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 and cleanup

Let us read the data and have a first look.

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

```
summary(crime_data_raw)
tail(crime_data_raw, n=8)
```

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.

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

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.

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.

```
# remove NA rows
crime_data = na.omit(crime_data_raw)
# convert factor to numeric for variable prbconv
crime_data$prbconv = as.numeric(levels(crime_data$prbconv)[crime_data$prbconv])
crime_data = crime_data %>% dplyr::select(-c(year, county))
# remove duplicate record
duplicated(crime_data)[duplicated(crime_data)==TRUE]
```

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

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

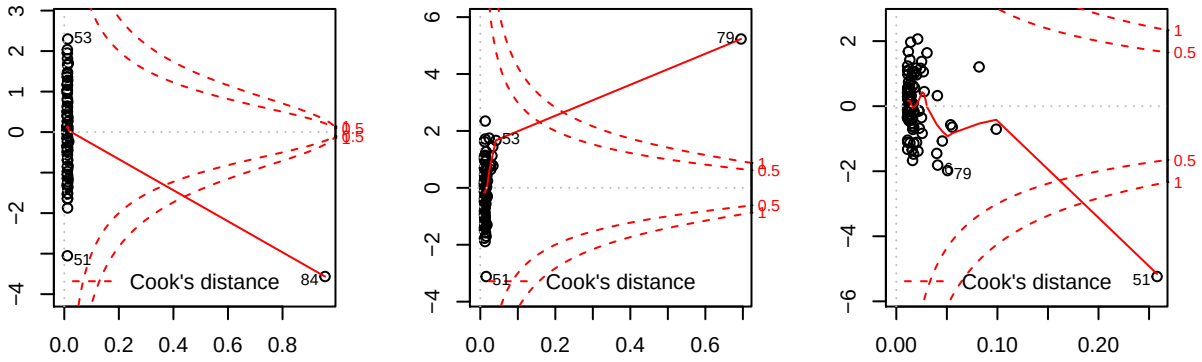
3. Influential outliers

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

We found the following observations to have outsized influence as measured by Cook's distance (> 1):

- *wser*, observation #84
- *density*, observation #79
- *polpc*, observation #51

```
par(mfrow=c(2,3), mai=c(0.35,0.35,0.35,0.35))
m = lm(log(crime_data$crmrte) ~ crime_data$wser)
plot(m, which=5, caption=NA)
m = lm(log(crime_data$crmrte) ~ log(crime_data$density))
plot(m, which=5, caption=NA)
m = lm(log(crime_data$crmrte) ~ log(crime_data$polpc))
plot(m, which=5, caption=NA)
```



Finally, observation #25 causes outliers in the final model fit due to a lot of influence. It shows very high crime rate, very high police per capita, very low density, highest tax per capita, very low minority. It is an observation with too many quirks and may merit separate investigation.

Here are all the outlier observations:

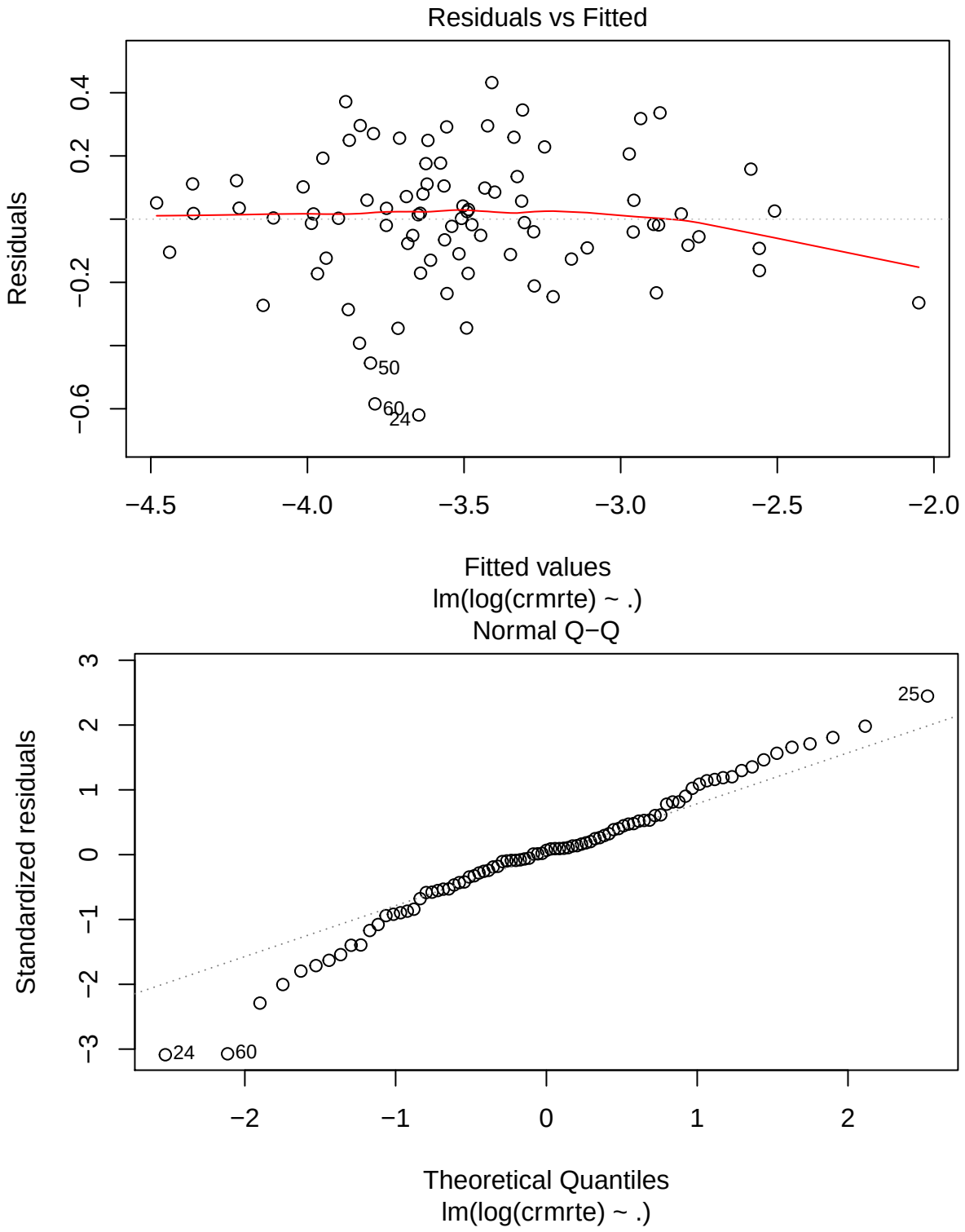
```
crime_data %>%
  slice(c(25, 51, 79, 84)) %>%
  select(crmrte, density, central, urban, polpc, wser, taxpc, pctmin80)
```

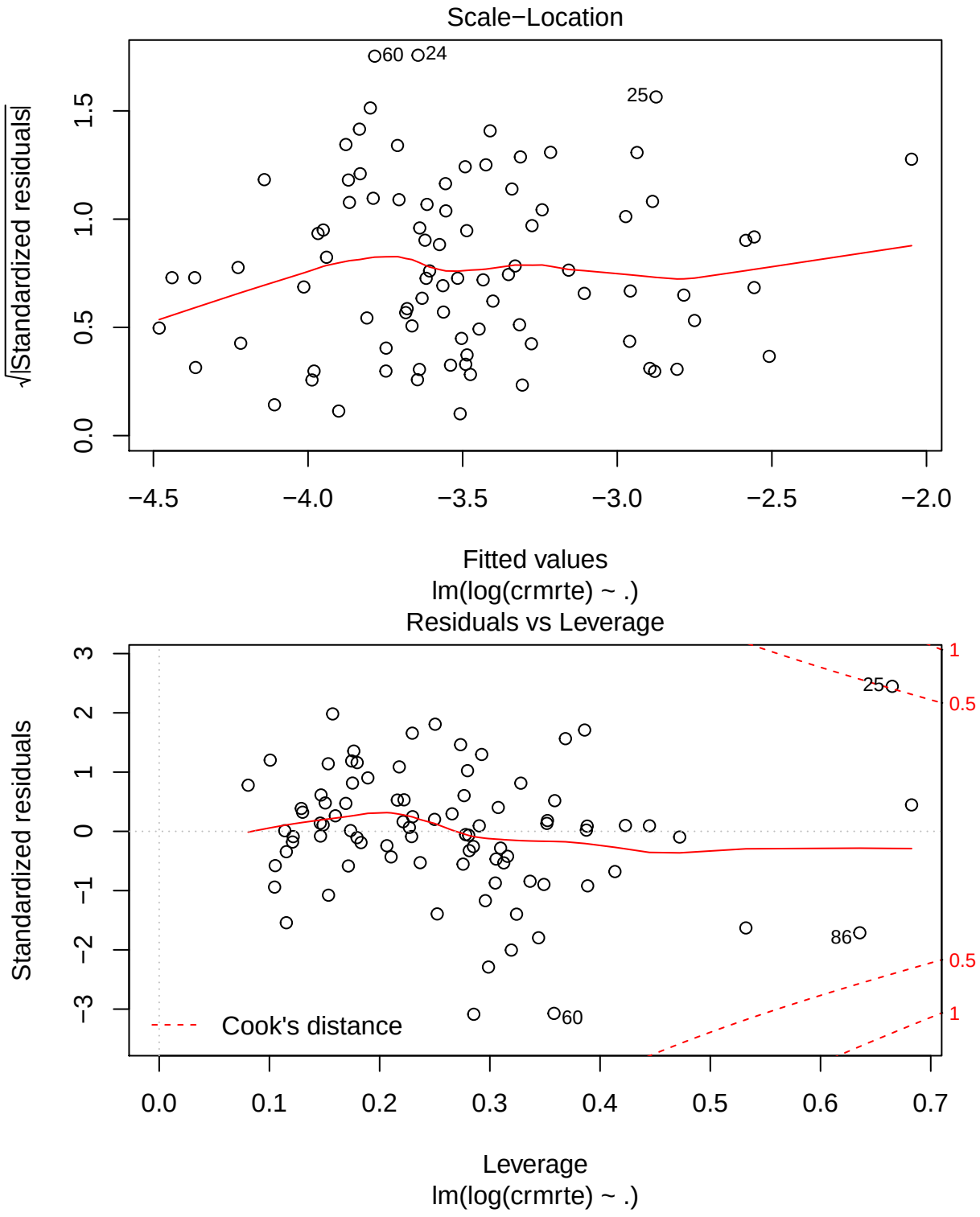
```
## # A tibble: 4 x 8
##    crmrte   density central urban   polpc  wser taxpc pctmin80
##     <dbl>     <dbl>   <int> <int>   <dbl> <dbl> <dbl>    <dbl>
## 1 0.0790  0.512           0     0 0.00401   221 120       6.50
## 2 0.00553 0.386           0     0 0.00905   245  28.2     1.28
## 3 0.0140  0.0000203       0     0 0.00316   204  37.7    25.4
## 4 0.0109  0.389           1     0 0.00122  2177  40.8    64.3
```

The first observation is a quirky one with many unusual variable values. The second observation above has very low crime rate at very high police per capita. The third has extremely low density. The fourth has extremely high wages in the service industry.

The observations are questionable and affect our model because of their high influence. We will remove them.

```
crime_data <- crime_data %>% slice(-c(51, 79, 84))
m = lm(log(crmrte) ~ ., data=crime_data)
plot(m)
```


Residuals vs Fitted
Residuals
0.4
0.2
−0.2
−0.6
50
60
24
−4.5
−4.0
−3.5
−3.0
−2.5
−2.0
Fitted values
lm(log(crmrte) ~ .)
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
25
24
60
−2
−1
0
1
2
Theoretical Quantiles
lm(log(crmrte) ~ .)



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(2,4), mai=c(0.35,0.35,0.35,0.35))
  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",
       col=c("green", "red"),
       xlab="predictor (green)",
       ylab="crime (red)")
  if (is.numeric(col))
    abline(m, col="blue")
  if (plot_model == TRUE) {
    # To reduce #pages in report
    # plot(m, which=5, caption=NA)
  }
}
```

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.

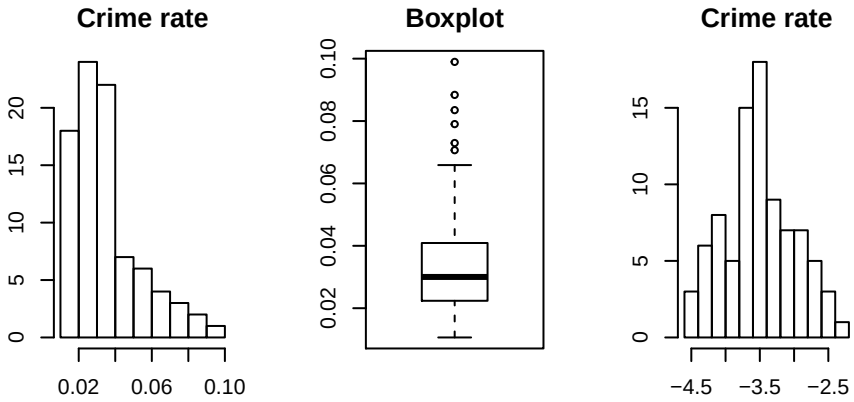
We will now try to get a sense of each variable in the dataset.

Single variable analysis

Crime rate

Crime rate is the key dependent variable of interest.

```
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
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.

Crime rate is mostly low, but there are some observations that show high crime rate (positive outliers). This causes skew.

```
crime_data %>% filter(crmrte > 0.07) %>% select(density, central, urban, prbarr, prbconv)
```

```
## # A tibble: 6 x 5
##   density central urban prbarr prbconv
##     <dbl>   <int> <int>  <dbl>   <dbl>
## 1   3.93        0     1  0.155   0.260
## 2   0.512       0     0  0.225   0.208
## 3   5.67        1     1  0.133   0.459
## 4   8.83        1     1  0.149   0.348
## 5   6.29        0     1  0.237   0.393
## 6   1.57        1     0  0.183   0.343
```

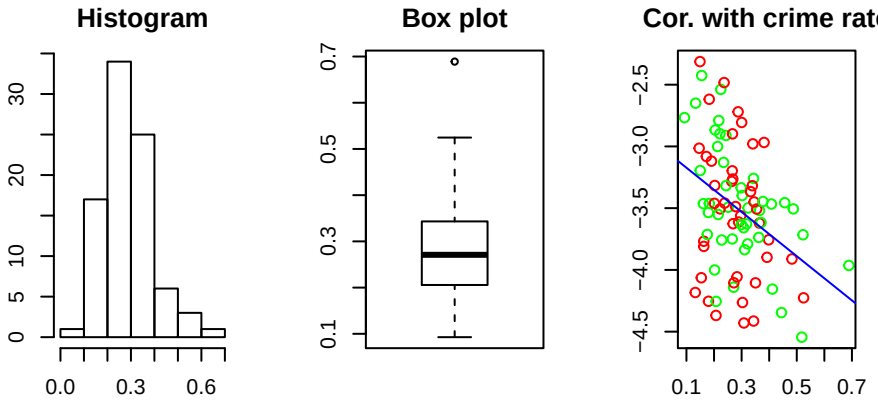
We see that high-crime areas are in central, urban North Carolina.

Probability of arrest

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

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

```
## NULL
```



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.

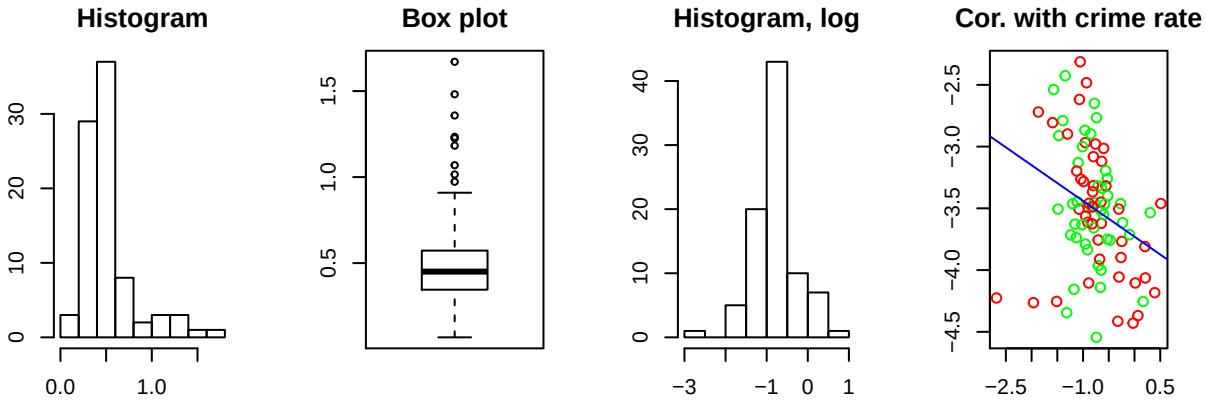
We will include *prbarr* in our model.

Probability of conviction

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

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

```
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*:

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

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

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

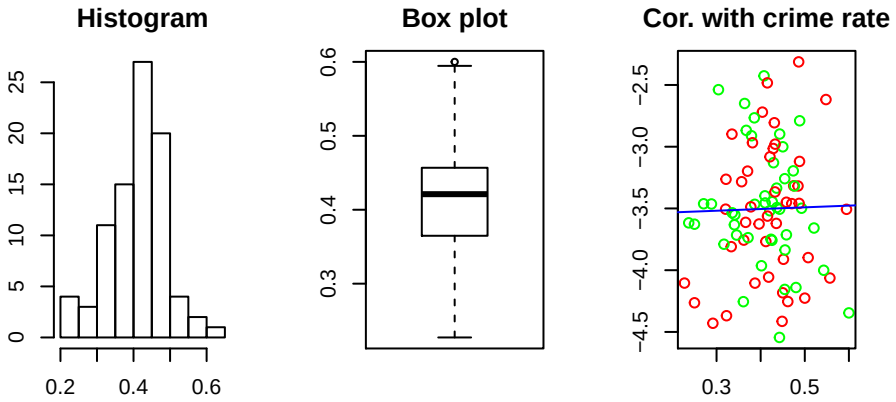
```
## [1] -0.2855235
```

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

Probability of prison sentence

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

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



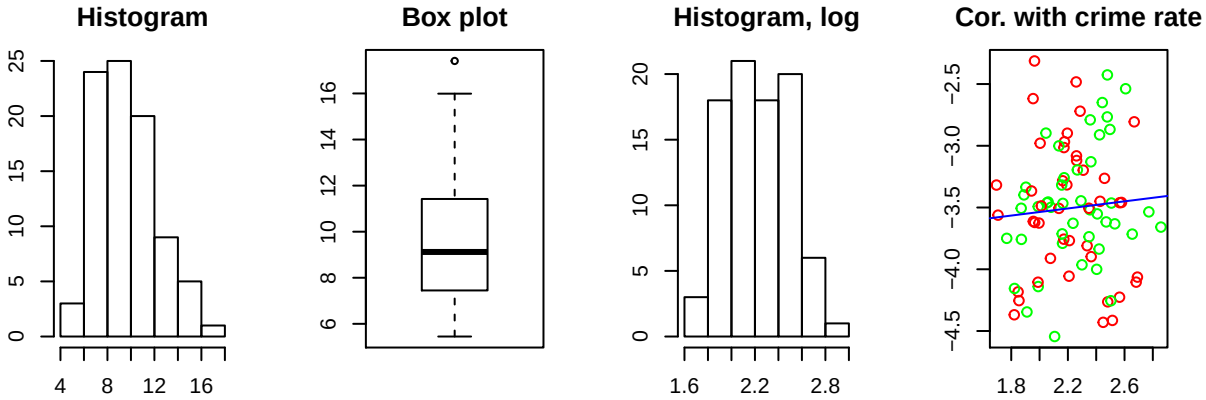
This histogram plot looks fairly normal. However, correlation is almost nonexistent wrt crime rate. This is interesting, because whether a crime results in spending time in prison does not seem to affect crime. This can shape government policies on whether to send criminals to prison or to find alternative 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.

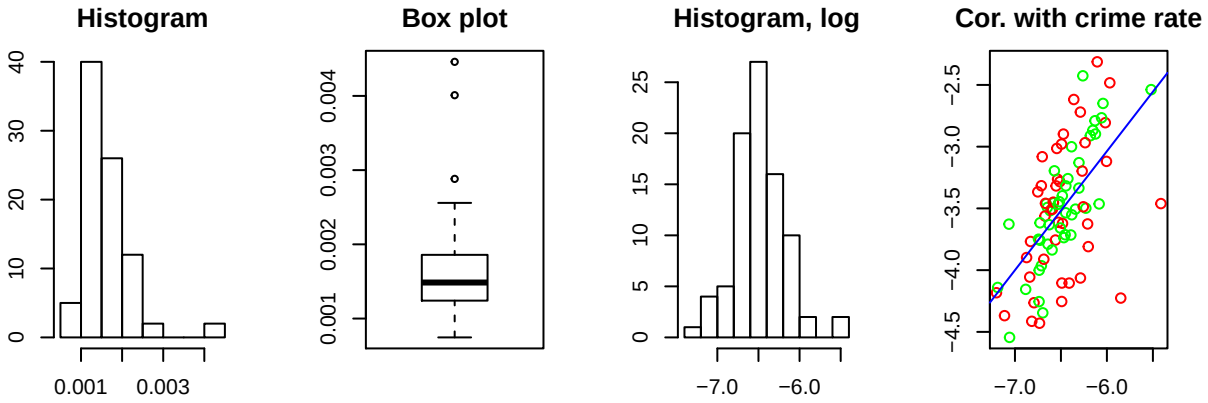
Police per capita has positive skew. Taking log helps:

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

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

```
## NULL
```

```
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. We will include it in the second model.

Population density

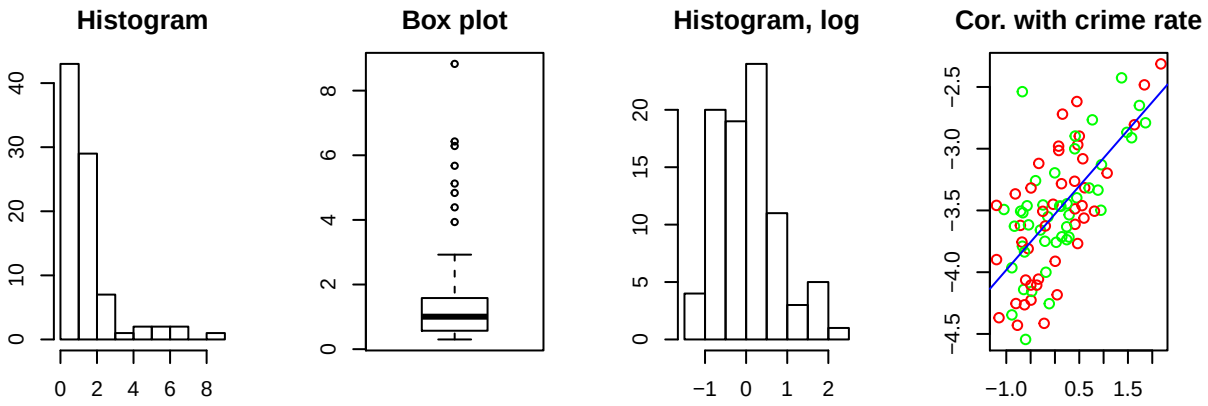
Note: In our first pass, we found an influential outlier with very low density and removed it, as mentioned in the section on outliers.

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

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

```
## NULL
```

```
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.

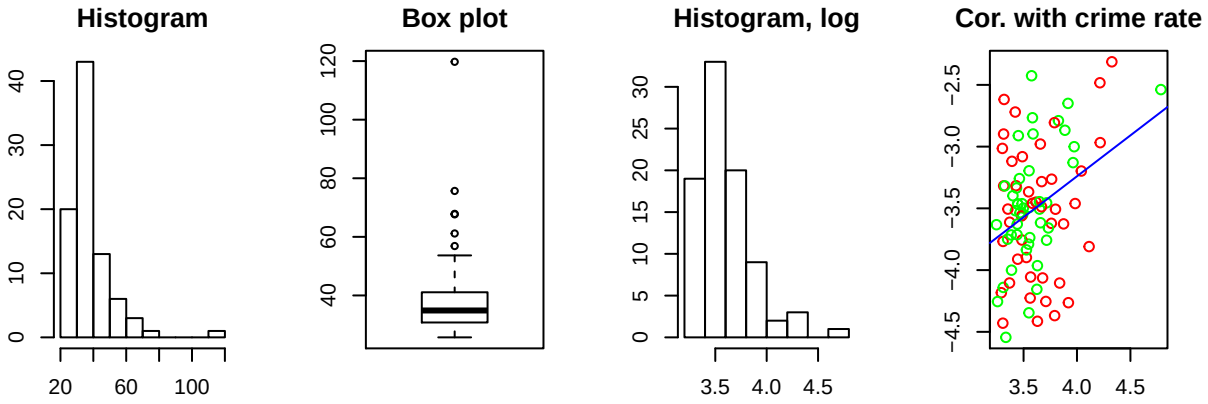
Tax revenue per capita

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

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

```
## NULL
```

```
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.

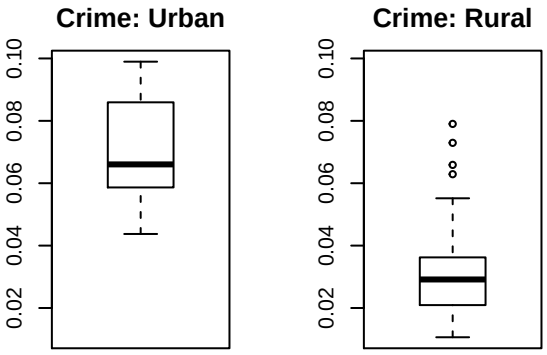
We will *not* include this variable in a first model.

Urban population

```
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(2,4), mai=c(0.35,0.35,0.35,0.35))
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. However, 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.

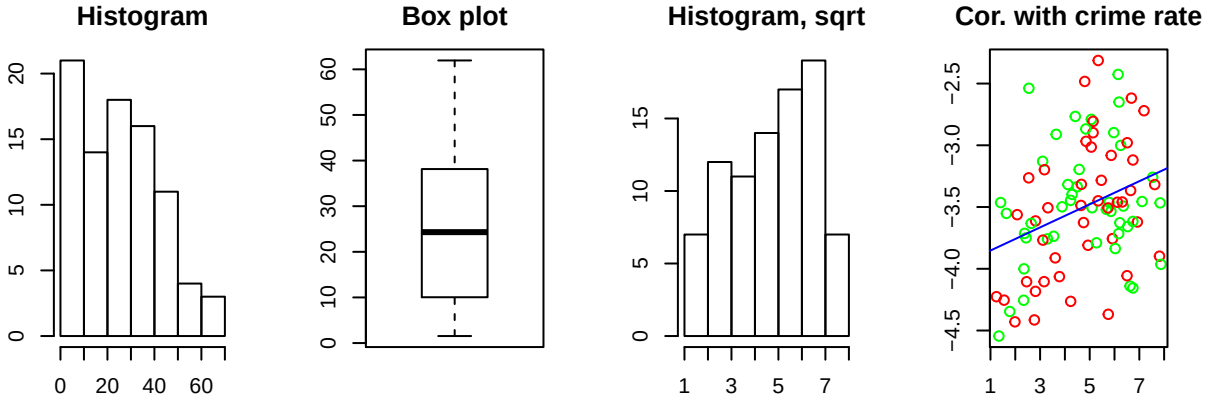
Percent minority

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

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

```
## NULL
```

```
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(3,4), mai=c(0.35,0.35,0.35,0.35))
hist(crime_data$wcon, title="wcon")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter

## Warning in axis(2, ...): "title" is not a graphical parameter
```

```
hist(crime_data$wloc, title="wloc")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter

## Warning in axis(2, ...): "title" is not a graphical parameter
```

```
hist(crime_data$wtrd, title="wtrd")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter

## Warning in axis(2, ...): "title" is not a graphical parameter
```

```
hist(crime_data$wtuc, title="wtuc")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter

## Warning in axis(2, ...): "title" is not a graphical parameter
```

```
hist(crime_data$wfir, title="wfir")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter

## Warning in axis(2, ...): "title" is not a graphical parameter
```

```
hist(crime_data$wser, title="wser")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter

## Warning in axis(2, ...): "title" is not a graphical parameter
```

```r
hist(crime_data$wmfg, title="wmfg")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter
```

```
## Warning in axis(1, ...): "title" is not a graphical parameter
```

```
## Warning in axis(2, ...): "title" is not a graphical parameter
```

```r
hist(crime_data$wfed, title="wfed")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter
```

```
## Warning in axis(1, ...): "title" is not a graphical parameter
```

```
## Warning in axis(2, ...): "title" is not a graphical parameter
```

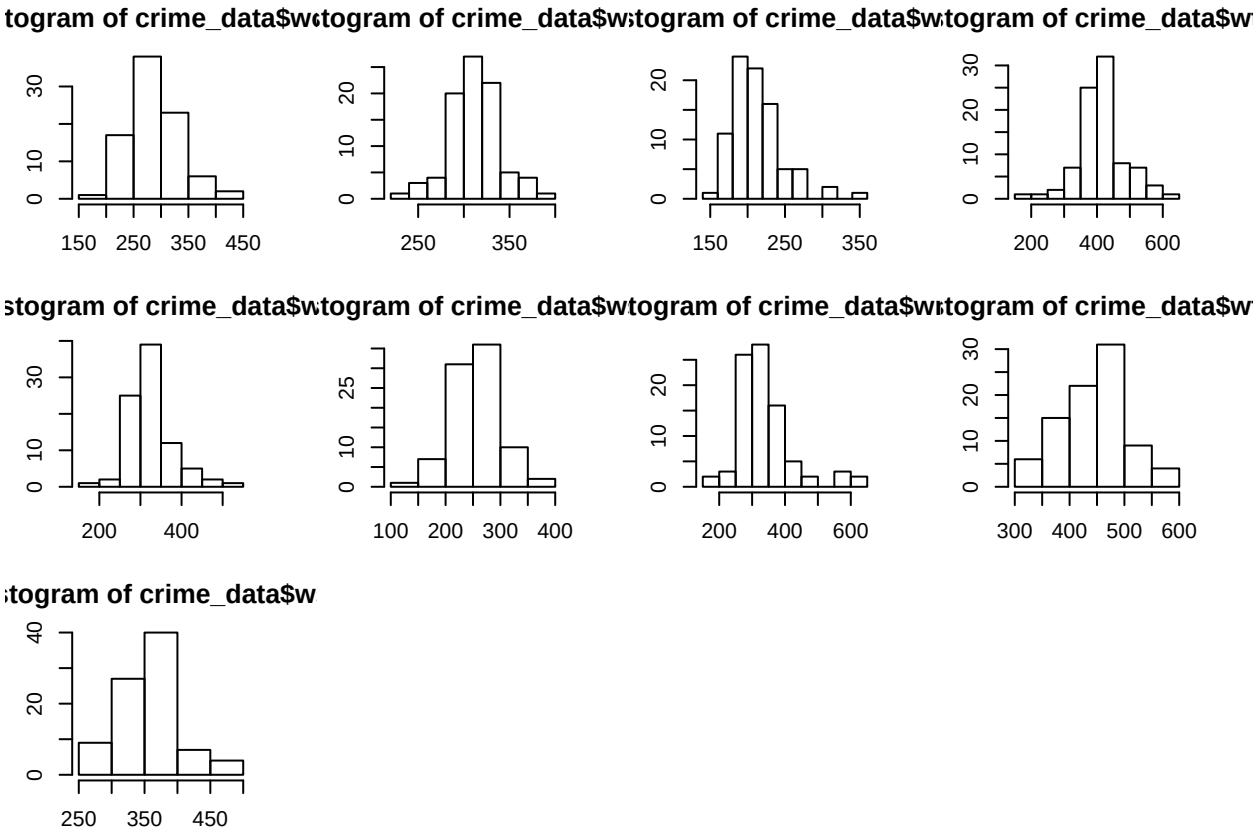
```r
hist(crime_data$wsta, title="wsta")
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter
```

```
## Warning in axis(1, ...): "title" is not a graphical parameter
```

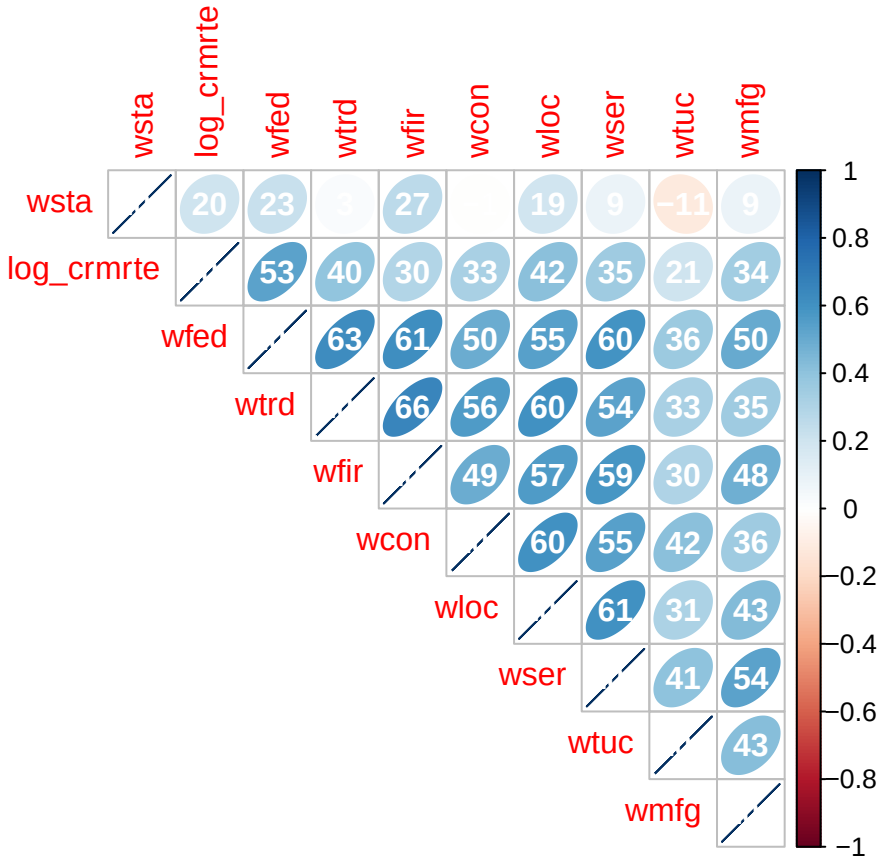
```
## Warning in axis(2, ...): "title" is not a graphical parameter
```



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("log_crmrte", "wcon", "wloc", "wtrd", "wtuc", "wfir",
              "wser", "wmfg", "wfed", "wsta")
corrplot(cor(crime_data[, wage_cols]), type="upper", diag=TRUE, addCoef.col="white", addCoefasPercent =
```



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.

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.

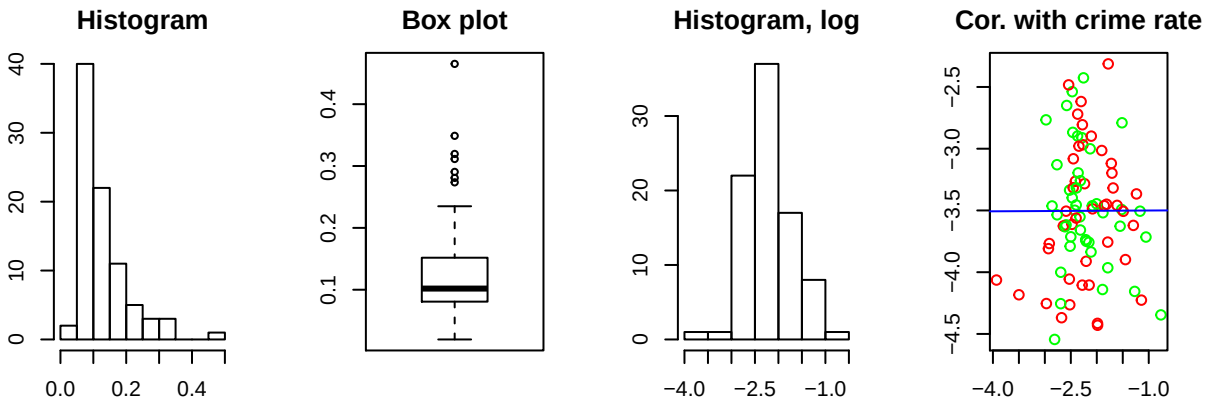
We will *not* include wages in a first model. We will use $wfed$ as a proxy variable for all wages, for a second model we will propose.

Offense Mix

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

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

```
## NULL
```



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.

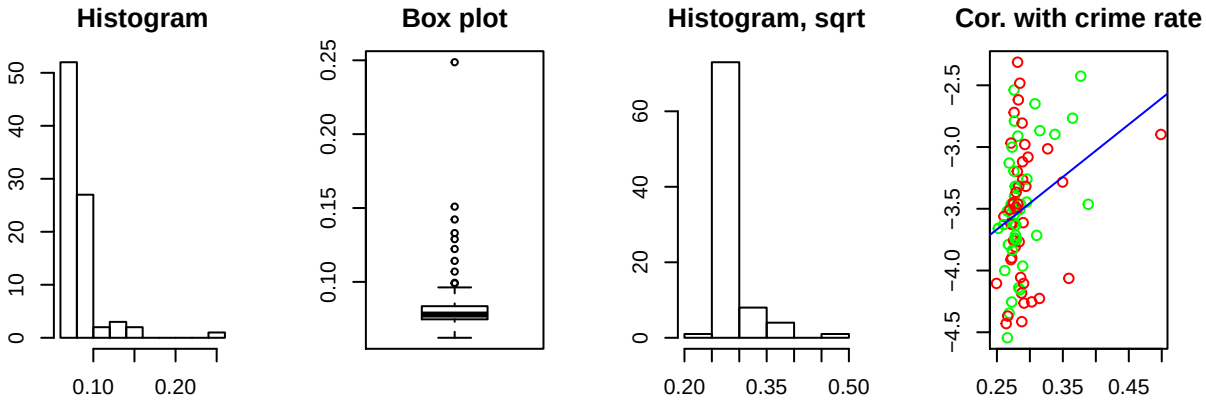
Percent of young males

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

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

```
## NULL
```

```
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 may also see the effect of omitted variables like youth unemployment or low education levels.

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

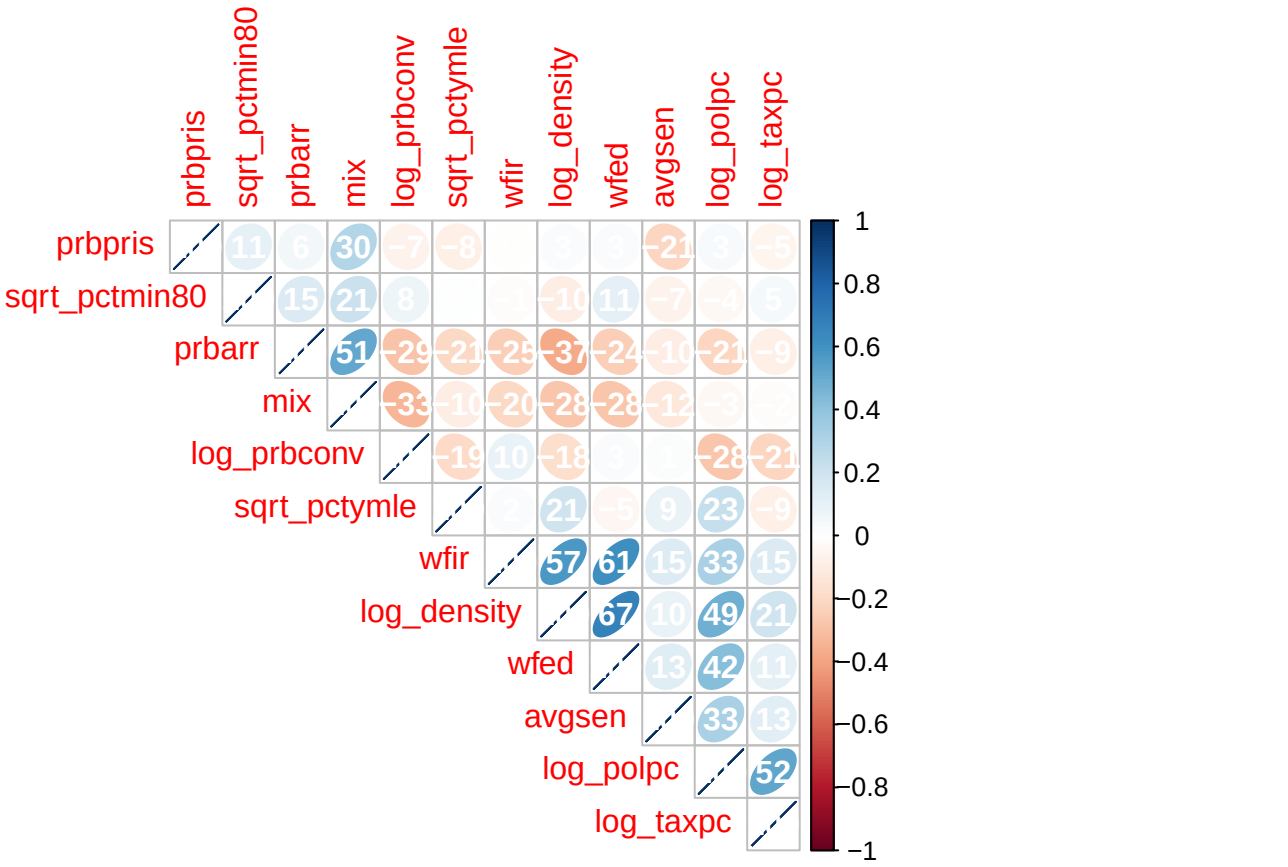
We will use these to come up with separate models, based on different factors, later in this analysis. [TODO: Saurav]

5. Correlation Analysis

Let us check for correlations across predictor pairs.

The correlation plot between the different predictors is as follows:

```
corrplot(
  cor(crime_data[,
                         c("prbarr", "log_prbconv", "prbpris", "avgsen",
                           "log_polpc", "log_density", "log_taxpc", "sqrt_pctmin80", "mix",
                           "sqrt_pctymle", "wfir", "wfed")
                         ]),
         type = "upper",
  diag=TRUE, addCoef.col="white", addCoefasPercent = TRUE, order="hclust", method="ellipse")
```



The following correlations are worth a remark:

- We do not see high correlations (>0.7), both positive or negative. This helps us avoid multicollinearity issues.
- Wages correlate highly (0.67) with population density, probably because it is tuned to cost of living.

- Police per capita correlates highly with density (0.56) and tax revenue (0.44), perhaps indicating federal laws on police counts.
- Probability of arrest correlates negatively (-0.38) with density, indicating that criminals get away with crime in populous areas. It also correlates negatively (-0.30) with probability of conviction, indicating higher chance of false arrests when a large number of arrests are made.

6. Model development

Summary of variables

Here is a summary table of variables we will use in our models.

Variable	Transform?	Model1?	Model2?	Model3?	Remarks
county	N/A				Unused
year	N/A				Unused
prbarr		Y	Y	Y	Causal
prbconv	log	Y	Y	Y	Causal
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					Omit var: wealth
wtuc					Proxied
wtrd					-“-
wfir				Y	
wser					
wmfg					
wfed			Y	Y	Proxy
wsta					Proxied
wloc					Proxied
mix	log			Y	No corr. found
pctymle	sqrt	Y	Y	Y	Causal, weak cor

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

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

- We believe that the following can directly cause higher crime: high density, higher percentage of minorities, higher percentage of young men.
- We also think the following cause lower crime: high probability of arrest and conviction.

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

Let us fit the model:

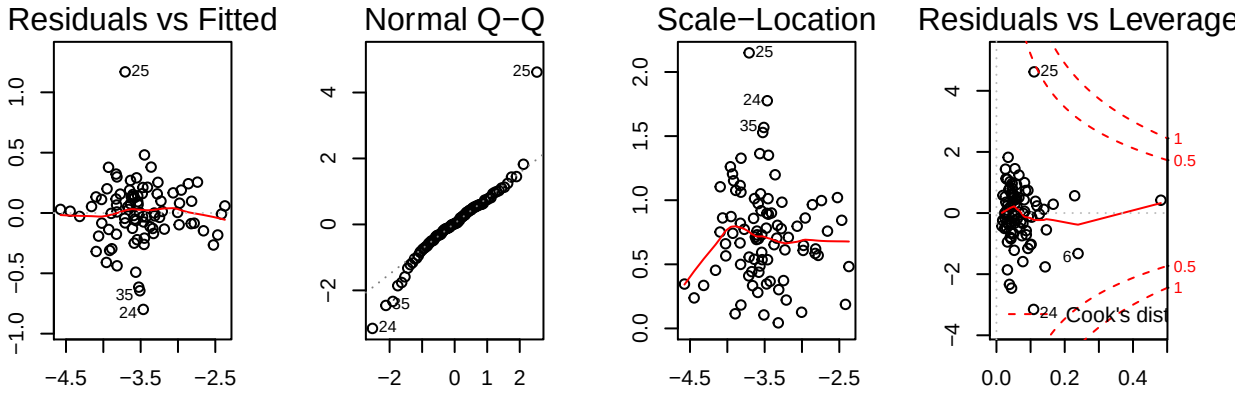
```
model1 = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
              sqrt_pctmin80 + sqrt_pctymle, data=crime_data)
summary(model1)
```

```
##
## 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
```

```
AIC(model1)
```

```
## [1] 25.8433
```

```
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
plot(model1)
```



The model shows a fit of about 0.79 as measured by adjusted R^2. There are a few (non-normal) outliers at the bottom left of the QQ-plot due to data skew.

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. In this model we also include three variables that show positive correlation with crime rate, albeit not causal. We think they are all the result of wealthy, urban demographics: high police per capita, high tax revenue and high federal wages. It is an omitted variable.

We do not include outcome variables that absorb causal effect (by having negative correlation).

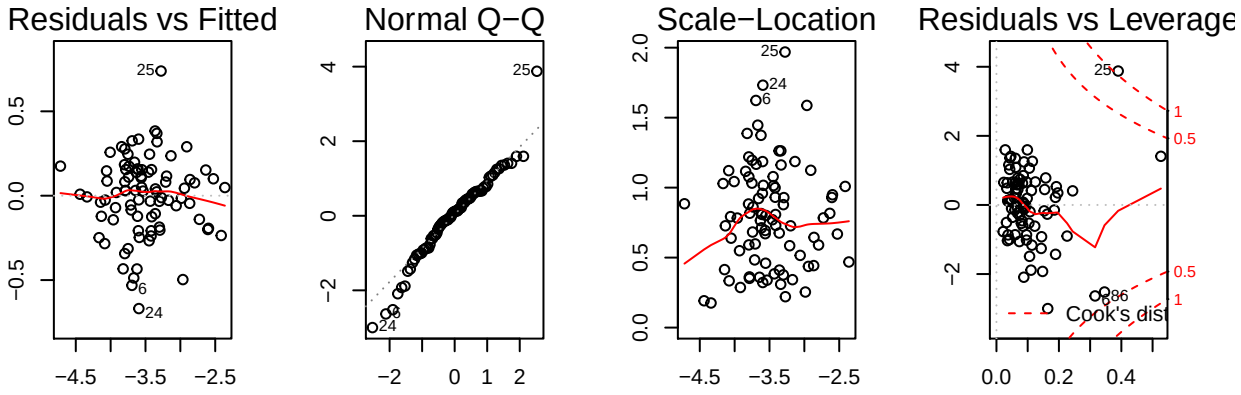
```
model2 = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
              sqrt_pctmin80 + sqrt_pctymle
            + log_polpc + log_taxpc + wfed,
            data=crime_data)
summary(model2)
```

```
##
## Call:
## lm(formula = log_crmrte ~ prbarr + log_prbconv + log_density +
##     sqrt_pctmin80 + sqrt_pctymle + log_polpc + log_taxpc + wfed,
##     data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66900 -0.13257  0.02751  0.15293  0.73879
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.9505273  1.3384221  -1.457  0.14904
## prbarr        -1.4205883  0.3039228  -4.674 1.21e-05 ***
## log_prbconv   -0.2586847  0.0603323  -4.288 5.12e-05 ***
## log_density    0.2640959  0.0554056   4.767 8.53e-06 ***
## sqrt_pctmin80  0.1236974  0.0156536   7.902 1.46e-11 ***
## sqrt_pctymle   0.5074692  0.9181660   0.553  0.58205
## log_polpc      0.3883345  0.1196693   3.245  0.00173 **
## log_taxpc      0.0640858  0.1245985   0.514  0.60847
## wfed           0.0004268  0.0006851   0.623  0.53513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.244 on 78 degrees of freedom
## Multiple R-squared:  0.7931, Adjusted R-squared:  0.7718
## F-statistic: 37.36 on 8 and 78 DF,  p-value: < 2.2e-16
```

```
AIC(model2)
```

```
## [1] 11.95387
```

```
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
plot(model2)
```



The fit improves to 0.808 (adjusted R^2). Although we added it, the wage $wfed$ is not significant. *taxpc* is

not significant either.

We will now build a third model that includes almost all the variables, for the sake of completeness and comparison. We only exclude wages because their distribution is highly alike.

```
model3 = lm(log_crmrte ~ prbarr + log_prbconv + log_density +
              sqrt_pctmin80 + sqrt_pctymle
            + log_polpc + log_taxpc
            + prbpris + avgsen + wfir + wfed + mix,
            data=crime_data)
summary(model3)
```

```
##
## Call:
## lm(formula = log_crmrte ~ prbarr + log_prbconv + log_density +
##     sqrt_pctmin80 + sqrt_pctymle + log_polpc + log_taxpc + prbpris +
##     avgsen + wfir + wfed + mix, data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60074 -0.12388  0.03179  0.13204  0.73809
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.7421830  1.4123267  -0.526 0.600804
## prbarr        -1.3854283  0.3120115  -4.440 3.09e-05 ***
## log_prbconv   -0.2280003  0.0619757  -3.679 0.000442 ***
## log_density    0.2909365  0.0559732   5.198 1.72e-06 ***
## sqrt_pctmin80  0.1223591  0.0157687   7.760 3.69e-11 ***
## sqrt_pctymle   0.4555590  0.9047906   0.503 0.616111
## log_polpc      0.4722139  0.1238017   3.814 0.000281 ***
## log_taxpc      0.0584928  0.1232063   0.475 0.636360
## prbpris       -0.4434790  0.3645861  -1.216 0.227702
## avgsen        -0.0206691  0.0109452  -1.888 0.062889 .
## wfir          -0.0013613  0.0006380  -2.134 0.036180 *
## wfed           0.0009056  0.0007211   1.256 0.213148
## mix           -0.0156420  0.4495366  -0.035 0.972336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2358 on 74 degrees of freedom
## Multiple R-squared:  0.8167, Adjusted R-squared:  0.787
## F-statistic: 27.48 on 12 and 74 DF,  p-value: < 2.2e-16
```

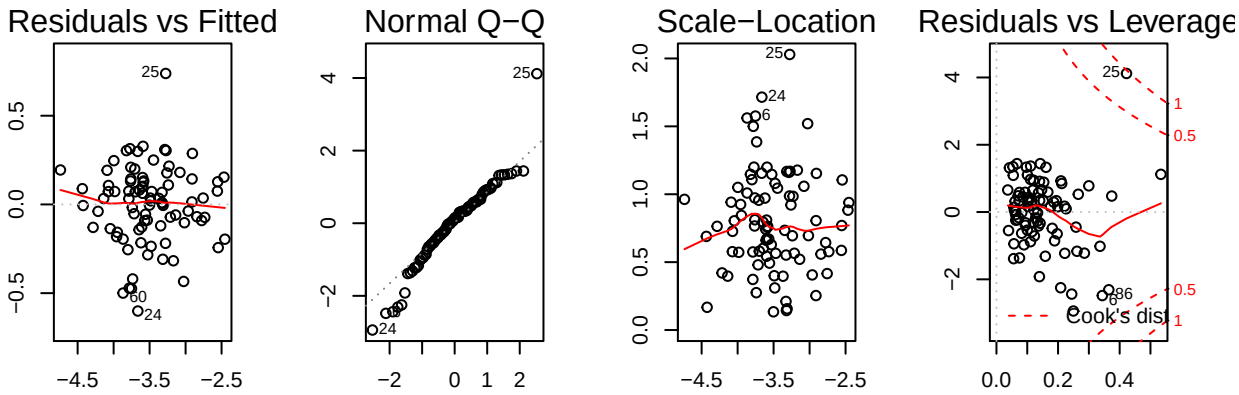
```
vif(model3)
```

```
##        prbarr   log_prbconv   log_density sqrt_pctmin80  sqrt_pctymle
##      1.713807      1.657098      2.805479      1.238806      1.445825
##     log_polpc     log_taxpc       prbpris        avgsen          wfir
##      2.420267      1.684817      1.200621      1.219175      1.850701
##          wfed           mix
##      2.845098      1.828567
```

```
AIC(model3)
```

```
## [1] 9.397448
```

```
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
plot(model3)
```



This tops fit at about 0.828 (adjusted R^2). It also says *avgsen* is significant.

The AIC scores also reflect the fit: the last model has the best (least) score, whereas the first model has the worst.

% 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 - 15:10:08

7. Omitted variables

We have talked about omitted variables as part of our EDA. Here we continue the discussion and also talk about the direction of bias.

We saw that crime rate was correlated positively with wages, tax revenue and police per capita. None of these can cause crime. We instead suspected urban petty crimes to be the cause of those correlations. Similarly, we can think of several omitted variables that can help us develop causal models:

- Education: What is the average number of years of education? Uneducated people are more likely to get into crime for fast money. Higher amounts of education should correlate negatively with crime rate to a large extent.
- Unemployment: What fraction of the population is unemployed? Lack of income can drive people to crime. Lower amounts of unemployment should correlate negatively with crime rate to a large extent.
- Poverty: What is the average family income? If income is low for a family as a unit, that may lead to crime. Higher poverty correlates with higher crime rate to a large extent.
- Family size: What is the average family size? If the family size is large, then the income may not be sufficient. Large family sizes may correlate with crime rate to a small degree.
- Alcoholism and substance abuse: Patient data from hospitals and police data can be a good marker to judge this. High number of cases of drug abuse correlates with higher crime rate to a large extent.
- Demographics: Too many single men? Single parenting? Orphaned children? Segregated neighborhoods? These can lead to trauma and crime. They may correlate with crime rate to a small degree.

8. Conclusion

Based on the above study, here are the conclusions we would like to offer the political campaign:

Table 2: Regression model summary

	Dependent variable:		
	log_crmrte		
	(1)	(2)	(3)
prbarr	-1.586^{***}	-1.421^{***}	-1.385^{***}
	(0.330)	(0.304)	(0.312)
log_prbconv	-0.316^{***}	-0.259^{***}	-0.228^{***}
	(0.063)	(0.060)	(0.062)
log_density	0.356^{***}	0.264^{***}	0.291^{***}
	(0.043)	(0.055)	(0.056)
sqrt_pctmin80	0.130^{***}	0.124^{***}	0.122^{***}
	(0.016)	(0.016)	(0.016)
sqrt_pctymle	0.576	0.507	0.456
	(0.914)	(0.918)	(0.905)
log_polpc		0.388^{***}	0.472^{***}
		(0.120)	(0.124)
log_taxpc		0.064	0.058
		(0.125)	(0.123)
prbpris			-0.443
			(0.365)
avgsen			-0.021^{*}
			(0.011)
wfir			-0.001^{**}
			(0.001)
wfed		0.0004	0.001
		(0.001)	(0.001)
mix			-0.016
			(0.450)
Constant	-4.100^{***}	-1.951	-0.742
	(0.291)	(1.338)	(1.412)
Observations	87	87	87
R^2	0.740	0.793	0.817
Adjusted R^2	0.724	0.772	0.787
Residual Std. Error	0.268 (df = 81)	0.244 (df = 78)	0.236 (df = 74)
F Statistic	46.084^{***} (df = 5; 81)	37.364^{***} (df = 8; 78)	27.476^{***} (df = 12; 74)
Note:			*p<0.1; **p<0.05; ***p<0.01

- Crime rate is most correlated (0.7) with population density. Median crime in urban areas is more than double those in rural areas. Our policies should make our cities safer, in order to reduce crime significantly.
- As probability of arrest and/or conviction increase, crime decreases. Law enforcement is a good deterrent to crime.
- Prison by itself does not correlate with crime, nor does the length of prison sentence. We should use this data to argue for shorter sentences and alternatives to prison, such as reform and counselling. We should be careful to continue to provide strong disincentives to crime, however.
- Police per capita correlates positively with crime: this may mean we have not improved the effectiveness of our police in crime-infested areas. This may also be due to omitted variables and is worth exploring further.
- Similarly, wealth as proxied by tax revenue begets crime. This may also suggest that we have the money and should be able to reroute tax dollars better to fight crime in high-crime areas.
- Minorities have moderate correlation with crime. Our policies should address integration of minorities into the mainstream and reduce segregation.
- Similarly, we should investigate correlation of crime with young men. If these men are driven to crime due to lack of education or unemployment (omitted variables in this dataset), we should pay attention to reforming our education or job market.