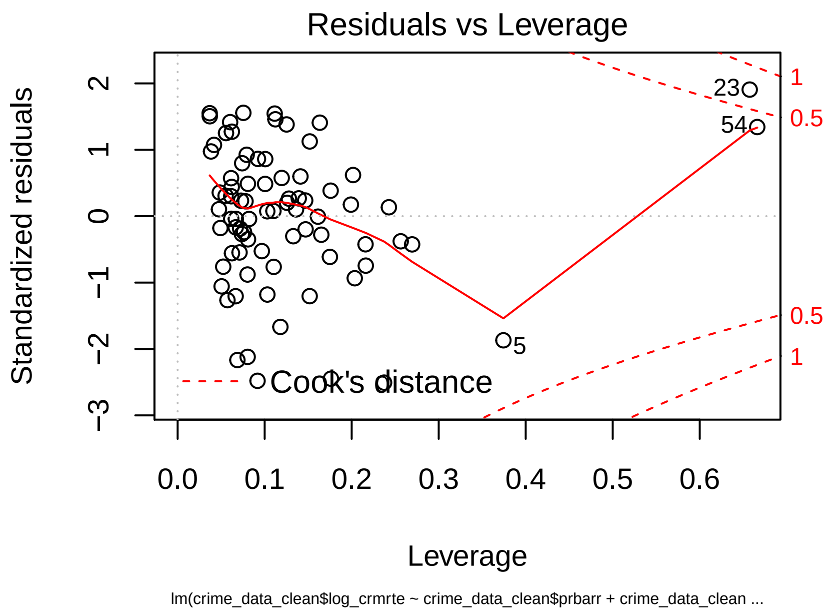

```
plot(model3, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 2)
```




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




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



The inclusion of more explanatory variables improves the mean, making it closer to zero, even though there is some violation of the zero conditional mean requirement.

The Q-Q plot shows significant deviations from normality. Model3 is better than model1 in terms of heteroskedasticity. However, there are definitely a couple of outliers that might skew results.

Comparing of model1 and model3 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(model1),
                      robust_se(model3))

stargazer(model1, model3,
          star.cutoffs =c(0.05,0.01, 0.001),
          se = robust_errors,
          type = 'text',
          font.size = 'small',
          float = FALSE)
```

```
##
## ==================================================================
##                                   Dependent variable:
##                        -------------------------------------------
##                                       log_crmrte
##                               (1)                     (2)
## ------------------------------------------------------------------
## prbarr                     -2.574***               -1.545***
##                             (0.535)                 (0.357)
##
## prbconv_fix                -0.978**                 -0.293
##                             (0.328)                 (0.238)
##
```

```
## prbpris                                              -0.361
##                                                     (0.393)
##
## avgsen                                               -0.017
##                                                     (0.012)
##
## polpc                                               301.725*
##                                                    (129.909)
##
## log_density                                          0.286***
##                                                     (0.064)
##
## taxpc                                                0.004
##                                                     (0.004)
##
## pctmin80                                             0.014***
##                                                     (0.002)
##
## pctymle                                              1.130
##                                                     (2.259)
##
## Constant                     -2.260***              -3.650***
##                               (0.259)               (0.471)
##
## -----------------------------------------------------------------
## Observations                    79                     79
## R2                             0.373                  0.827
## Adjusted R2                    0.357                  0.804
## Residual Std. Error    0.403 (df = 76)        0.222 (df = 69)
## F Statistic         22.619*** (df = 2; 76) 36.629*** (df = 9; 69)
## =================================================================
## Note:                               *p<0.05; **p<0.01; ***p<0.001
```

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

```
##        df       AIC
## model1  4 85.627029
## model3 11 -2.044801
```

As we can see, model3 has significanlty improved on the AIC, R2 and Residual SE, but there are some p-values that are not significant now (prbconv_fix, polpc). There is likely a more optimized model that has fewer coefficients that we can derive out of the Model1 and Model3 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_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$log_density +
##     crime_data_clean$taxpc + crime_data_clean$pctmin80 + crime_data_clean$pctymle)
##
## Standardized Coefficients::
##                  (Intercept)      crime_data_clean$prbarr
##                   0.00000000                  -0.32747628
## crime_data_clean$prbconv_fix     crime_data_clean$prbpris
##                  -0.10072985                  -0.05264131
##      crime_data_clean$avgsen       crime_data_clean$polpc
```

```
##                   -0.07972709                   0.30646668
## crime_data_clean$log_density        crime_data_clean$taxpc
##                    0.43990931                   0.09349452
##    crime_data_clean$pctmin80      crime_data_clean$pctymle
##                    0.45722169                   0.05433308
```

Based on the above we may consider removing those with lower (e.g. <0.1) gain.

Model2 - 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 probconv_fix variable seems to of lower significance in the model3, possible it correlates heavily with other variables and therefore decreasing in significance.

We therefore select the following for our second model:

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

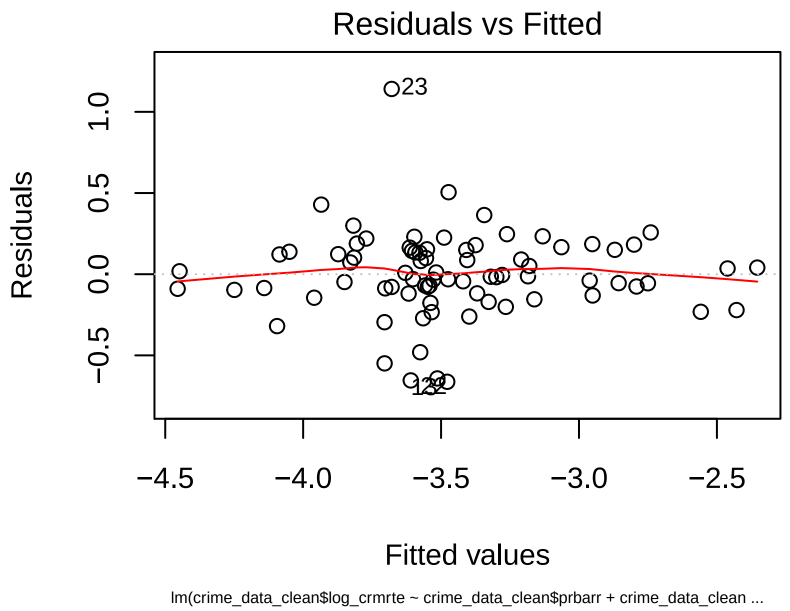
Note that we elected to remove the police per capita variable ("polpc") from this model as we believe it is likely an effect rather than a cause, and heavily correlated to other variables in the regression.

Creating model2 out of these variables:

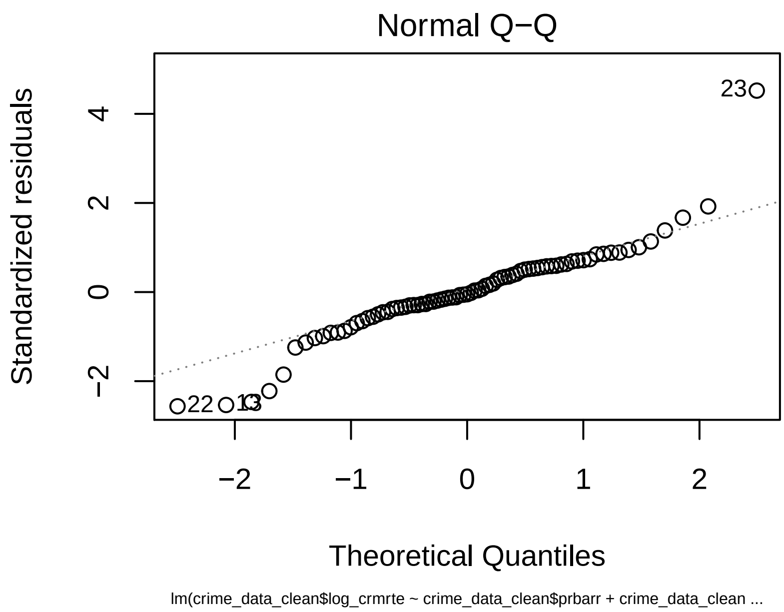
```
model2 = lm(crime_data_clean$log_crmrte ~
              crime_data_clean$prbarr +
              crime_data_clean$log_density +
              crime_data_clean$prbconv_fix +
              crime_data_clean$pctmin80)
```

We can then plot Model3 to evaluate OLS assumptions:

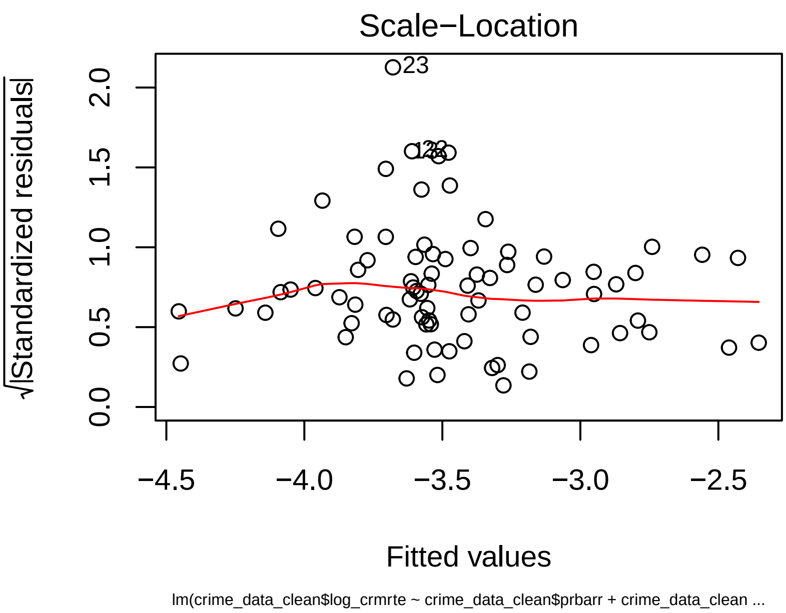
```
plot(model2, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 1)
```



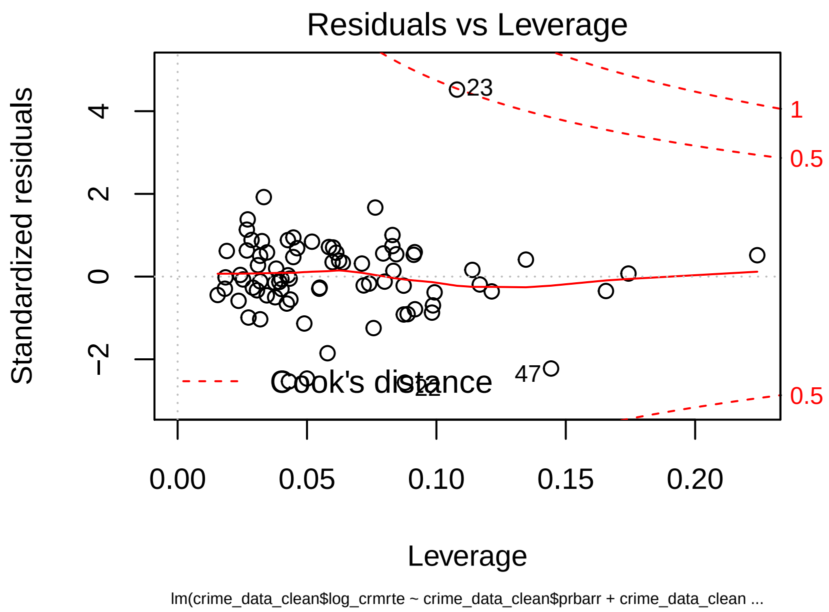
```r
plot(model2, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5, which = 2)
```




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




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



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 = 'text',
          font.size = 'small',
          float = FALSE)
```

```
##
## ========================================================================================
##                                              Dependent variable:
##                          ---------------------------------------------------------------------
##                                                  log_crmrte
##                               (1)                    (2)                    (3)
## ----------------------------------------------------------------------------------------
## prbarr                     -2.574***              -1.859***              -1.545***
##                            (0.535)                (0.397)                (0.357)
##
## log_density                                        0.348***               0.286***
##                                                   (0.070)                (0.064)
##
## taxpc                                                                     0.004
##                                                                          (0.004)
##
## prbconv_fix                -0.978**               -0.711**               -0.293
##                            (0.328)                (0.274)                (0.238)
##
## prbpris                                                                  -0.361
##                                                                          (0.393)
##
## avgsen                                                                   -0.017
```

```
##                                                                (0.012)
##
## polpc                                                          301.725*
##                                                               (129.909)
##
## pctmin80                                       0.013***         0.014***
##                                                (0.002)          (0.002)
##
## pctymle                                                          1.130
##                                                                 (2.259)
##
## Constant                   -2.260***          -2.966***        -3.650***
##                             (0.259)            (0.272)          (0.471)
##
## -------------------------------------------------------------------------------------
## Observations                  79                 79               79
## R2                           0.373              0.732            0.827
## Adjusted R2                  0.357              0.718            0.804
## Residual Std. Error    0.403 (df = 76)     0.267 (df = 74)    0.222 (df = 69)
## F Statistic        22.619*** (df = 2; 76) 50.557*** (df = 4; 74) 36.629*** (df = 9; 69)
## =====================================================================================
## Note:                                              *p<0.05; **p<0.01; ***p<0.001
```

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

```
##        df       AIC
## model1  4 85.627029
## model2  6 22.466835
## model3 11 -2.044801
```

Model2 shows significant improvement over Model1 with a better AIC, much better R2 and lower Residual SE.

Model2 also improves over Model3 in several areas including:

- Residuals vs Fitted plot shows that it is pretty close to satisfying the zero conditional mean requirement.
- The Q-Q plot is more normal than Model3. So the coefficients are more robust. We see that the p-values are all very significant unlike Model3's p-values showing that the coefficients are more consistent
- There are no major outliers in the residuals vs leverage plot unlike Model3.

5. Discussion - Model Specification & Omitted Variables

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

1. 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.
2. 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.
3. 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.

4. 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.
5. 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.
6. 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.
7. More detailed crime information: 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?