Lab3 Draft, w203: Statistics for Data Science

Avinash Chandrasekaran, Deepak Nagaraj, Saurav Datta

March 31, 2018

1. Introduction

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

The data provided consists of 25 variables and 91 different observations collected in a given year. Moreover the dataset obtained is a single cross-section of data collected from variety of different sources. For the analysis made in this research, we will assume that the data collected from different counties in NC were randomly sampled.

Our primary analysis of data will include ordinary least squares regressions to make casual estimates and we will clearly explain how omitted variabled may affect our conclusions. We begin our research by conducting exploratory analysis of the dataset to gain a better understanding of the variables.

2. Data Input

Let us read the data and have a first look.

```r
# Read the csv file
crime_data_raw = read.csv("crime_v2.csv")
```

Empty rows

There appears to be 6 rows of NA's across all variables. We can simply use na.omit(), because the number of all-NA rows matches the count on all the variables.

```r
# Remove NA rows
crime_data = na.omit(crime_data_raw)
```

Column formatting

We also notice that 'prbconv' is a factor while the rest of the variables are numeric.

```r
# convert factor to numeric for variable prbconv
crime_data$prbconv = as.numeric(levels(crime_data$prbconv)[crime_data$prbconv])
```

Unused variables

County and Year variables just represent the different counties and the year the data was collected. Year is always 87. Hence, we can safely remove these from the dataset for further analysis.

```r
crime_data = crime_data %>% dplyr::select(-c(year, county))
```

Duplicate records

We also noticed a duplicate record (record #89) in the dataset. As this could potentially affect our regression analysis, we will remove the duplicate record.

```
duplicated(crime_data)[duplicated(crime_data) == TRUE]
```

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

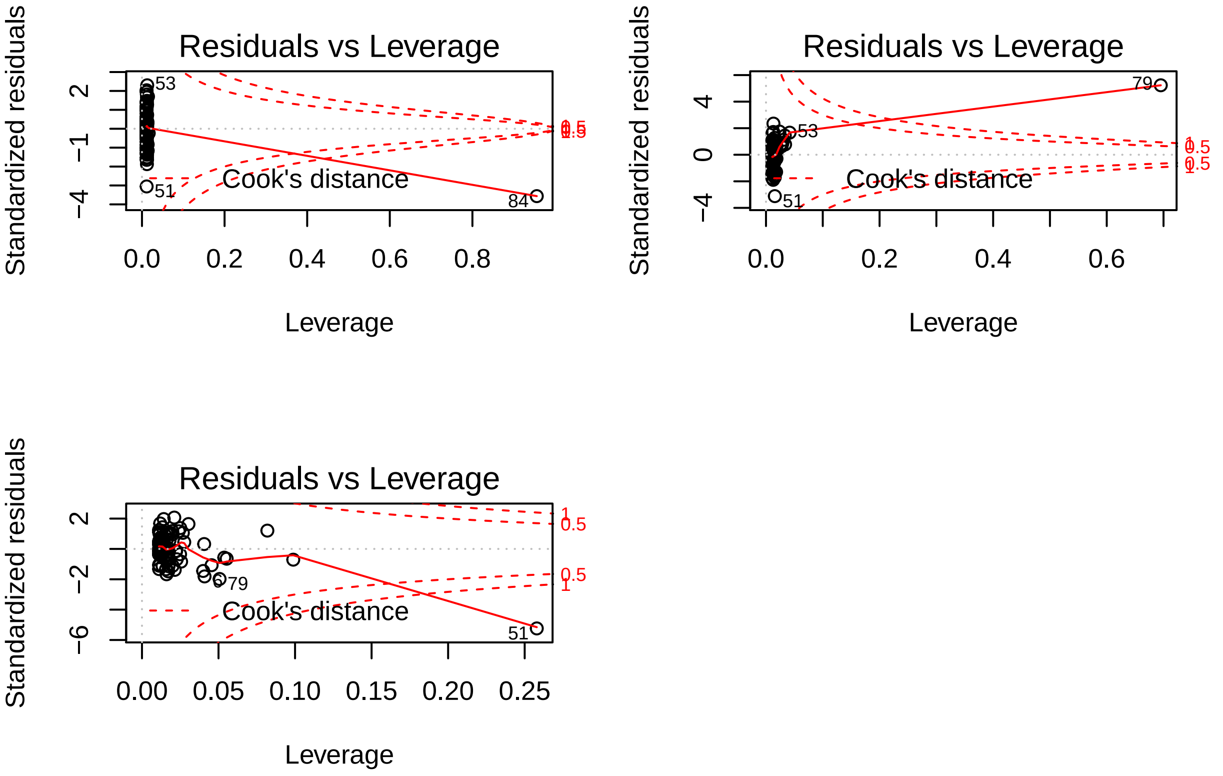
```
crime_data = distinct(crime_data)
```

3. Influential outliers

This section was filled in after a first pass at the variables. This is so that we remove any malformed observations in the beginning, and also to show the removed observations.

There is a large outlier for wages in service industry, *wser*. Observation #84 has very large influence as shown by Cook's distance.

```
par(mfrow = c(2, 2))
m = lm(log(crime_data$crmrte) ~ crime_data$wser)
plot(m, which = 5)
m = lm(log(crime_data$crmrte) ~ log(crime_data$density))
plot(m, which = 5)
m = lm(log(crime_data$crmrte) ~ log(crime_data$polpc))
plot(m, which = 5)
```



Next, let us consider *density*. Observation #79 has Cook's distance beyond 1, meaning extreme leverage:

Finally, we consider *polpc*. Police per capita has positive skew. Taking log helps, but we still see a very large outlier. Fitting a model, we see that observation #51 has Cook's distance beyond 1. This is a lot of leverage:

Here are all the outlier observations:

```
outlier_data = crime_data %>% slice(c(51, 79, 84))
structure(outlier_data)
```

```
## # A tibble: 3 x 23
##    crmrte prbarr prbconv prbpris avgsen   polpc   density taxpc  west
##     <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>     <dbl> <dbl> <int>
## 1 0.00553  1.09    1.50    0.500  20.7  0.00905 0.386      28.2     1
## 2 0.0140   0.530   0.328   0.150   6.64 0.00316 0.0000203  37.7     1
## 3 0.0109   0.195   2.12    0.443   5.38 0.00122 0.389      40.8     0
## # ... with 14 more variables: central <int>, urban <int>, pctmin80 <dbl>,
## #   wcon <dbl>, wtuc <dbl>, wtrd <dbl>, wfir <dbl>, wser <dbl>,
## #   wmfg <dbl>, wfed <dbl>, wsta <dbl>, wloc <dbl>, mix <dbl>,
## #   pctymle <dbl>
```

The first observation above has very low crime rate at very high police per capita. The second has extremely low density. The third has extremely high wages in the service industry.

The observations are questionable and affect our model because of their high influence, as measured by Cook's distance. We will remove them.

```
crime_data <- crime_data %>% slice(-c(51, 79, 84))
```

4. Exploratory Data Analysis

```
# Utility function to describe a column variable
f_describe_col = function(col, do_log = FALSE, plot_model = FALSE, do_sqrt = FALSE) {
    y = log(crime_data$crmrte)
    par(mfrow = c(2, 2))
    if (is.numeric(col)) {
        hist(col, main = "Histogram")
        boxplot(col, main = "Box plot")
    }
    if (do_log == TRUE) {
        x = log(col)
        hist(x, main = "Histogram, log")
    } else if (do_sqrt == TRUE) {
        x = sqrt(col)
        hist(x, main = "Histogram, sqrt")
    } else {
        x = col
    }
    if (is.numeric(col)) {
        print(paste("Correlation: ", signif(cor(x, y), 3)))
    }
    m = lm(y ~ x)
    plot(x, y, main = "Cor. with crime rate")
    if (is.numeric(col))
        abline(m, col = "blue")
    if (plot_model == TRUE) {
        if (do_log == TRUE) {
            m = lm(y ~ x)
        }
        plot(m, which = 5)
    }
}
```

We will start with an explanatory note on transformations. Any skew in the original data may cause the residuals not to follow normal distribution. If this happens, it violates an assumption of the LS regression model: we will not be able to draw inferences from our model. Hence it is important to ensure our residuals to follow normal distribution as much as possible, and to transform our predictors if that helps.

We will now try to get a sense of each variable in the dataset.

Single variable analysis

There are a total of 90 observations across 23 different variables. We will now explore each of the variables collected in the data.

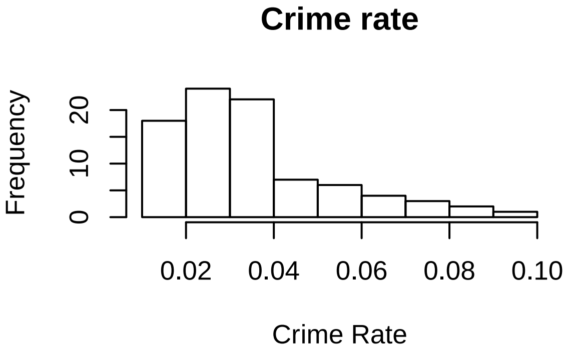
Crime rate

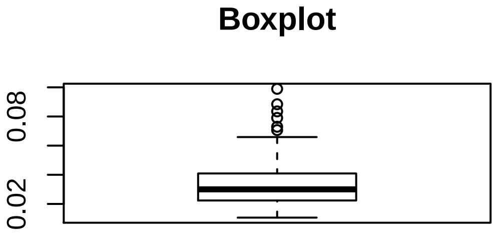
Crime rate is the key dependent variable of interest.

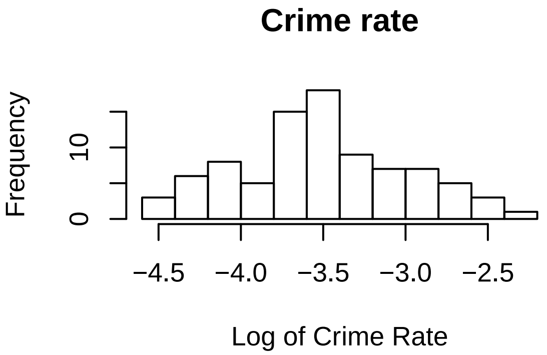
```
summary(crime_data$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.01062 0.02238 0.03002 0.03432 0.04090 0.09897
```

```
par(mfrow = c(2, 2))
hist(crime_data$crmrte, main = "Crime rate", xlab = "Crime Rate")
boxplot(crime_data$crmrte, main = "Boxplot")
hist(log(crime_data$crmrte), main = "Crime rate", xlab = "Log of Crime Rate")
crime_data$log_crmrte = log(crime_data$crmrte)
```





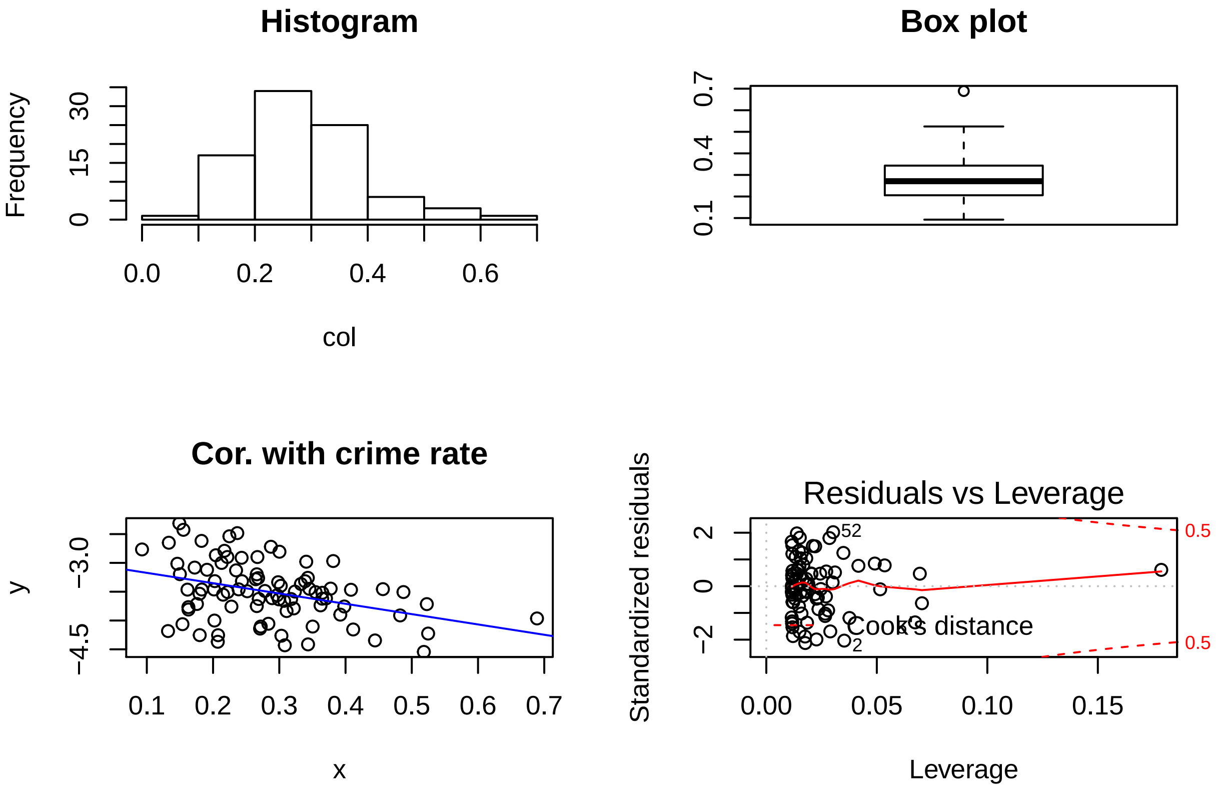


Looking at the histogram, the distribution is positively skewed to the left. We can take the log transformation which makes the variable appear more normally distributed.

Probability of arrest

```
f_describe_col(crime_data$prbarr, plot_model = TRUE)
```

```
## [1] "Correlation:  -0.374"
```



The plot looks fairly normal; there is only one outlier.

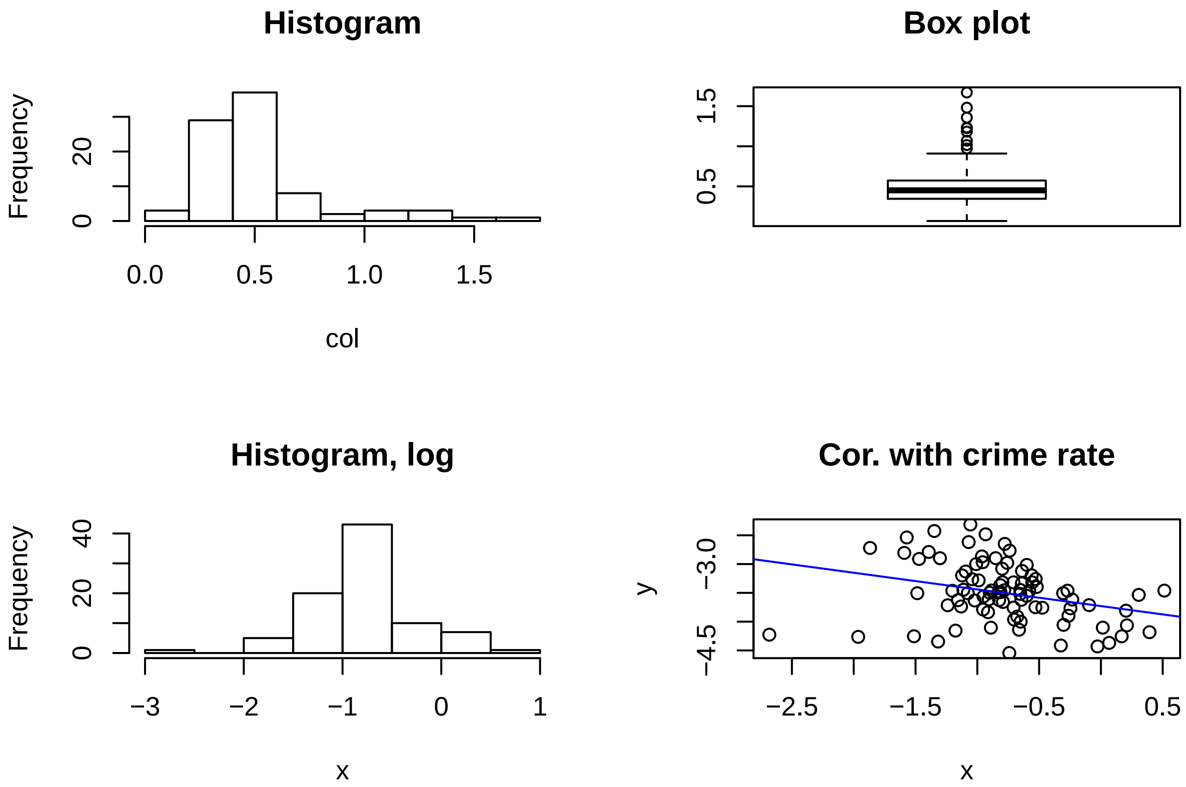
There is fairly negative correlation of -0.37: as probability of arrests increases, crime rate goes down. It may be that arrests are a deterrent, indicating causality.

We will include *prbarr* in our model.

Probability of conviction

```
f_describe_col(crime_data$prbconv, do_log = TRUE)
```

```
## [1] "Correlation:  -0.299"
```



```r
crime_data$log_prbconv = log(crime_data$prbconv)
```

This variable has quite a bit of left skew. It also has many outliers after the 3rd quartile. There are a few beyond 1 as well. Again, this is because we are not looking at a real probability but a ratio of convictions to arrests. It is possible, although perhaps uncommon, that a suspect is arrested once but convicted on multiple charges.

Taking a log transform improves the skew, although the spread is still quite a bit. There are no outliers with large influence as measured by Cook's distance.

There is moderate negative correlation with crime rate of -0.3. As convictions go up, crime rate goes down. Since we have already considered *prbarr*, let us check if *prbconv* has high correlation with *prbarr*:

```r
print(cor(crime_data$prbarr, crime_data$prbconv))
```

```
## [1] -0.296224
```

```r
print(cor(crime_data$prbarr, crime_data$log_prbconv))
```

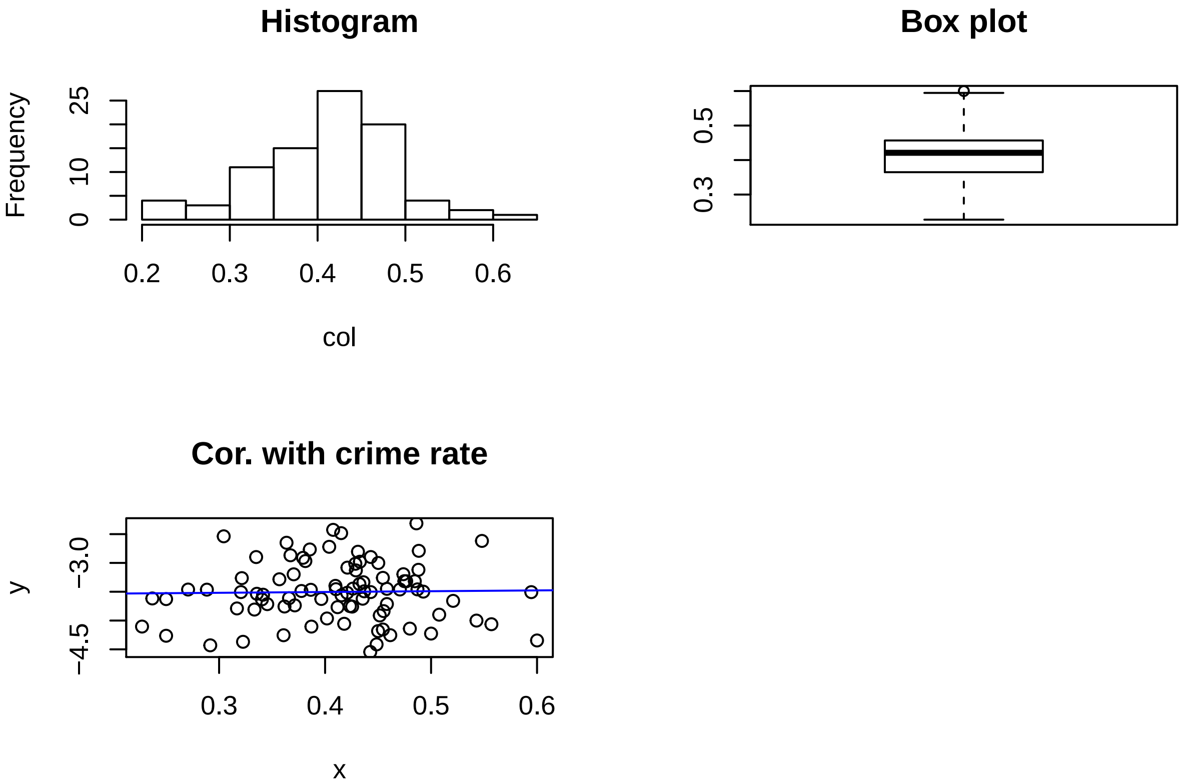
```
## [1] -0.2855235
```

Not much. We will include *log_prbconv* in our model.

Probability of prison sentence

```r
f_describe_col(crime_data$prbpris)
```

```
## [1] "Correlation:  0.0206"
```



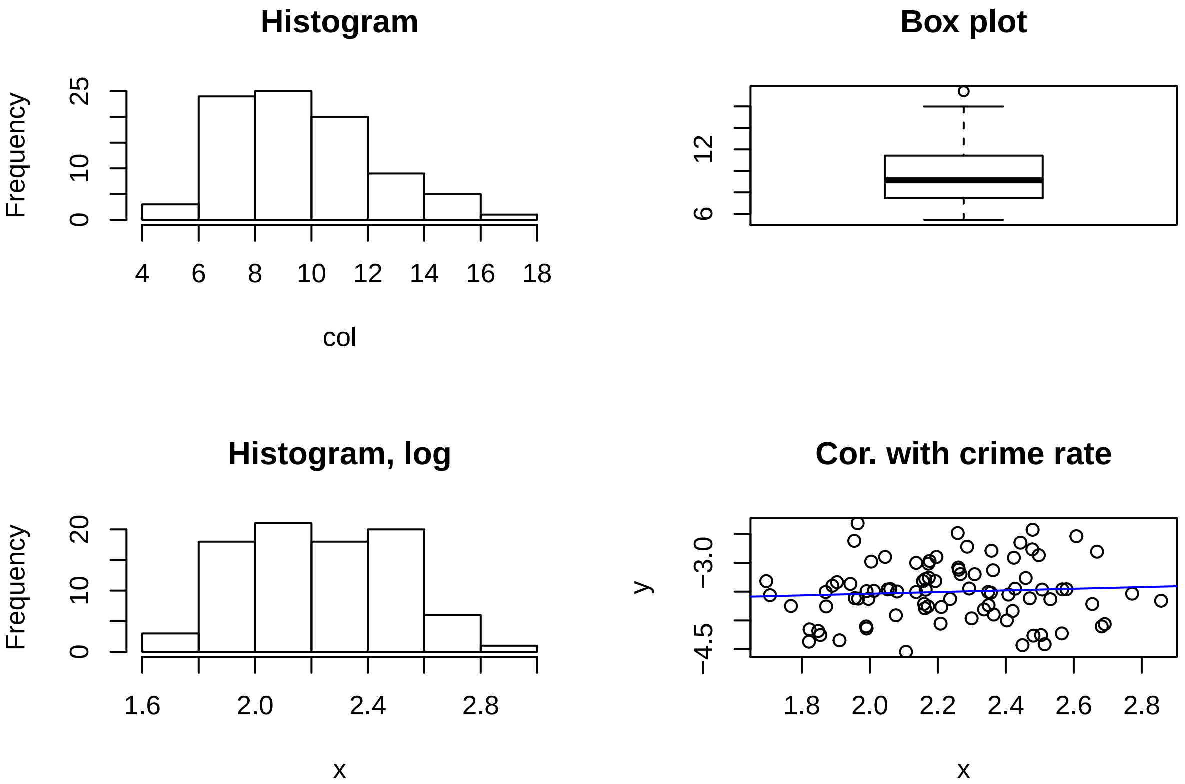
This histogram plot looks fairly normal and we don't observe any outliers. However, correlation is almost nonexistent wrt crime rate. This is interesting, because whether a crime results in spending time in prison does not seem to affect crime. This can shape government policies on whether to send criminals to prison or to find alternative ways to reform them.

We will *not* consider this variable in our model.

Average sentence duration

```
f_describe_col(crime_data$avgsen, do_log = TRUE)
```

```
## [1] "Correlation:  0.0741"
```



The average sentence in days looks slightly positive skewed, which we can correct with a log transform. But correlation is absent with respect to crime rate. It is interesting because we would expect that longer sentences would deter crime.

Perhaps we can use this data to make a policy recommendation to reduce sentences over long periods of time, or to be more lenient in pardoning criminals already serving long sentences.

We will *not* consider this variable in our model.

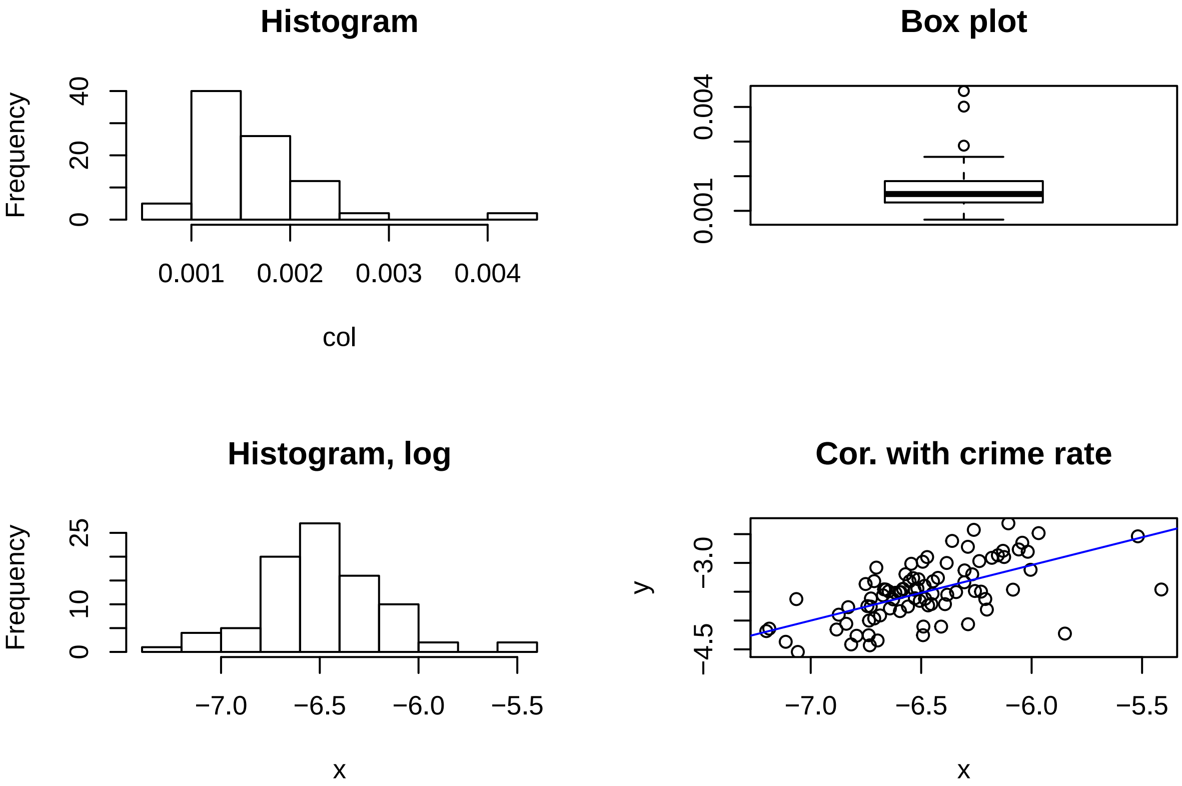
Police per capita

Note: In our first pass, we found an influential outlier with very low crime rate, even at very high police per capita. We removed it, as mentioned in the section on outliers.

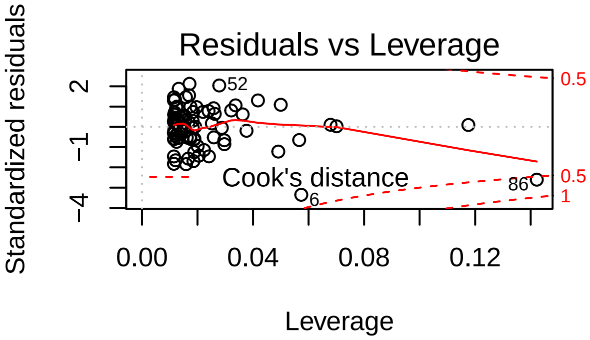
Police per capita has positive skew. Taking log helps:

```
f_describe_col(crime_data$polpc, do_log = TRUE, plot_model = TRUE)
```

```
## [1] "Correlation:  0.603"
```



```r
crime_data$log_polpc = log(crime_data$polpc)
```



The distribution looks better now. We see fairly strong positive correlation of 0.6 with crime rate: high number of police per capita is associated with high crime rate. It is probably a cause, rather than a result. More police may have been deployed to deal with higher amount of crime. If that is the case, it is worth questioning further why the additional police has not lowered the crime rate: are they ineffective?

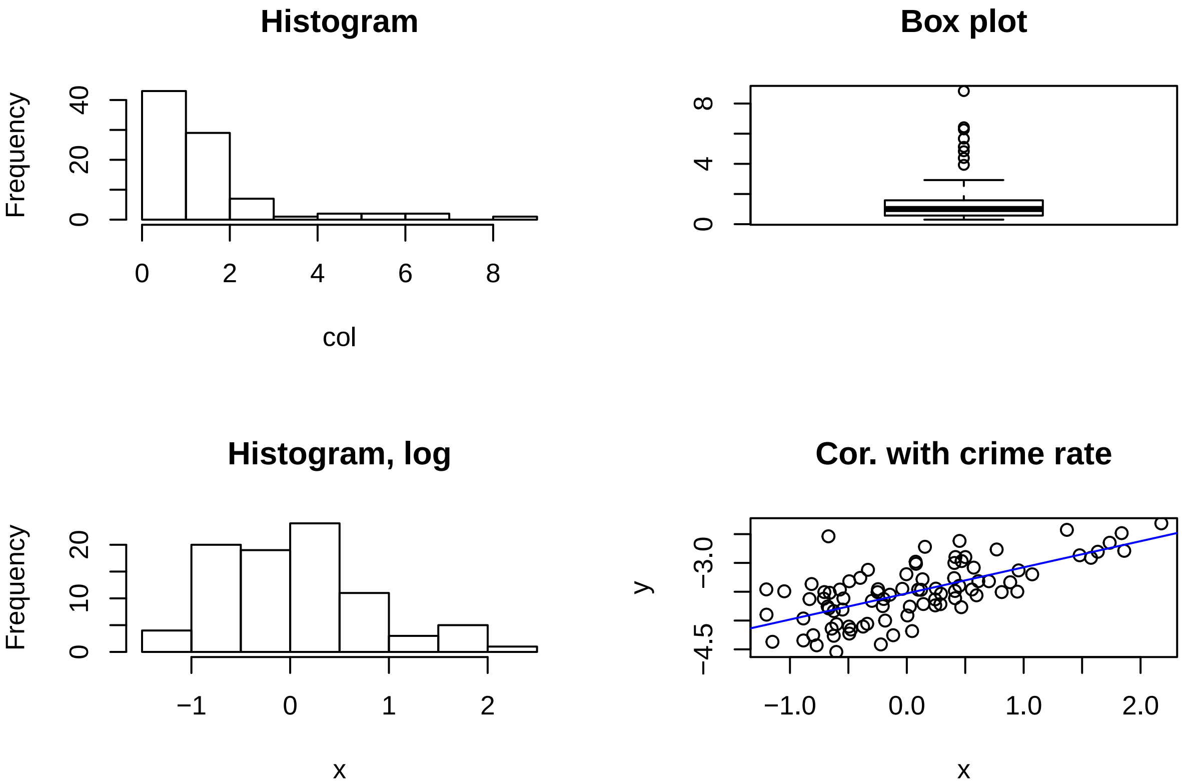
For our first model, we will *not* include this variable.

Population density

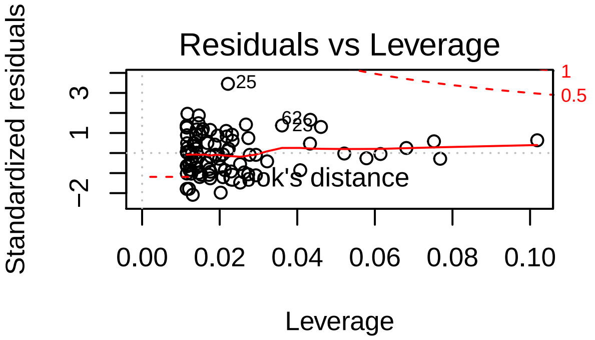
Note: In our first pass, we found an influential outlier with very low density and removed it, as mentioned in the section on outliers.

```r
f_describe_col(crime_data$density, do_log = TRUE, plot_model = TRUE)
```

```
## [1] "Correlation:  0.675"
```



```r
crime_data$log_density = log(crime_data$density)
```



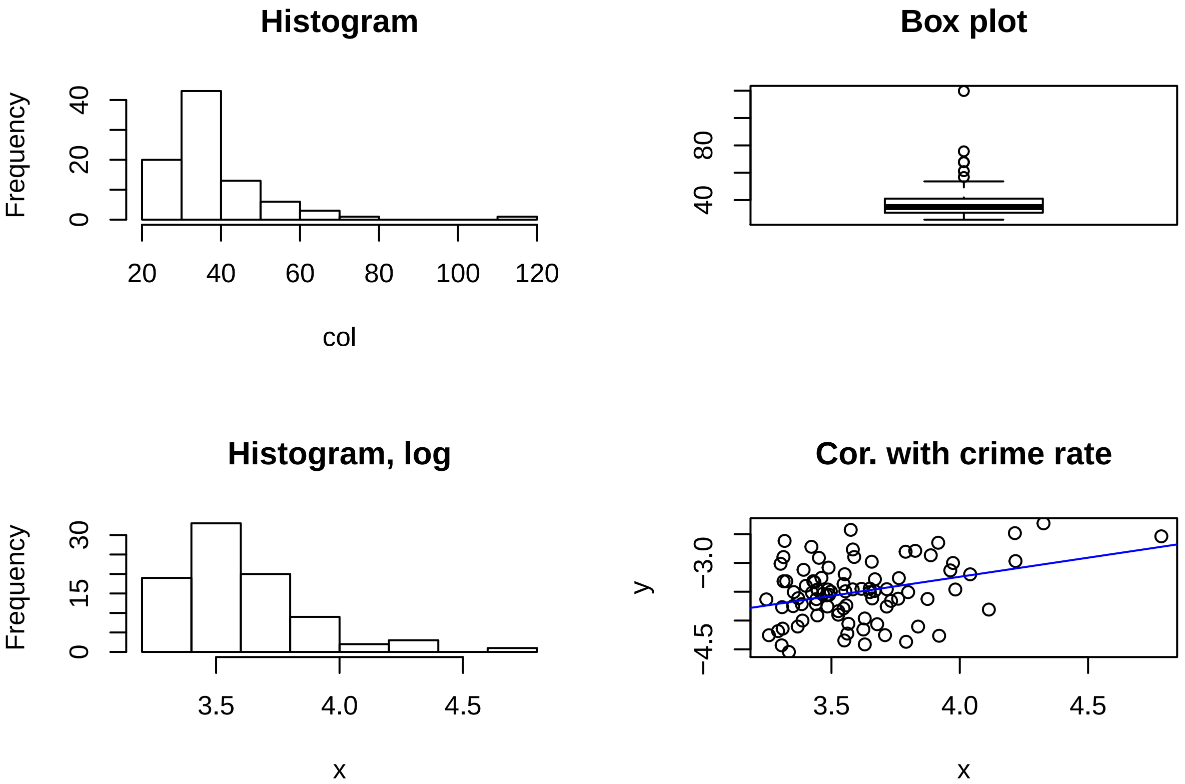
The histogram of density shows quite a bit positive skew. The log transformation shows a more promising normal distribution. There are no outliers with large leverage as measured by Cook's distance.

We see high positive correlation with crime rate. It may be that high population density indicates greater scope for hiding or cooperation in order to commit crime, indicating causality. We will surely consider this variable in our model.

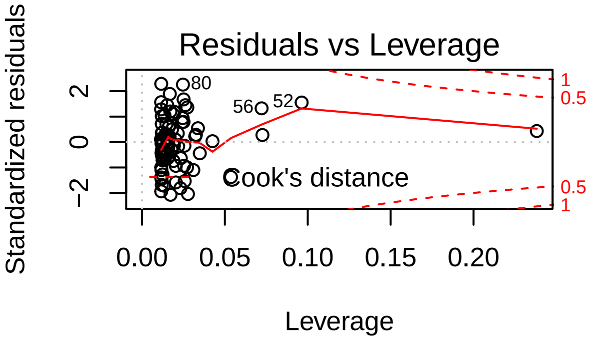
Tax revenue per capita

```r
f_describe_col(crime_data$taxpc, do_log = TRUE, plot_model = TRUE)
```

```
## [1] "Correlation:  0.347"
```



```r
crime_data$log_taxpc = log(crime_data$taxpc)
```



Tax revenue also shows positive skew, with one outlier indicating high tax revenue per capita (>100). It does not show a lot of leverage, however, so we will keep the value.

We also see considerable positive correlation with crime rate. It may be that tax revenue is a proxy for wealth, and high amount of wealth attracts crime. On the other hand, it is worth checking if we are spending tax dollars wisely in combating crime: if that were the case, counties with higher tax revenue would probably see lower crime.

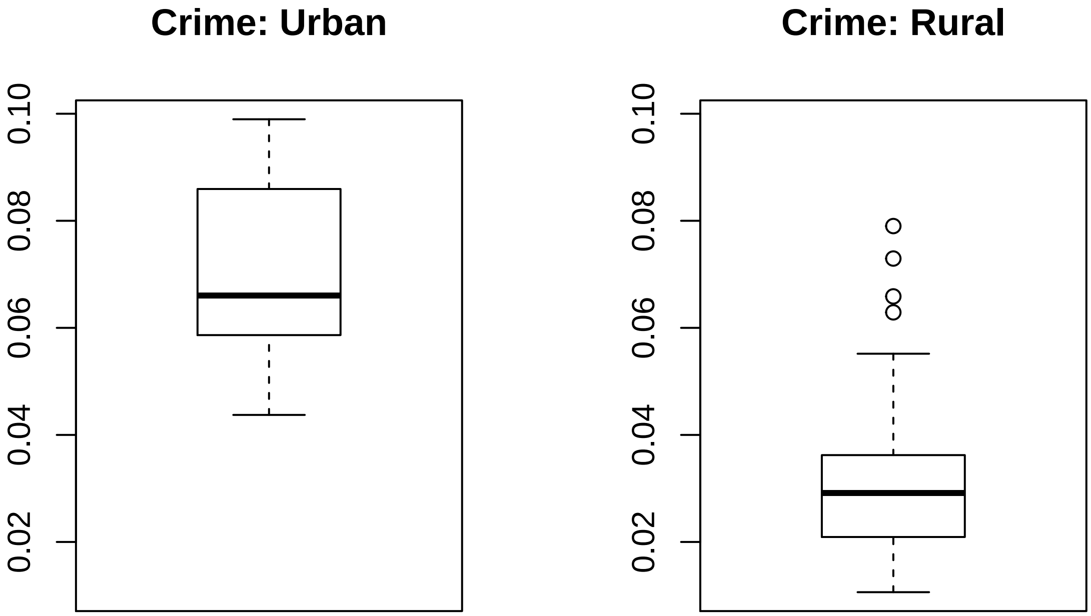
We will *not* include this variable in a first model.

Urban population

```r
print(length(crime_data$urban[crime_data$urban == 1]))
```

```
## [1] 8
```

```
urban_crime_data = crime_data %>% filter(urban == 1) %>% dplyr::select(-urban)
rural_crime_data = crime_data %>% filter(urban == 0) %>% dplyr::select(-urban)
par(mfrow = c(1, 2))
lmts = range(urban_crime_data$crmrte, rural_crime_data$crmrte)
boxplot(urban_crime_data$crmrte, main = "Crime: Urban", ylim = lmts)
boxplot(rural_crime_data$crmrte, main = "Crime: Rural", ylim = lmts)
```



It is worth noting that there are only 8 observations classified urban in this dataset. Median crime rate in urban regions is double that of rural regions.

Let us fit a model and see if our variable is salient.

```
print(cor(crime_data$urban, log(crime_data$crmrte)))
```

```
## [1] 0.513772
```

```
m = lm(log(crmrte) ~ factor(urban), data = crime_data)
print(summary(m))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ factor(urban), data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95860 -0.23686  0.03449  0.26254  1.04801
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.5861     0.0496 -72.307  < 2e-16 ***
## factor(urban)1   0.9030     0.1636   5.521 3.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4408 on 85 degrees of freedom
## Multiple R-squared:  0.264,  Adjusted R-squared:  0.2553
## F-statistic: 30.48 on 1 and 85 DF,  p-value: 3.593e-07
```

We do see a strong correlation between observations classified "urban" and crime rate, and the same is reflected by the low p-value in the model summary.

Let us check if there is correlation between "urban" and "density":

```
cor(crime_data$density, crime_data$urban)
```

```
## [1] 0.8218822
```

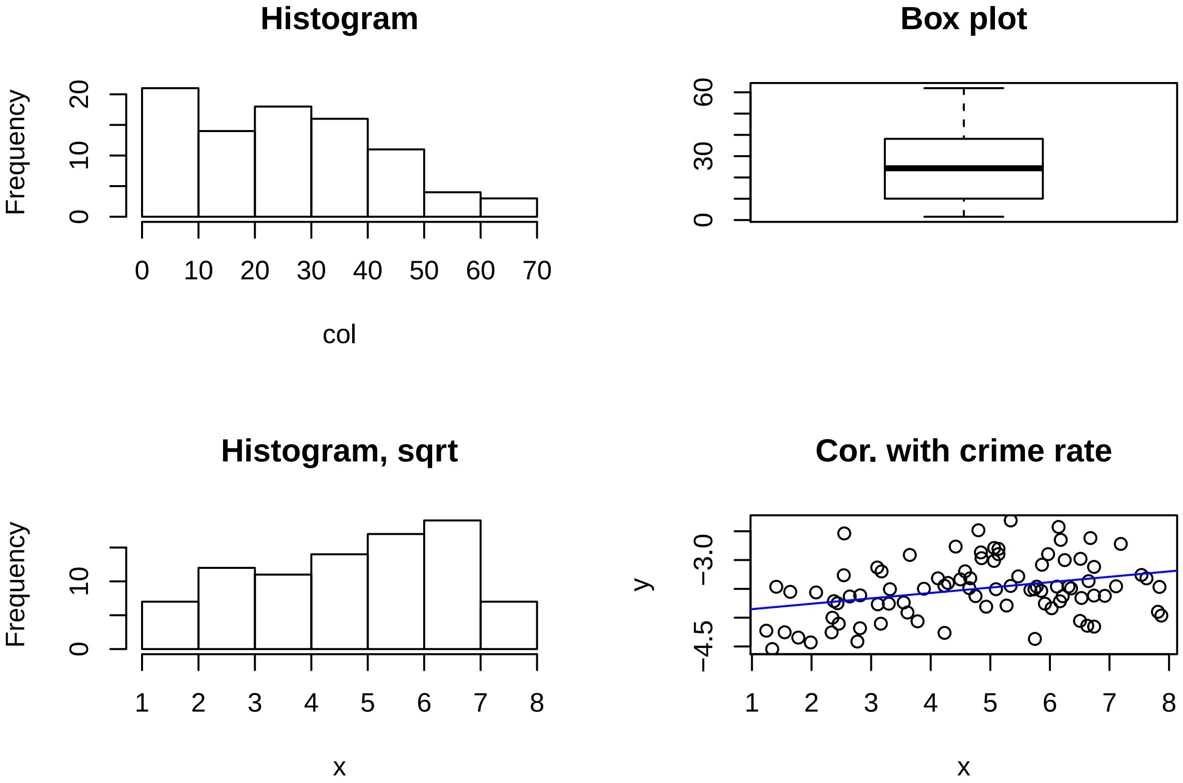
This is quite high, so we run a risk of multicollinearity.

Therefore, and since we have already selected density (with an additional advantage of more number of observations), we will *not* include this variable in our model.

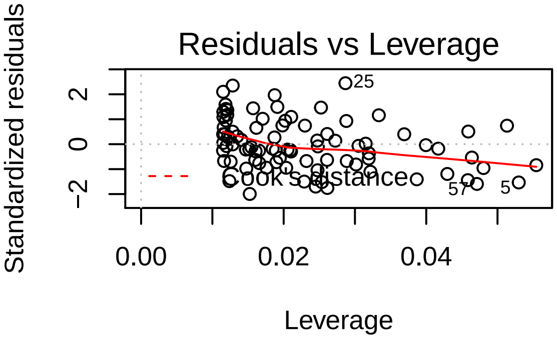
Percent minority

```
f_describe_col(crime_data$pctmin80, plot_model = TRUE, do_sqrt = TRUE)
```

```
## [1] "Correlation:  0.329"
```



```
crime_data$sqrt_pctmin80 = sqrt(crime_data$pctmin80)
```



Minority percentage has positive skew, but no outliers. Taking square root reshapes the distribution nicely.

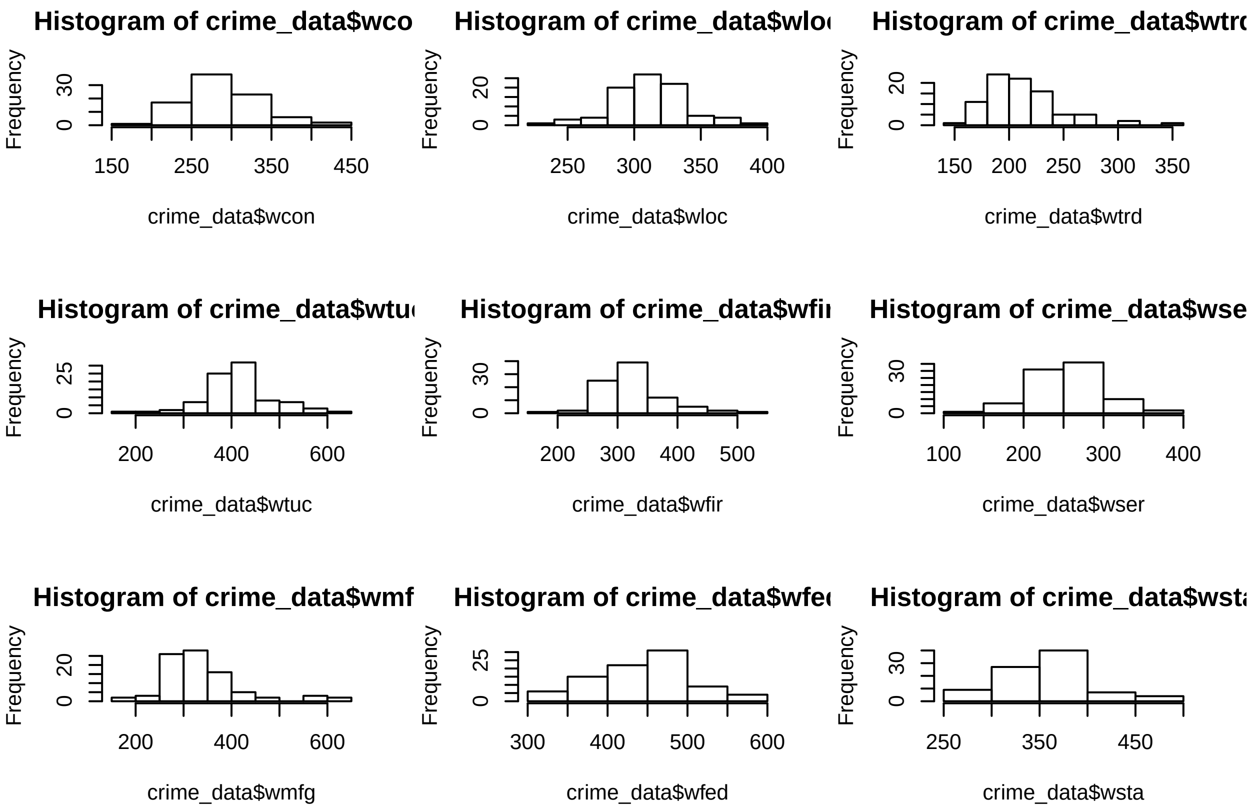
There is a fair amount of positive correlation with crime rate (0.27). It may be that as minorities increase, there is loss of social homogeneity and/or hate crime.

We will include this (transformed) variable in our model.

Wage distribution

Note: In our first pass, we found an influential outlier in services wages and removed it, as mentioned in the section on outliers.

```
par(mfrow = c(3, 3))
hist(crime_data$wcon)
hist(crime_data$wloc)
hist(crime_data$wtrd)
hist(crime_data$wtuc)
hist(crime_data$wfir)
hist(crime_data$wser)
hist(crime_data$wmfg)
hist(crime_data$wfed)
hist(crime_data$wsta)
```



Most of the wage variables conform to normal distributions. We do not have to worry about transformations.

Let us look which of them have high correlation with crime rate, considering all those with $R > 0.25$ (arbitrarily).

```
wage_cols = c("wcon", "wloc", "wtrd", "wtuc", "wfir", "wser", "wmfg", "wfed", "wsta")
cor(log(crime_data$crmrte), crime_data[, wage_cols])
```

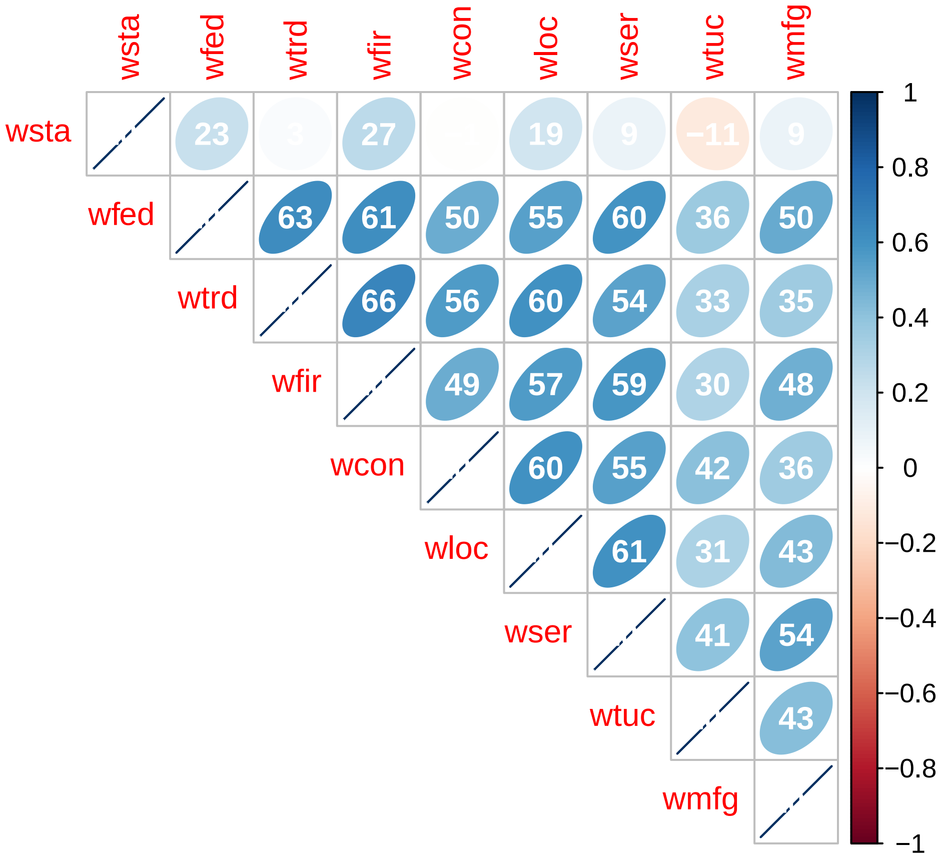
```
##           wcon      wloc     wtrd      wtuc      wfir      wser      wmfg
## [1,] 0.3296627 0.4173795 0.396715 0.2092351 0.2975497 0.3528975 0.3441131
##           wfed      wsta
## [1,] 0.5306463 0.2000234
```

This eliminates wsta and wtuc, but we are still left with 7 categories.

- wfed (0.50)
- wcon (0.37)
- wtrd (0.37)
- wser (0.34)
- wloc (0.28)
- wmfg (0.28)
- wfir (0.27)

As a different approach, let us check if the wages have high correlation among them. This will allow us to eliminate possible multi-collinearity.

```
corrplot(cor(crime_data[, wage_cols]), type = "upper", diag = TRUE, addCoef.col = "white",
    addCoefasPercent = TRUE, order = "hclust", method = "ellipse")
```



Indeed, a lot of the wage categories above have a high degree of correlation among them, but all are less than 70. We cannot eliminate any wage categories this way.

As a third approach, let us check for variance inflation instead:

```
m = lm(log(crmrte) ~ wfed + wcon + wtrd + wser + wloc + wmfg + wfir + wsta + wtuc,
    data = crime_data)
print(vif(m))
```

```
##     wfed     wcon     wtrd     wser     wloc     wmfg     wfir     wsta
## 2.304902 1.961994 2.541301 2.248561 2.247945 1.692487 2.452187 1.274484
##     wtuc
## 1.431049
```

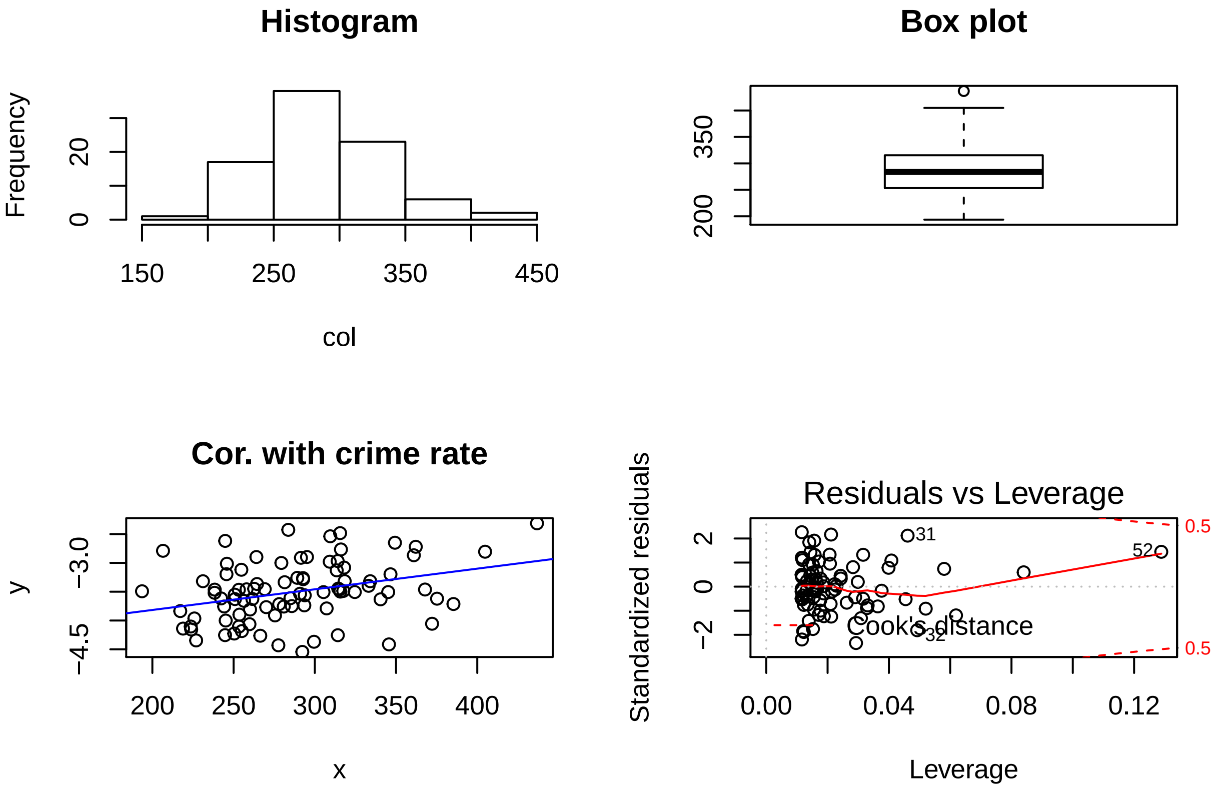
Again, no VIF is above 5. This procedure also does not eliminate any wage categories.

A general remark is in order for the positive correlation of crime across the wage categories. Higher wages may indicate higher wealth or a different omitted variable, and cannot be causal in and of themselves.

We will *not* include wages in a first model.

```
f_describe_col(crime_data$wcon, plot_model = TRUE)
```

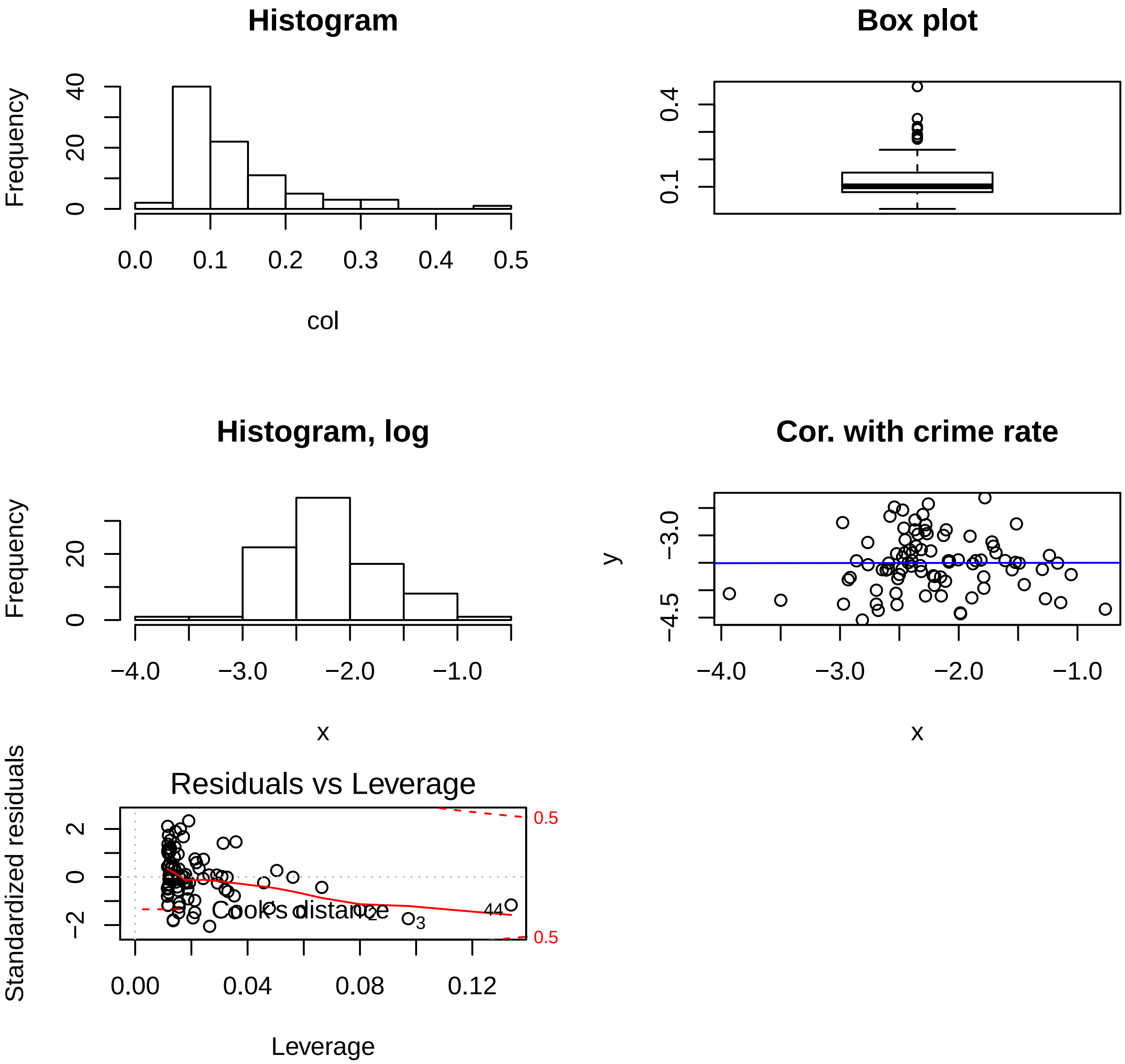
```
## [1] "Correlation:  0.33"
```



Offense Mix

```r
f_describe_col(crime_data$mix, do_log = TRUE, plot_model = TRUE)
```

```
## [1] "Correlation:  0.00224"
```



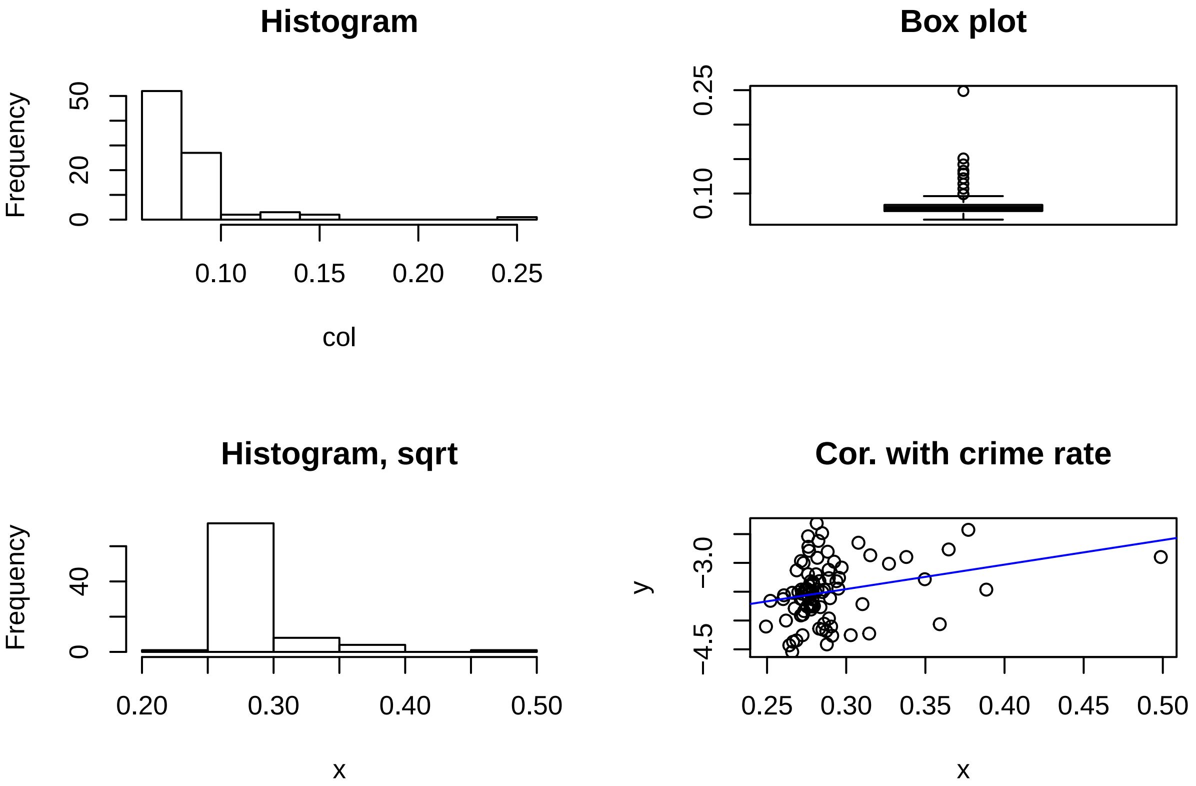
Offense mix does not seem to have any correlation with crime rate. The distribution is skewed, but a log transform fixes it. Outliers exist, but none have leverage as detected by Cook's distance.

We will *not* include offense mix in our models.

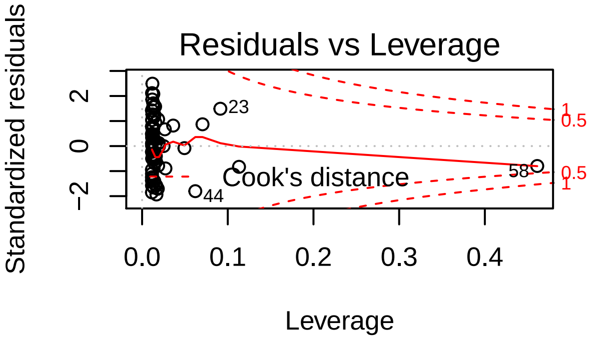
Percent of young males

```
f_describe_col(crime_data$pctymle, do_sqrt = TRUE, plot_model = TRUE)
```

```
## [1] "Correlation:  0.281"
```



```r
crime_data$sqrt_pctymle = sqrt(crime_data$pctymle)
```



We see moderate positive correlation with higher percentage of young males. There is positive skew, which we correct by taking a square root. Boxplot shows outliers, but none has outsized influence (Cook's distance < 0.5).

A high percentage of young males can indicate higher aggressiveness and risk, causing higher rate of crime. We will include this variable in our model.

Categorical variables

We have the following categorical variables in the dataset:

- Direction: west, central, other
- Urban or rural

We will use these to come up with separate models, based on different factors, later in this analysis. [TODO: Saurav]

Summary of variables

Here is a summary table of variables used in our models.

Variable	Transform?	Model1?	Model2?	Model3?	Remarks
county	N/A				Unused
year	N/A				Unused
prbarr		Y	Y	Y	
prbconv	log	Y	Y	Y	
prbpris				Y	No corr. found
avgsen				Y	No corr. found
polpc	log		Y	Y	Effect, not cause
density	log	Y	Y	Y	Causal
taxpc	log		Y	Y	Omit var: wealth
west	N/A				Categ, sep. model
central	N/A				Categ, sep. model
urban					Cor. with density
pctmin80	sqrt	Y	Y	Y	Causal
wcon			Y	Y	Omit var: wealth
wtuc				Y	Low corr. found
wtrd			Y	Y	
wfir			Y	Y	
wser			Y	Y	
wmfg			Y	Y	
wfed			Y	Y	
wsta				Y	Low corr. found
wloc			Y	Y	
mix	log			Y	No corr. found
pctymle	sqrt	Y	Y	Y	Causal, weak cor

5. Models

As outlined in the table above, here are three models we propose.

The first model includes what we think would be only causal variables:

$$\log\left(crmrte\right) = \beta_0 + \beta_1 \ prbarr + \beta_2 \ \log\left(prbconv\right) + \beta_3 \ \log\left(density\right) + \beta_4 \ \sqrt{pctmin80} + \beta_6 \ \sqrt{pctymle}$$

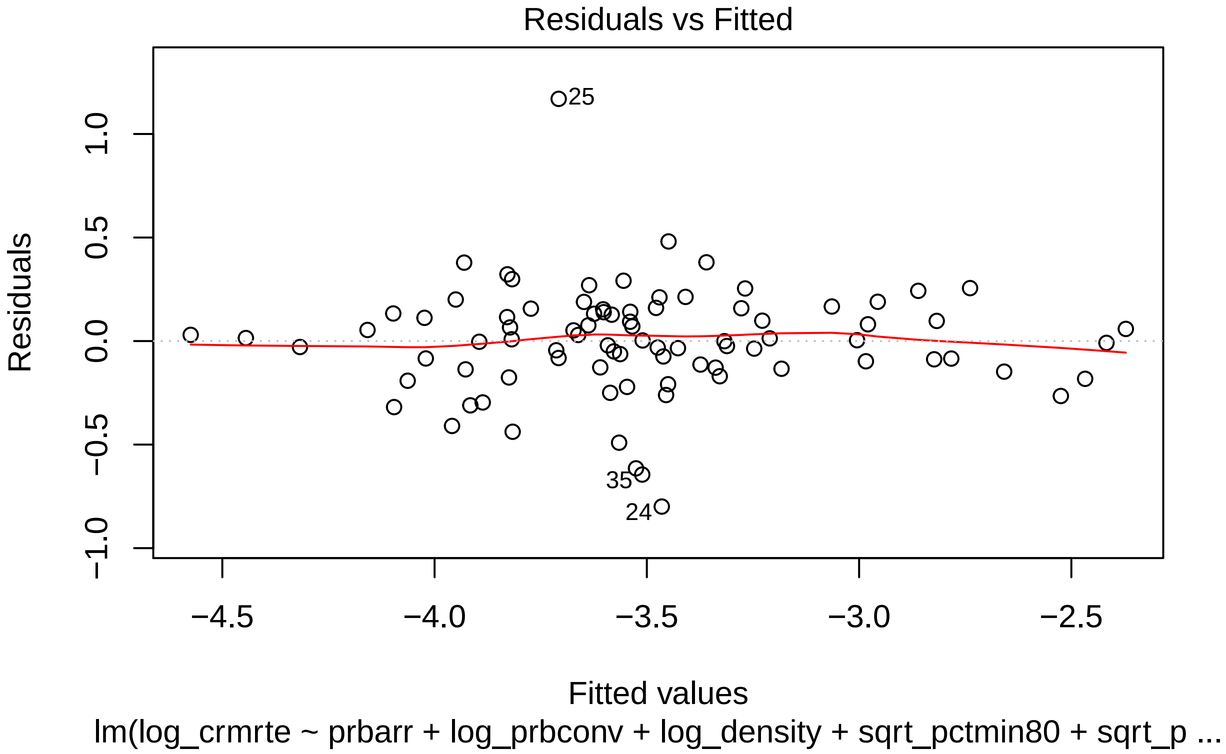
Let us fit the model:

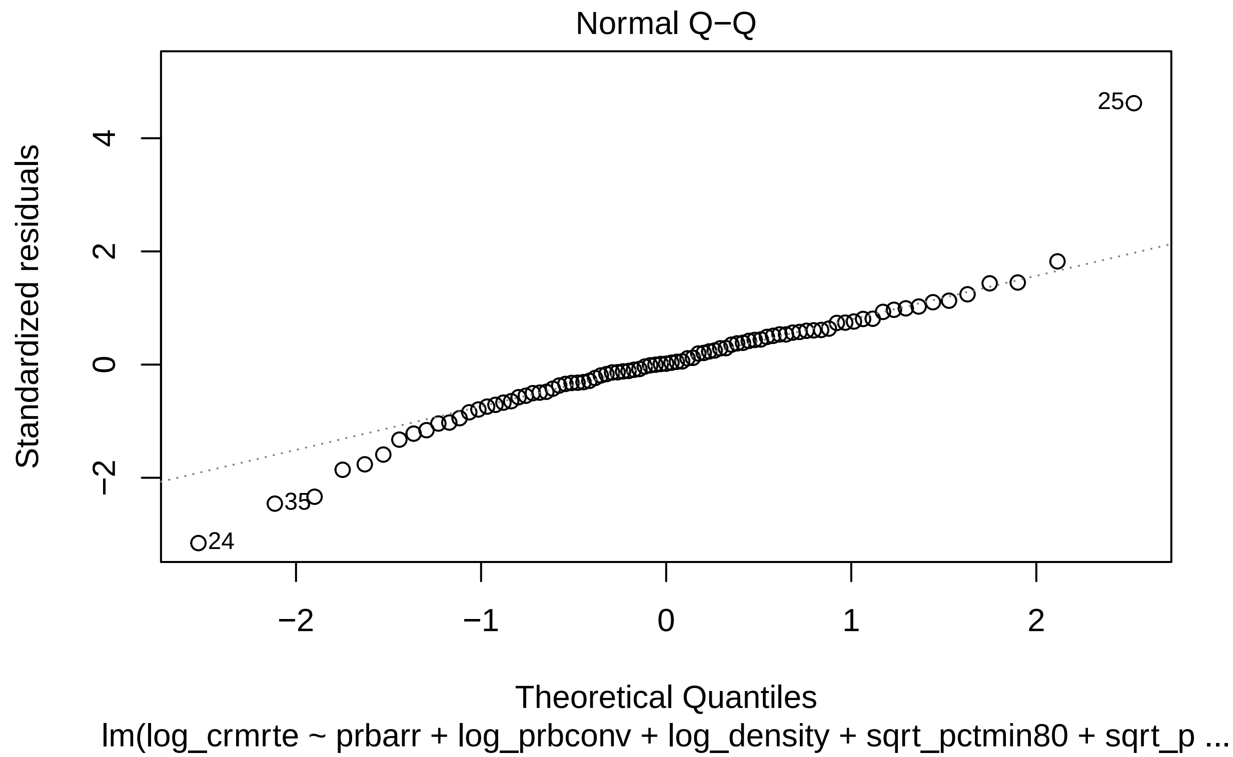
```r
model1 = lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_pctymle,
    data = crime_data)
summary(model1)
```

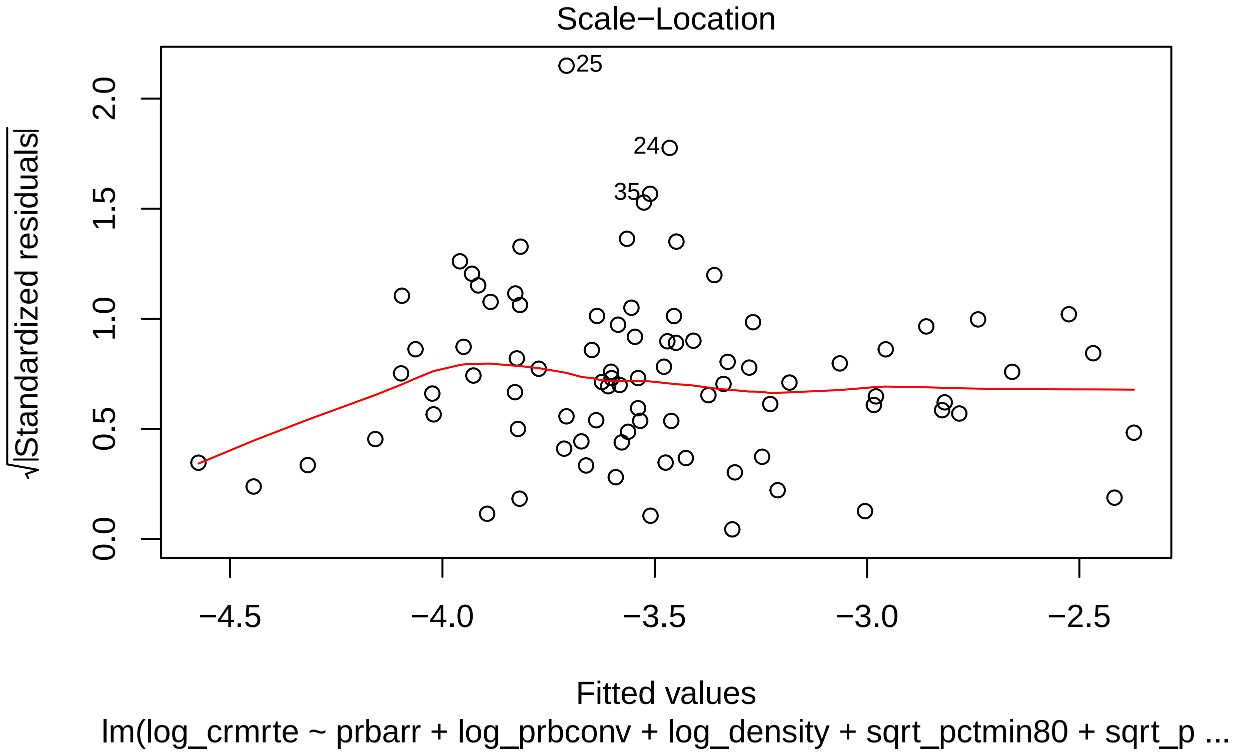
```
##
## Call:
## lm(formula = log_crmrte ~ prbarr + log_prbconv + log_density +
##     sqrt_pctmin80 + sqrt_pctymle, data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79921 -0.12778  0.00417  0.14039  1.16959
```

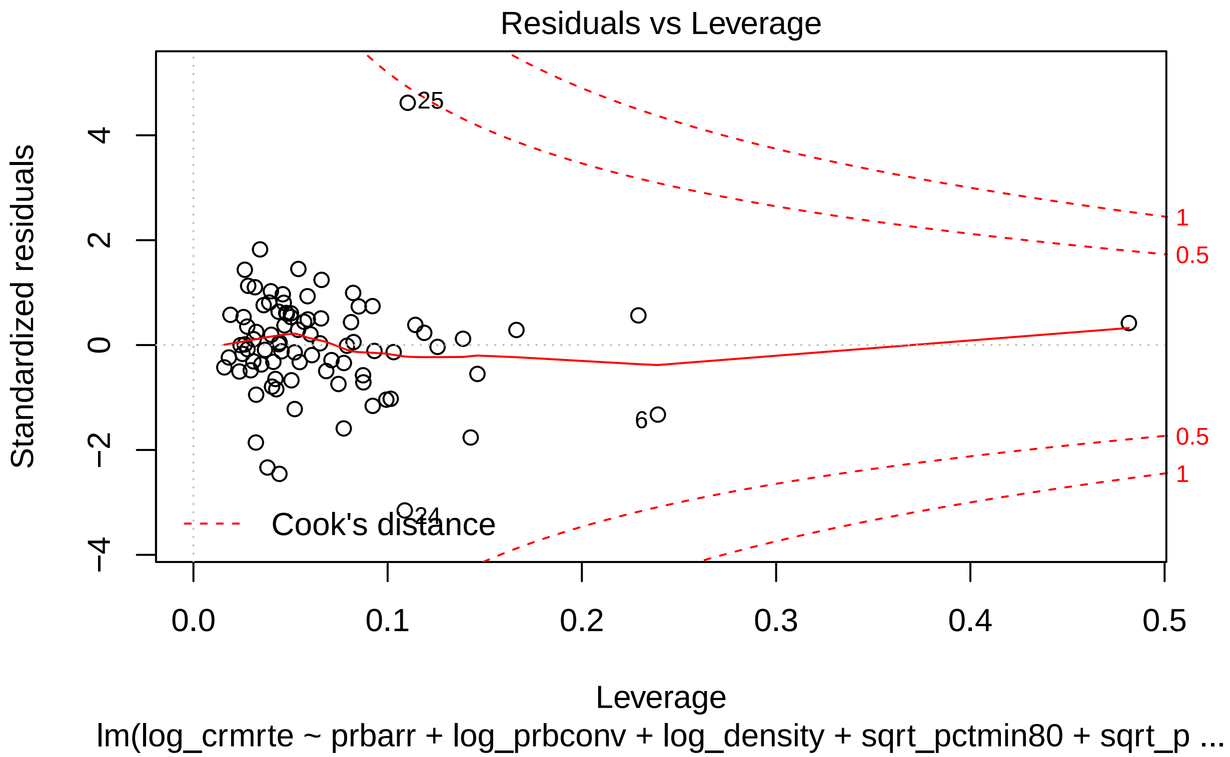
```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.09955    0.29059 -14.108  < 2e-16 ***
## prbarr        -1.58575    0.33018  -4.803 7.07e-06 ***
## log_prbconv   -0.31635    0.06276  -5.041 2.77e-06 ***
## log_density    0.35582    0.04337   8.205 2.96e-12 ***
## sqrt_pctmin80  0.12976    0.01650   7.866 1.38e-11 ***
## sqrt_pctymle   0.57581    0.91368   0.630     0.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2684 on 81 degrees of freedom
## Multiple R-squared:  0.7399, Adjusted R-squared:  0.7238
## F-statistic: 46.08 on 5 and 81 DF,  p-value: < 2.2e-16
```

```
plot(model1)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
−2
−1
0
1
2
25
35
24
Theoretical Quantiles
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
−4.5
−4.0
−3.5
−3.0
−2.5
25
24
35
Fitted values
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...



The model shows a fit of about 72% as measured by adjusted R^2. Observation #25 shows a large deviation from the normal, but is still not an outlier as per Cook's distance.

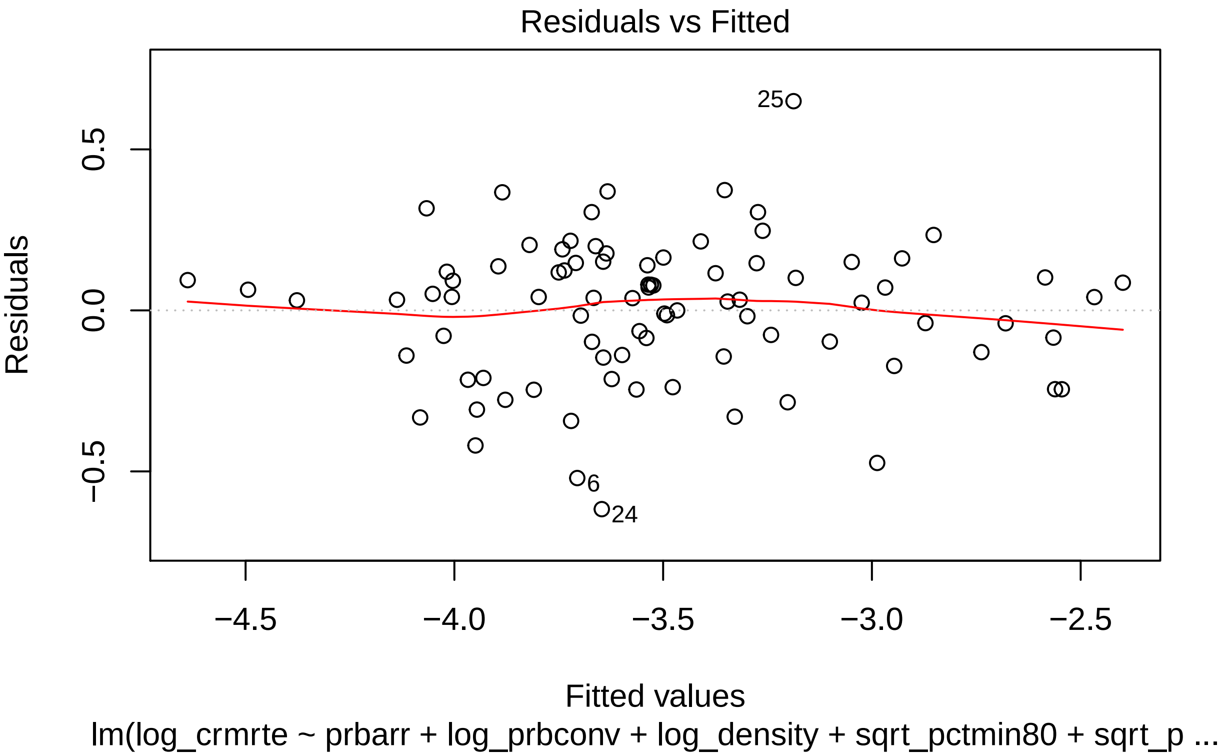
It is interesting that the model does not consider *pctymle* to be a significant predictor.

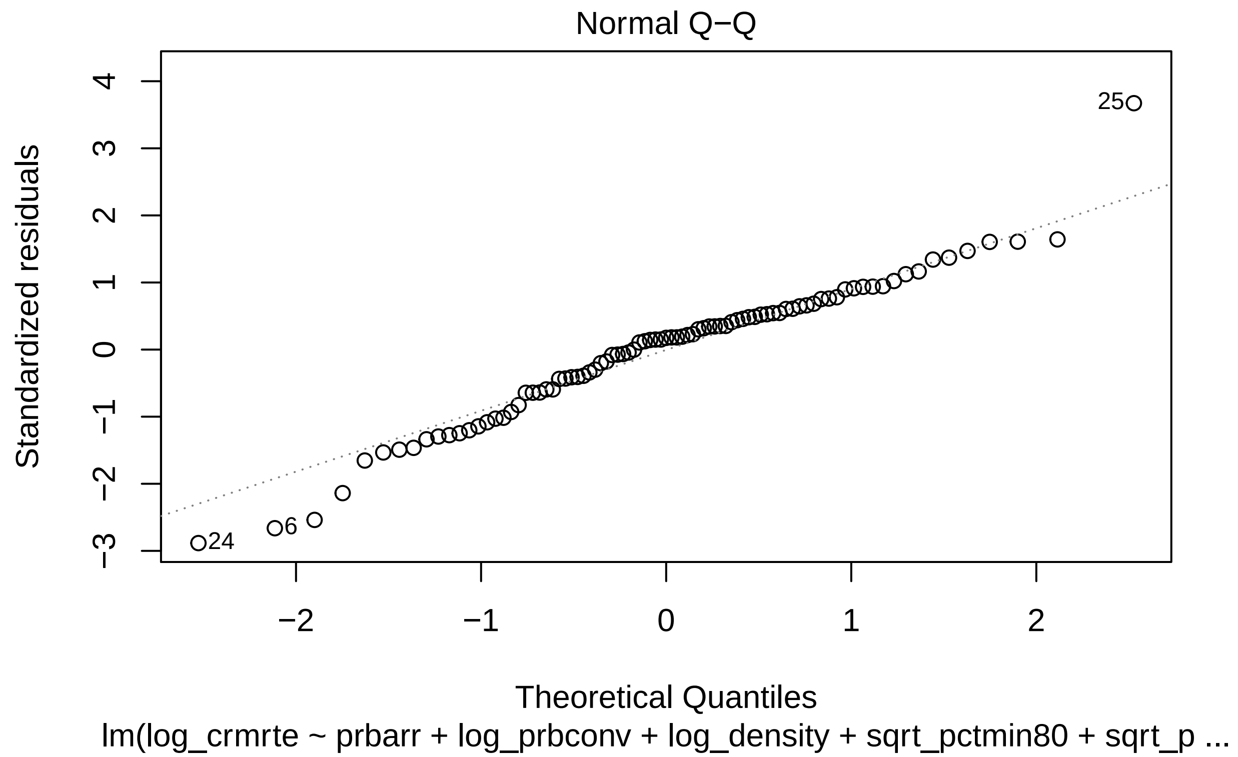
Next, let us include some more variables, as model 2.

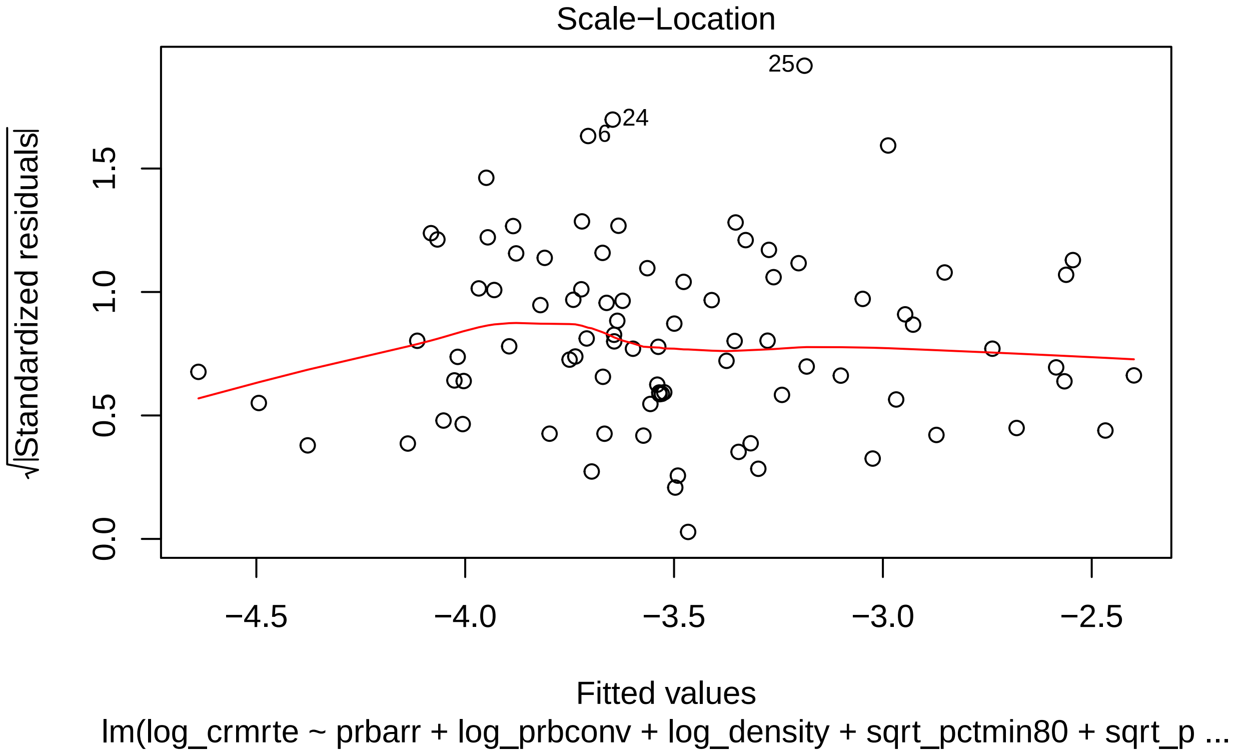
```
model2 = lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_pctymle +
    log_polpc + log_taxpc + wcon + wtrd + wfir + wser + wmfg + wfed + wloc, data = crime_data)
summary(model2)
```

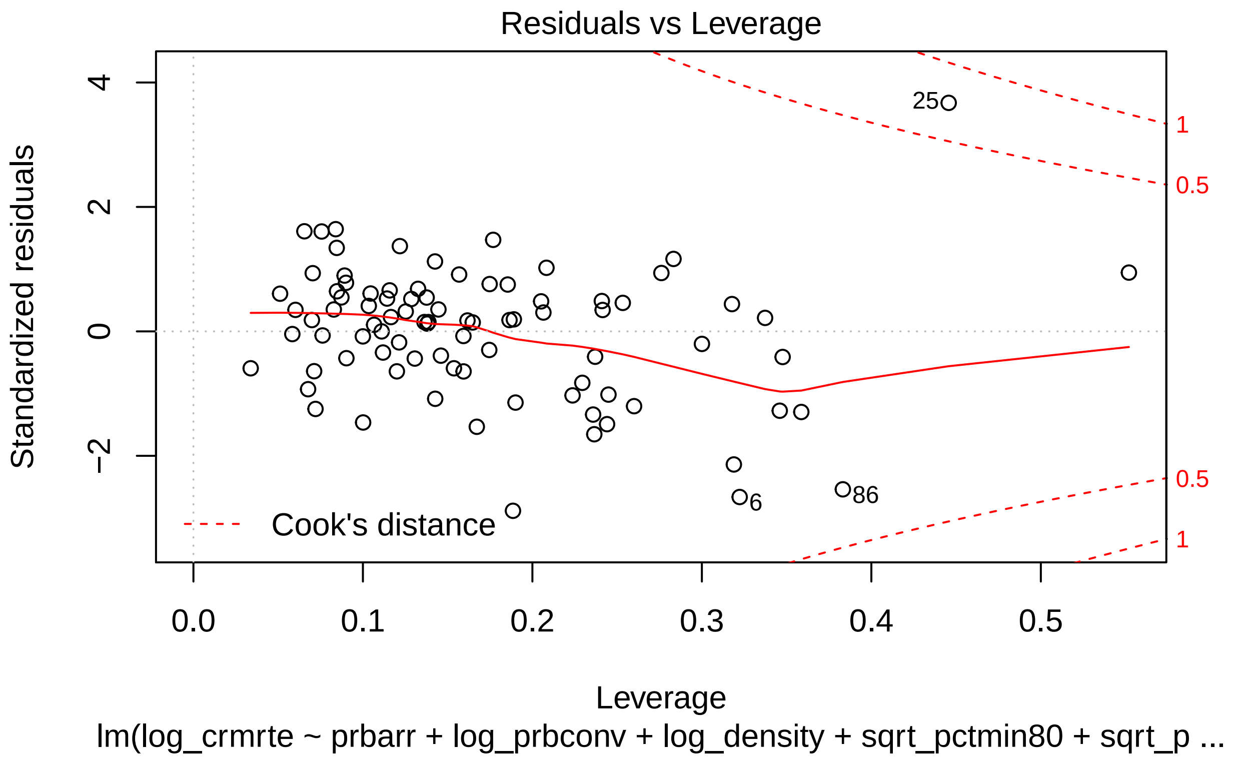
```
##
## Call:
## lm(formula = log_crmrte ~ prbarr + log_prbconv + log_density +
##     sqrt_pctmin80 + sqrt_pctymle + log_polpc + log_taxpc + wcon +
##     wtrd + wfir + wser + wmfg + wfed + wloc, data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61710 -0.13940  0.03807  0.13842  0.64958
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.8395094  1.3460863  -1.367 0.176015
## prbarr        -1.3874702  0.2975227  -4.663 1.40e-05 ***
## log_prbconv   -0.2140392  0.0618672  -3.460 0.000913 ***
## log_density    0.2941785  0.0588892   4.995 3.97e-06 ***
## sqrt_pctmin80  0.1157958  0.0159080   7.279 3.37e-10 ***
## sqrt_pctymle   0.9586338  0.9233789   1.038 0.302660
## log_polpc      0.4192690  0.1199977   3.494 0.000818 ***
## log_taxpc      0.0971032  0.1339898   0.725 0.470981
## wcon           0.0007047  0.0007492   0.941 0.350054
## wtrd           0.0009411  0.0012629   0.745 0.458577
## wfir          -0.0016378  0.0007354  -2.227 0.029074 *
## wser          -0.0017679  0.0009284  -1.904 0.060867 .
## wmfg           0.0001428  0.0003861   0.370 0.712507
## wfed           0.0009439  0.0007784   1.213 0.229189
## wloc           0.0006421  0.0013986   0.459 0.647546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2376 on 72 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.7837
## F-statistic: 23.26 on 14 and 72 DF,  p-value: < 2.2e-16
```

```
plot(model2)
```




Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
25
24
6
−2
−1
0
1
2
Theoretical Quantiles
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
25
24
6
−4.5
−4.0
−3.5
−3.0
−2.5
Fitted values
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...



The fit improves to 78% (adjusted R^2). The only wages that are significant are $wfir$ and $wser$. $taxpc$ is not significant either.

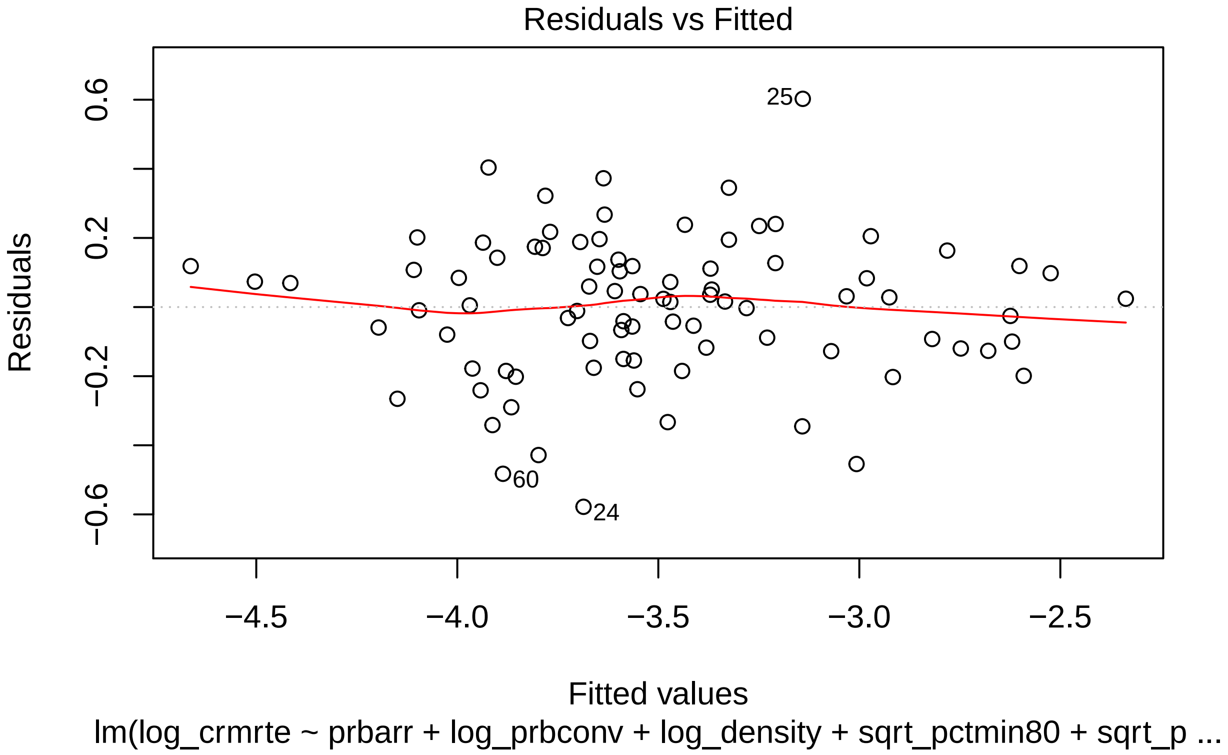
Observation #25 causes considerable problems to this model as well.

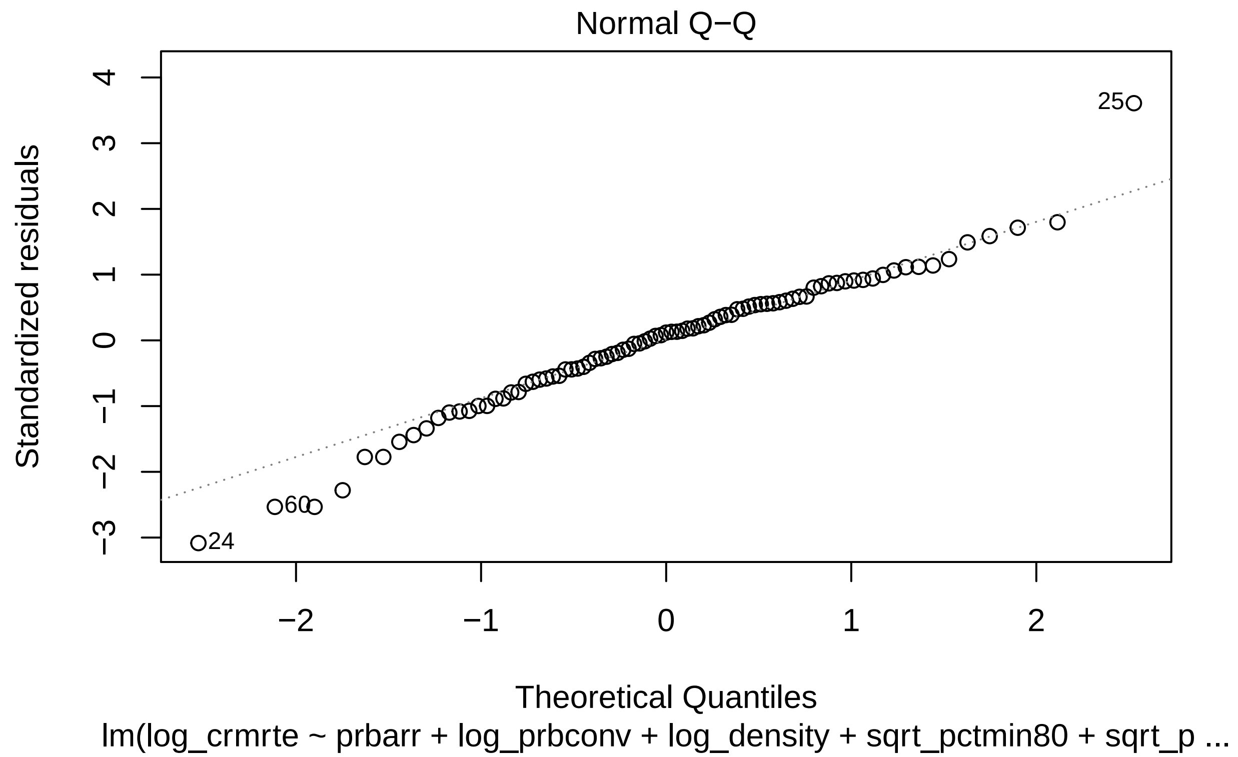
We will now build a third model that includes almost all the variables:

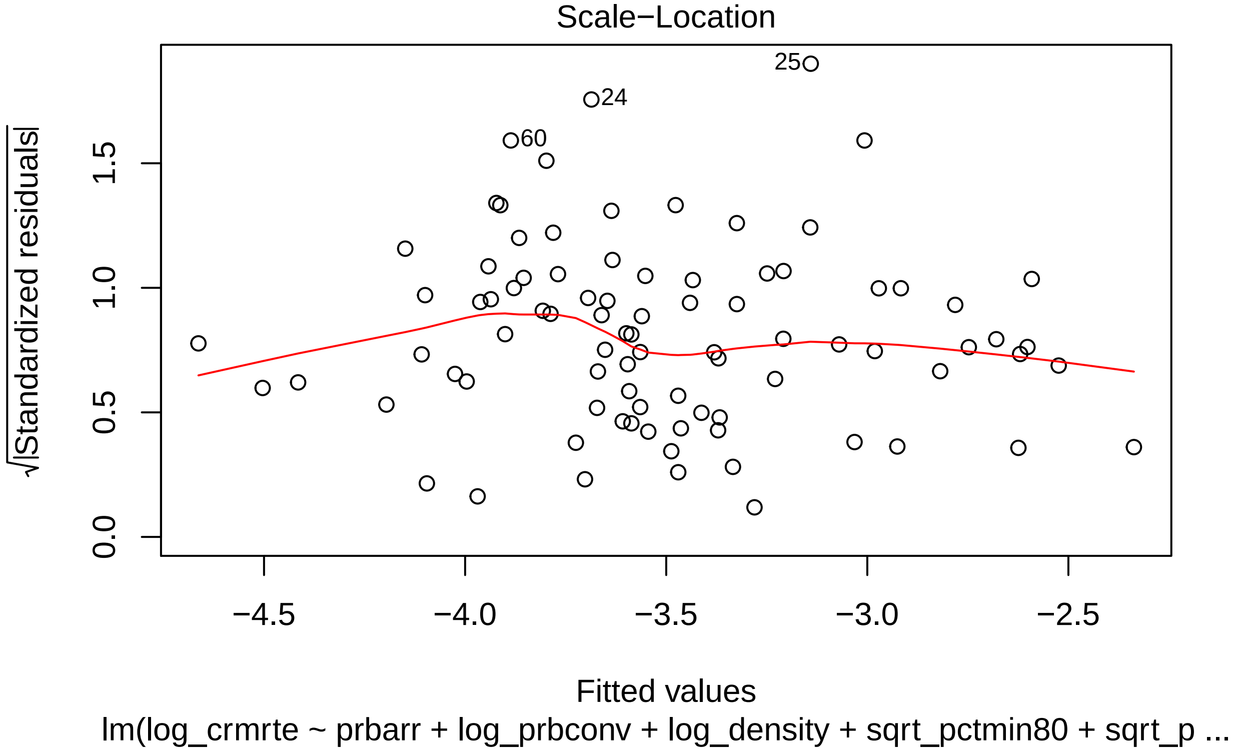
```
model3 = lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_pctymle +
    log_polpc + log_taxpc + wcon + wtrd + wfir + wser + wmfg + wfed + wloc + prbpris +
    avgsen + wtuc + wsta + mix, data = crime_data)
summary(model3)
```

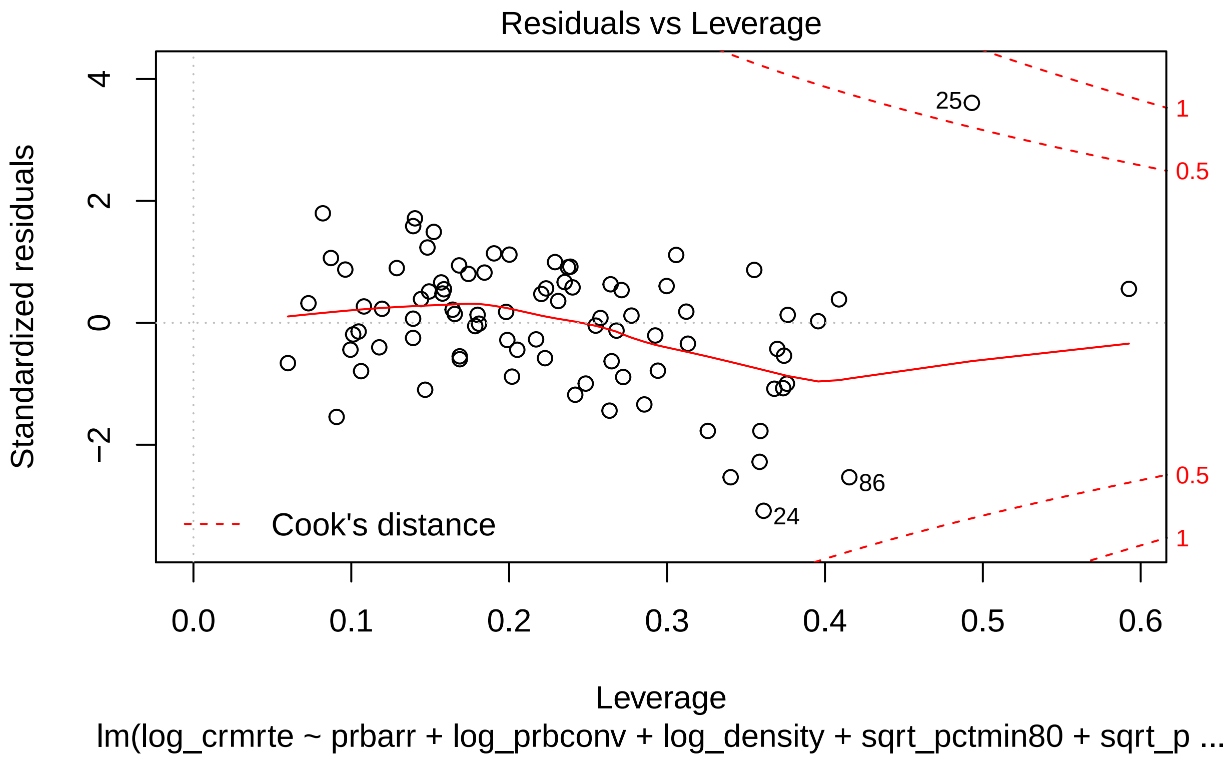
```
##
## Call:
## lm(formula = log_crmrte ~ prbarr + log_prbconv + log_density +
##     sqrt_pctmin80 + sqrt_pctymle + log_polpc + log_taxpc + wcon +
##     wtrd + wfir + wser + wmfg + wfed + wloc + prbpris + avgsen +
##     wtuc + wsta + mix, data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57800 -0.12329  0.02358  0.12301  0.60252
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.6957414  1.4559944  -0.478 0.634315
## prbarr        -1.3827745  0.3151706  -4.387 4.16e-05 ***
## log_prbconv   -0.2077696  0.0657308  -3.161 0.002362 **
## log_density    0.2862921  0.0583452   4.907 6.21e-06 ***
## sqrt_pctmin80  0.1161967  0.0163157   7.122 9.17e-10 ***
## sqrt_pctymle   1.1156541  0.9466294   1.179 0.242744
## log_polpc      0.5200949  0.1308611   3.974 0.000175 ***
## log_taxpc      0.0934828  0.1337919   0.699 0.487147
## wcon           0.0002746  0.0007772   0.353 0.724971
## wtrd           0.0013435  0.0013318   1.009 0.316717
## wfir          -0.0013795  0.0007766  -1.776 0.080212 .
## wser          -0.0022459  0.0009411  -2.386 0.019845 *
## wmfg          -0.0000987  0.0004027  -0.245 0.807108
## wfed           0.0010306  0.0008057   1.279 0.205296
## wloc           0.0007644  0.0014219   0.538 0.592645
## prbpris       -0.5620162  0.3789196  -1.483 0.142708
## avgsen        -0.0266055  0.0118630  -2.243 0.028227 *
## wtuc           0.0003366  0.0004252   0.792 0.431383
## wsta          -0.0002038  0.0007302  -0.279 0.781000
## mix           -0.1284043  0.4583911  -0.280 0.780249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2345 on 67 degrees of freedom
## Multiple R-squared:  0.8358, Adjusted R-squared:  0.7892
## F-statistic: 17.95 on 19 and 67 DF,  p-value: < 2.2e-16
```

```
plot(model3)
```




Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
25
60
24
−2
−1
0
1
2
Theoretical Quantiles
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
25
24
60
−4.5
−4.0
−3.5
−3.0
−2.5
Fitted values
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...



This tops fit at about 0.79 (adjusted R^2). It also says *avgsen* and *wser* are significant.

```
stargazer(model1, model2, model3, title = "Regression models")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 01, 2018 - 00:20:22

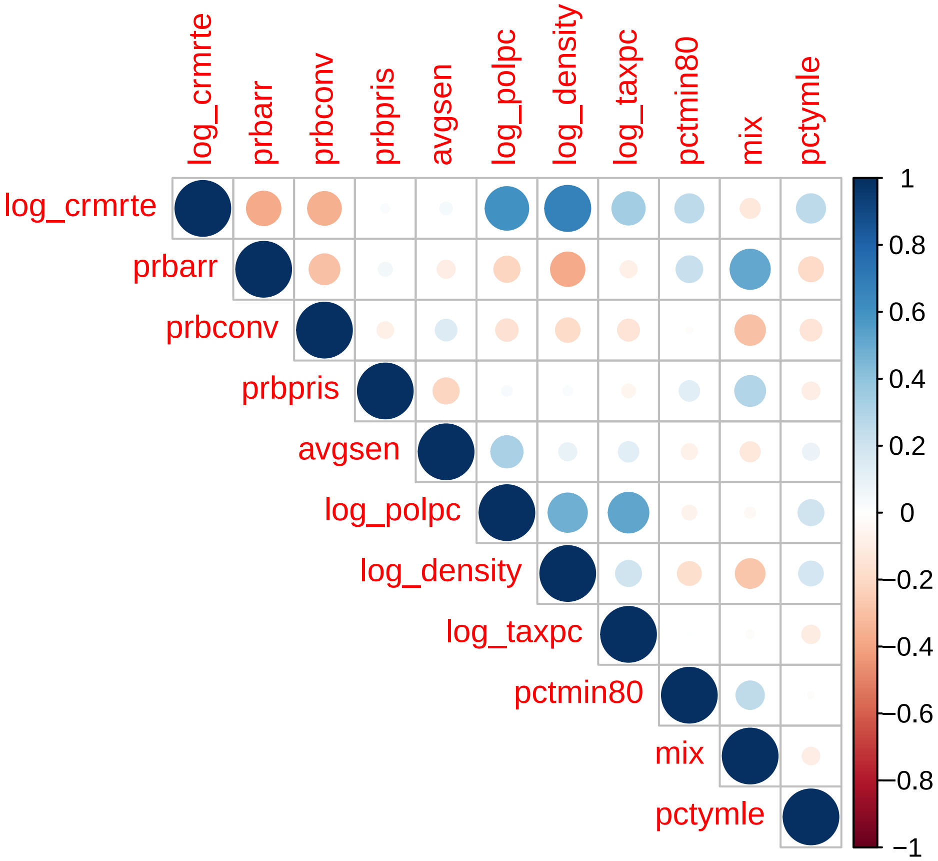
Bi-variate Analysis

The correlation plot between the different variables is as follows:

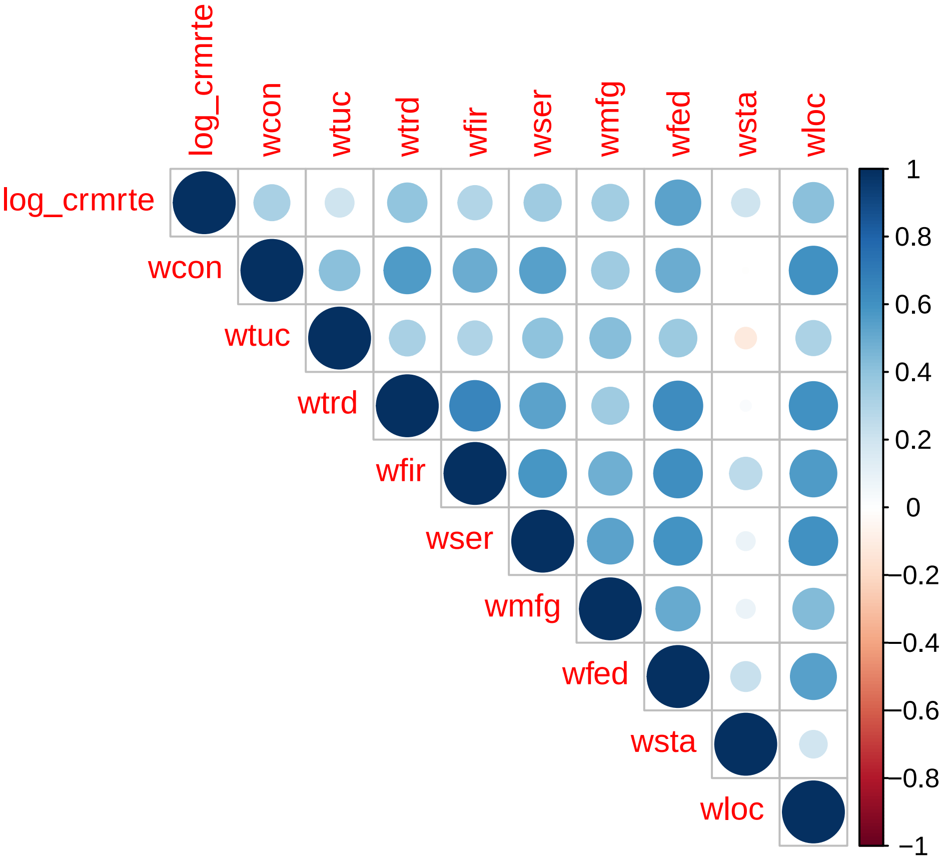
```
corrplot(cor(crime_data[, c("log_crmrte", "prbarr", "prbconv", "prbpris", "avgsen",
    "log_polpc", "log_density", "log_taxpc", "pctmin80", "mix", "pctymle")]), type = "upper")
```

Table 2: Regression models

	Dependent variable:		
	log_crmrte		
	(1)	(2)	(3)
prbarr	−1.586***	−1.387***	−1.383***
	(0.330)	(0.298)	(0.315)
log_prbconv	−0.316***	−0.214***	−0.208***
	(0.063)	(0.062)	(0.066)
log_density	0.356***	0.294***	0.286***
	(0.043)	(0.059)	(0.058)
sqrt_pctmin80	0.130***	0.116***	0.116***
	(0.016)	(0.016)	(0.016)
sqrt_pctymle	0.576	0.959	1.116
	(0.914)	(0.923)	(0.947)
log_polpc		0.419***	0.520***
		(0.120)	(0.131)
log_taxpc		0.097	0.093
		(0.134)	(0.134)
wcon		0.001	0.0003
		(0.001)	(0.001)
wtrd		0.001	0.001
		(0.001)	(0.001)
wfir		−0.002**	−0.001*
		(0.001)	(0.001)
wser		−0.002*	−0.002**
		(0.001)	(0.001)
wmfg		0.0001	−0.0001
		(0.0004)	(0.0004)
wfed		0.001	0.001
		(0.001)	(0.001)
wloc		0.001	0.001
		(0.001)	(0.001)
prbpris			−0.562
			(0.379)
avgsen			−0.027**
			(0.012)
wtuc			0.0003
			(0.0004)
wsta			−0.0002
			(0.001)



```
corrplot(cor(crime_data[, c("log_crmrte", "wcon", "wtuc", "wtrd", "wfir", "wser",
    "wmfg", "wfed", "wsta", "wloc")]), type = "upper")
```



We can see that there is a high positive correlation between:

- log of crime rate vs. log of policy per capita, log of tax revenue per capita, log of density and percent young male
- log of crime rate vs. most of the wage variables

And there is a high negative correlation between:

- log of crime rate vs. probability of arrests and conviction

The positive correlation observed makes sense for the following reasons:

1) More densely populated regions tends to observe more crimes
2) More wealthy areas (more wages and taxes) tend to have more crimes
3) More crimes leads to more police presence in a particular county to monitor and reduce crime rate

The negative correlations can be further observed using:

```
par(mfrow=c(1,2))
plot(crime_data$prbarr, crime_data$log_crmrte,
     main="Probability of arrest", ylab="Log Crime rate", xlab="Prob on arrest")
abline(lm(crime_data$log_crmrte ~ crime_data$prbarr))
plot(crime_data$prbconv, crime_data$log_crmrte,
     main="Probability of conviction vs. crime rate", ylab="Log Crime rate",
     xlab="Prob on conv")
abline(lm(crime_data$log_crmrte ~ crime_data$prbconv))
```

As seen above, as the probability of arrests and conviction go down, there are more criminals on the loose which leads to higher crime rates observed

TODO: Talk about other possible correlations here?

TODO: Discuss other interesting bi-variate analysis?

3. Model Specification and Assumptions

In our earlier analysis, we observed some key relationships between crime rate and other variables presented. Some of these variables had high positive correlation to crime rate while some others exhibited strong negative correlation.

For our first simple model, we will choose a subset of these variables that we believe are most important determinants of crime rate.

Model 1

$$log(CrimeRate) = \beta_0 + \beta_1 log(Density) + \beta_2(YoungMale) + \beta_3(Minority) + u$$

It is common knowledge that areas with higher density have more crime. Therefore we include that factor in our model. Similarly we hypothesized that crime rate is high among minority and young male population, so we round off our model with that factored in as well.

```
model1 = lm(log(crmrte) ~ (log_density)+pctymle+pctmin80, data=crime_data)
model1$coefficients
par(mfrow=c(2,2))
plot(model1)
AIC(model1)
summary(model1)$r.squared
summary(model1)$adj.r.squared
```

Model 2

high probability of arrests and conviction act as deterrents to crime.

$$log(CrimeRate) = \beta_0 + \beta_1 log(Density) + \beta_2(YoungMale) + \beta_3(Minority) + \beta_4(Conviction) + \beta_5(Arrest) + \beta_6(Tax) + u$$

```
model2 = lm(log(crmrte) ~ (log_density)+pctymle+pctmin80+adj_prbarr+
            adj_prbconv+taxpc, data=crime_data)
model2$coefficients
par(mfrow=c(2,2))
plot(model2)
AIC(model2)
summary(model2)$r.squared
```

Model 3

everything

```
model3 = lm(log(crmrte) ~ (log_density)+pctymle+pctmin80+adj_prbarr+adj_prbconv+
             taxpc+log_polpc, data=crime_data)
model3$coefficients
par(mfrow=c(2,2))
```

```
plot(model3)
AIC(model3)
summary(model3)$r.squared
```

```
stargazer(model1, model2, model3)
```

5. Discussion of omitted variables (Identify what you think are the 5-10 most important omitted variables that bias results you care about.)

Education

Unemployment

Poverty