Lab3 Final, w203: Statistics for Data Science

Avinash Chandrasekaran, Deepak Nagaraj, Saurav Datta

April 13, 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 97 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 causal estimates and we will clearly explain how omitted variables 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

Our first step is to detect anomalies such as missing and duplicate values and to clean up the dataset before deeper dive into regression analysis.

```
# 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 noticed 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)
```

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

- *prbconv*, observation #6: extremely low conviction rate
- *prbarr*, observation #51: ratio of arrests to offenses is 1.1. It doesn't seem likely that a person will be arrested more than once for the same offense (unless of course we mix in criminals escaping). And this outlier single handedly affects regression line on crime rate as well
- *density*, observation #79: extremely low density. Such small number of people per square mile likely is erroneous
- *wser*, observation #84: extremely high service wages. An order of magnitude larger that we deemed likely to be due to data entry error
- *polpc*, observation #89: large outlier causes skew

```
par(mfrow=c(2,4), mai=c(0.35,0.35,0.35,0.35))
#layout(matrix(c(1,2,3,4), 1, 4), heights = lcm(6))
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)
m = lm(log(crime_data$crmrte) ~ log(crime_data$prbconv))
plot(m, which=5, caption=NA)
```
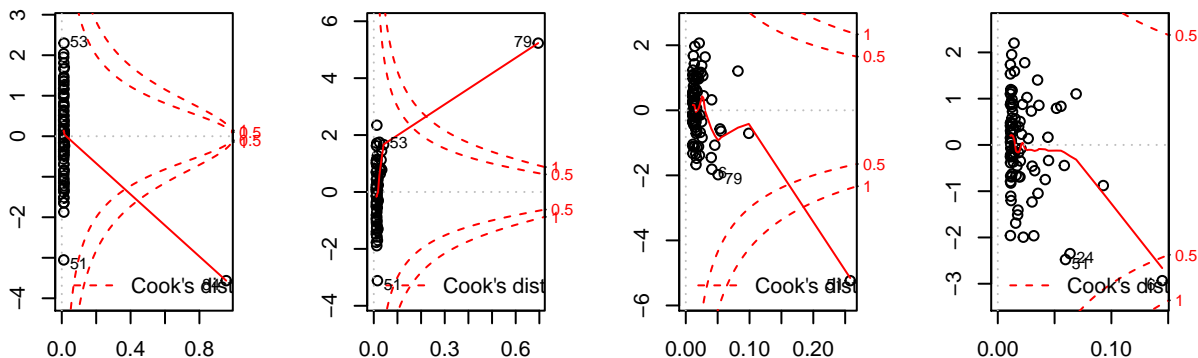

Finally, observation #25 causes outliers in the final model fit due to a lot of influence. It shows very high crime rate, very high police per capita, very low density, highest tax per capita, very low minority. It is an

observation with too many quirks and may merit separate investigation.

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

```
crime_data <- crime_data %>% slice(-c(6, 25, 51, 79, 84, 89))
```

3. 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) {
    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 do not see any top-coding or bottom-coding of variables. All the variables have their values as-is.

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)
```
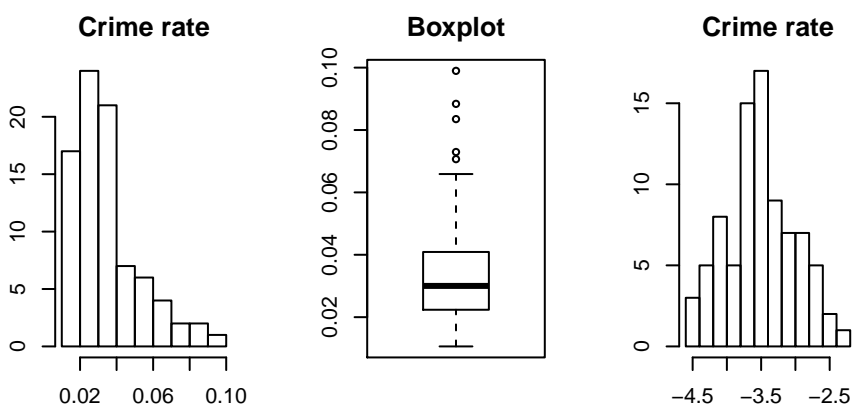

Looking at the histogram, the distribution is positively skewed to the left. We can take the log transformation which makes the variable appear more normally distributed.

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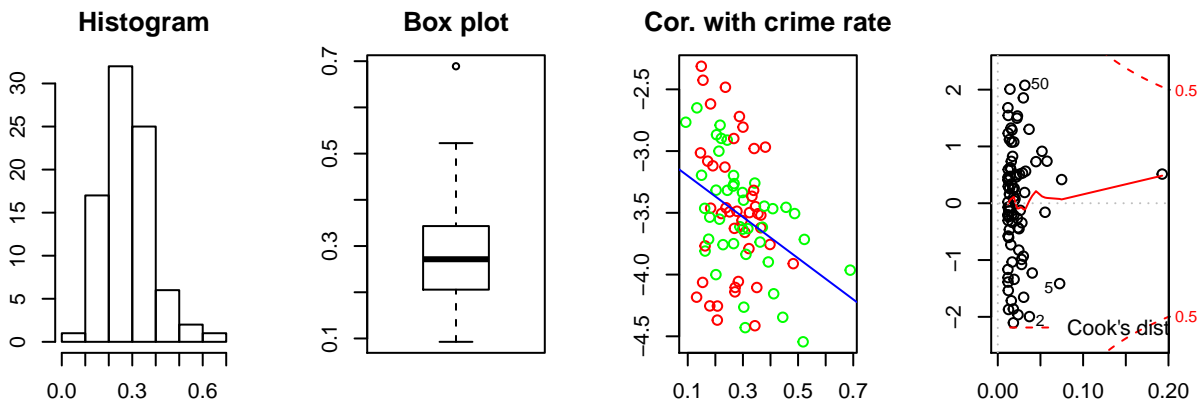
```
## 4    6.29       0     1  0.237   0.393
## 5    1.57       1     0  0.183   0.343
```

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:  -0.347"
```


The plot looks fairly normal; there is only one outlier that corresponds to the high arrest probability of 0.7. It is not influential, so we will keep it.
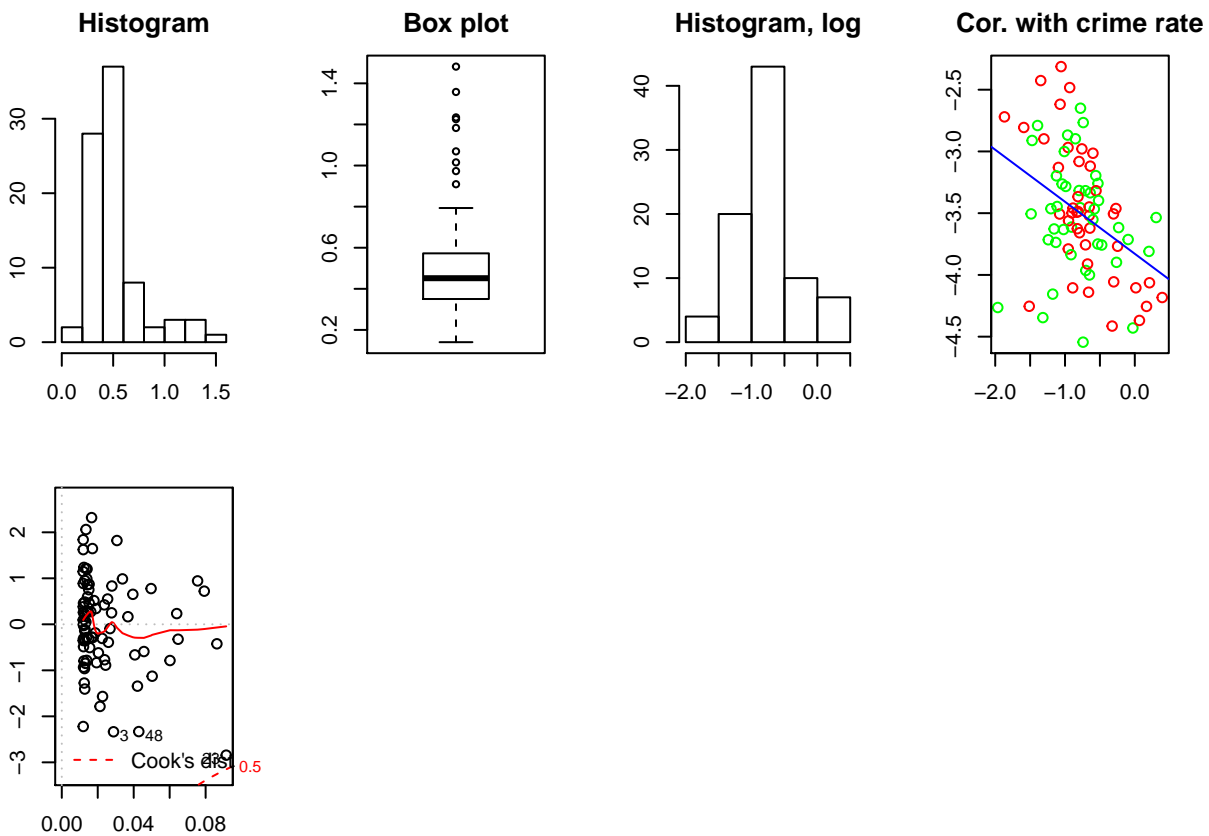
There is fairly negative correlation of -0.37: as probability of arrests increases, crime rate goes down. It may be that arrests are a deterrent, indicating causality.

We will include *prbarr* in our model.

Probability of conviction

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

```
## [1] "Correlation:  -0.391"
```

```
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] -0.2668945
```

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

```
## [1] -0.2133534
```

We don't see much correlation and so, will include *log_prbconv* in our model.
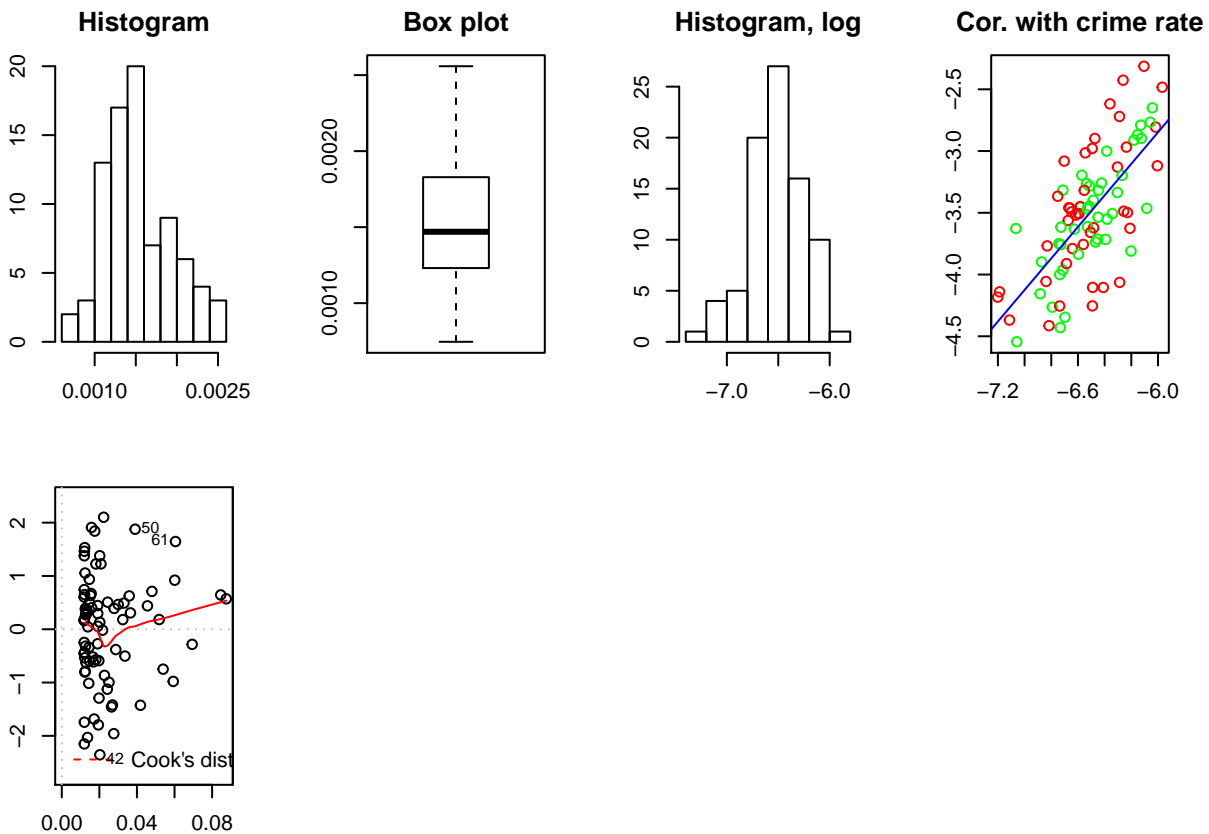
Police per capita

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. We observed that taking the log tranformation made the distribution more normal.

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

```
## [1] "Correlation:  0.695"
```

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

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

In our first pass, we found an influential outlier where observation #79 has Cook's distance beyond 1, meaning extreme leverage. The value of 0.000020342 seems like an unlikely value for people per square mile. We thus removed it from consideration.

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

[1] "Correlation: 0.715"

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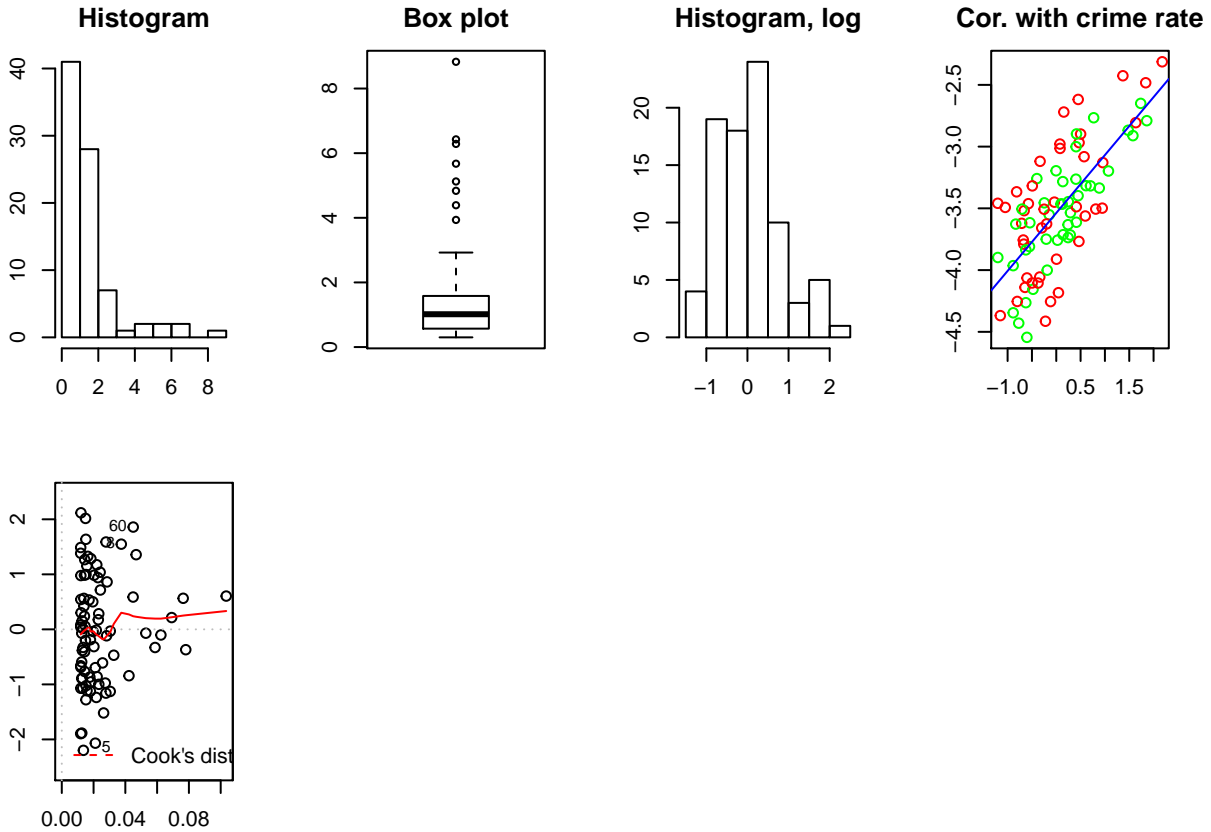

The histogram of density shows quite a bit positive skew. The log transformation shows a more promising normal distribution. There are no outliers with large leverage as measured by Cook's distance.
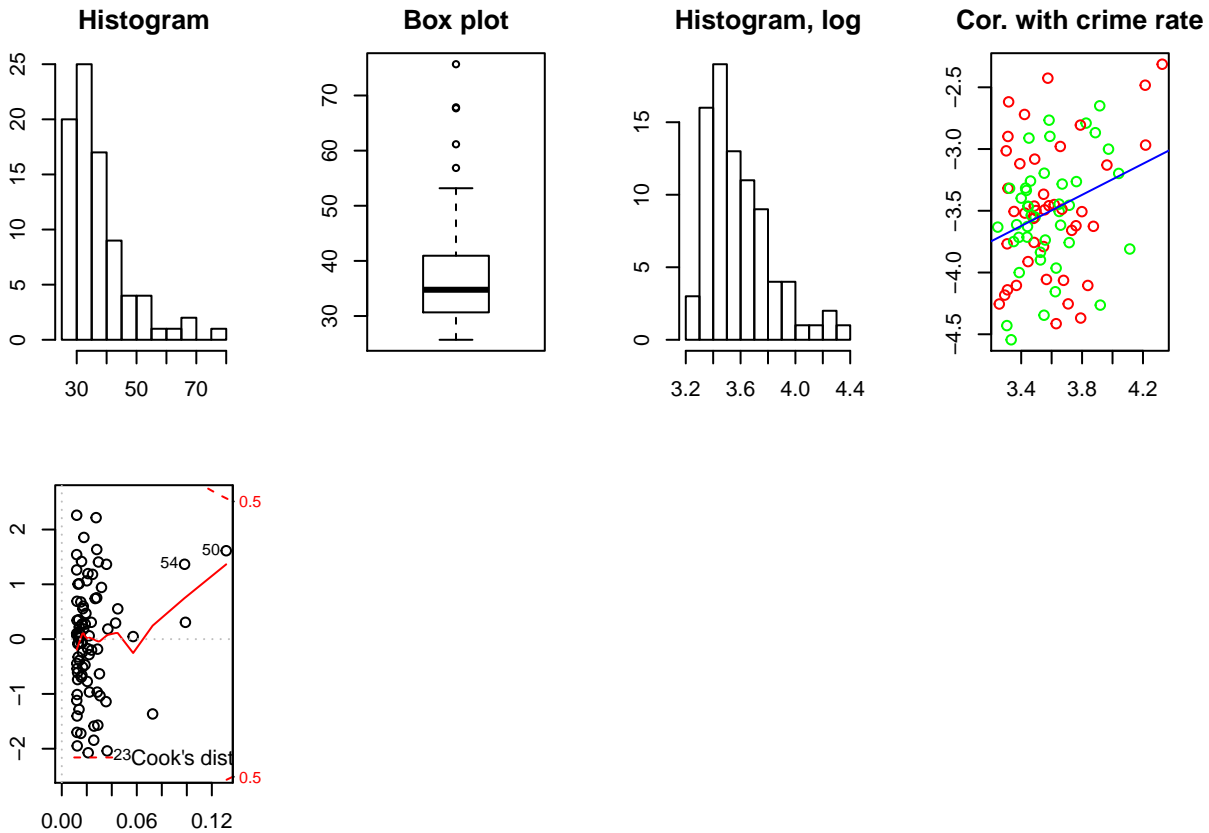
We see high positive correlation with crime rate. It may be that high population density indicates greater scope for hiding or cooperation in order to commit crime, indicating causality. We will surely consider this variable in our model.

Tax revenue per capita

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

[1] "Correlation: 0.295"

```
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.

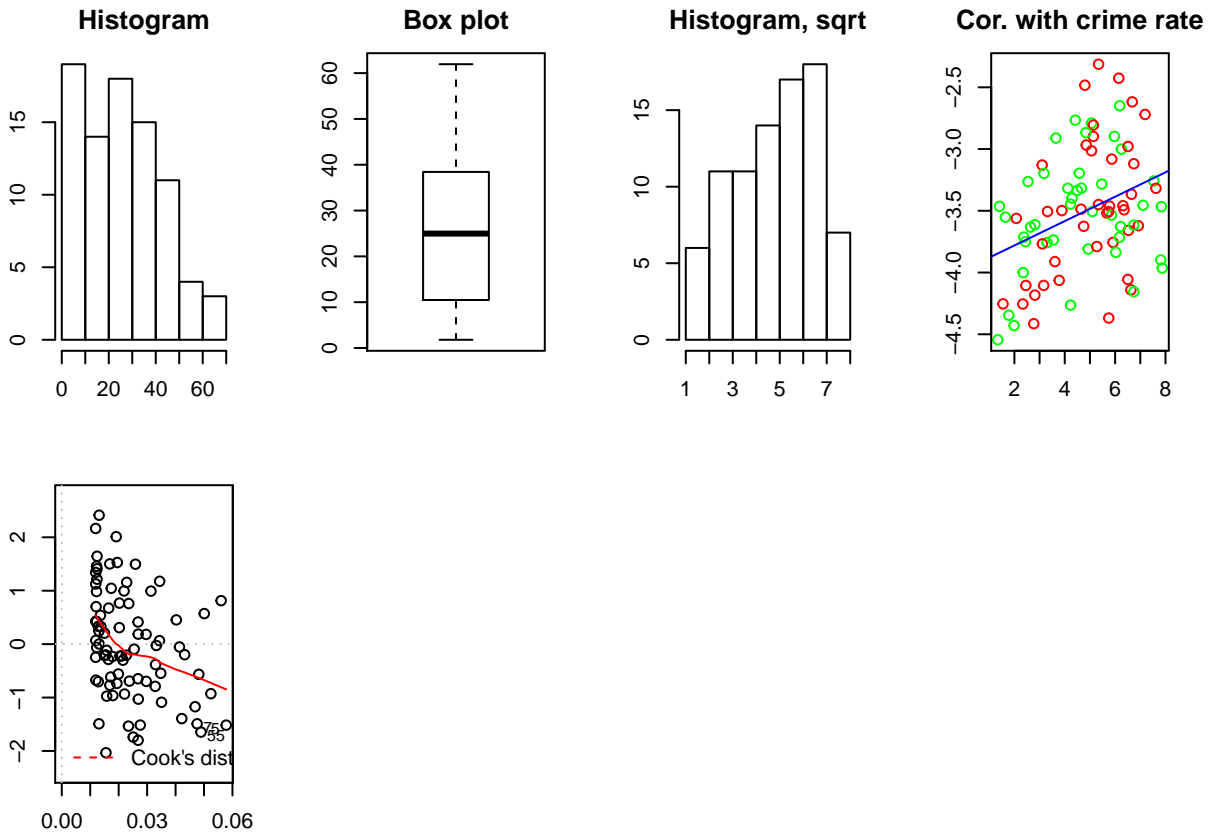We will *not* include this variable in a first model.

Percent minority

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

```
## [1] "Correlation:  0.346"
```

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

Histogram	Box plot	Histogram, sqrt	Cor. with crime rate

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

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.

```r
par(mfrow=c(3,4), mai=c(0.35,0.35,0.35,0.35))
hist(crime_data$wcon, main="wcon")
hist(crime_data$wloc, main="wloc")
hist(crime_data$wtrd, main="wtrd")
hist(crime_data$wtuc, main="wtuc")
hist(crime_data$wfir, main="wfir")
hist(crime_data$wser, main="wser")
hist(crime_data$wmfg, main="wmfg")
hist(crime_data$wfed, main="wfed")
hist(crime_data$wsta, main="wsta")
```
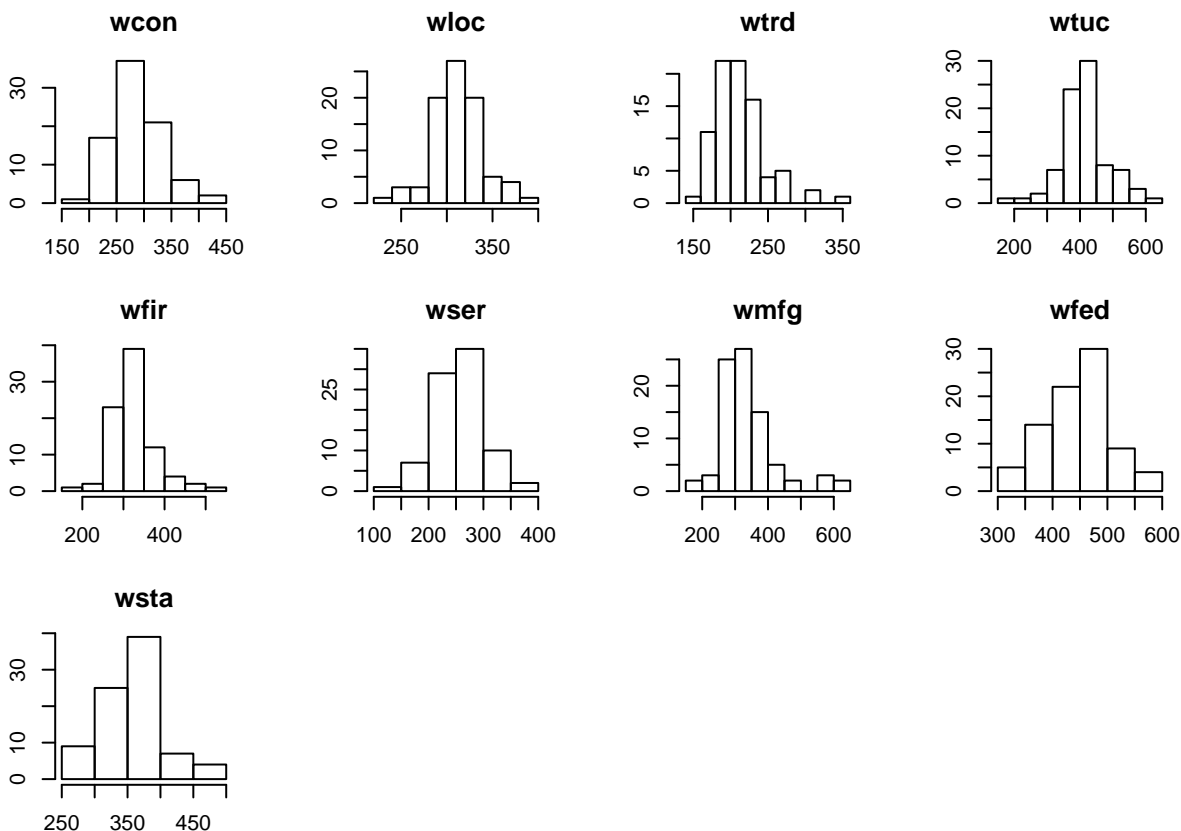
Most of the wage variables conform to normal distributions. We do not have to worry about transformations.

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

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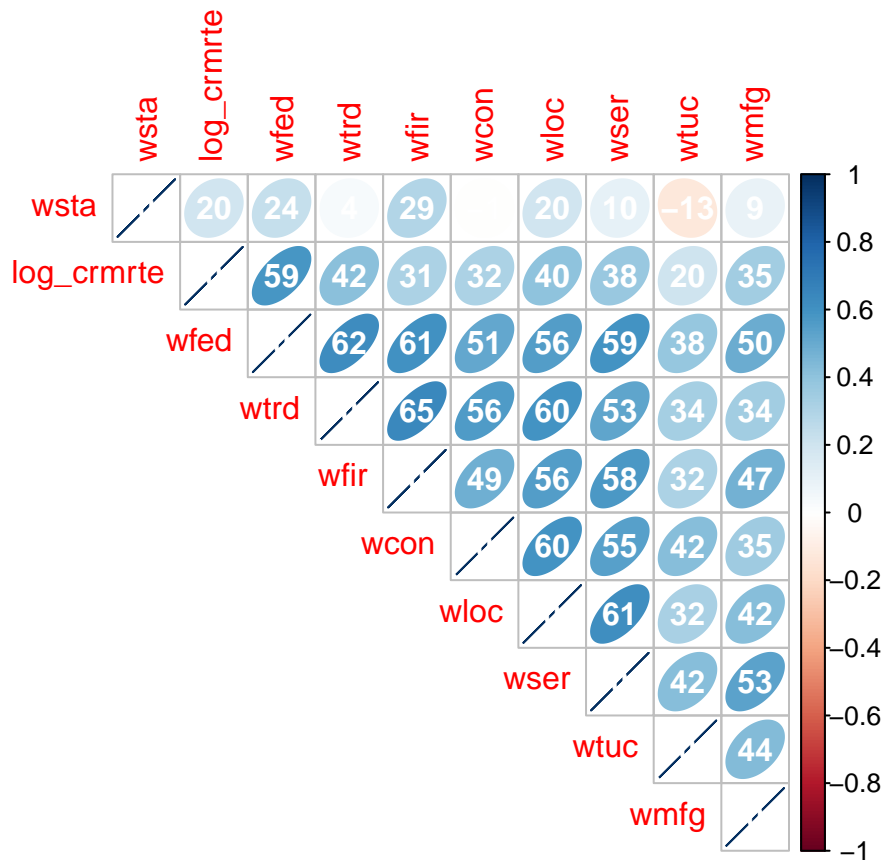
Indeed, a lot of the wage categories above have a high degree of correlation among them, but all are less than 70. We cannot eliminate any wage categories this way.

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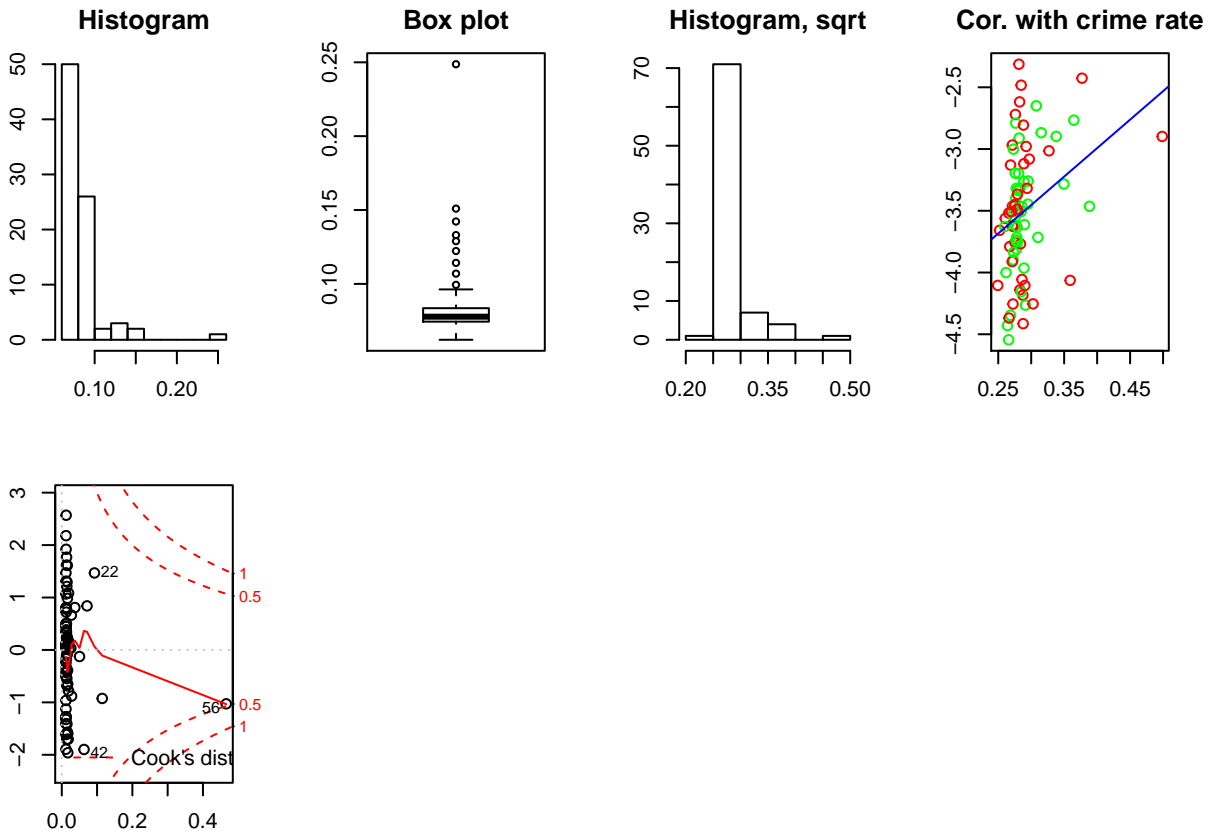
We will *not* include wages in a first model. We will use $wfed$ as a proxy variable for all wages, for a second model we will propose.

Percent of young males

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

```
## [1] "Correlation:  0.314"
```

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

Histogram   **Box plot**   **Histogram, sqrt**   **Cor. with crime rate**

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 > 1).

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.

Urban population

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

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

```r
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)
```
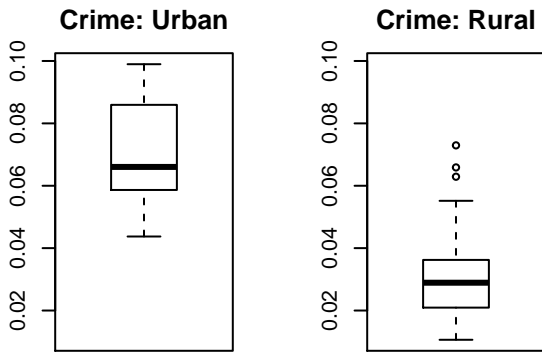
It is worth noting that there are only 8 observations classified urban in this dataset. However, median crime rate in urban regions is double that of rural regions.

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

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

```
## [1] 0.8231259
```

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.

Other variables ignored

The following variables show low to nonexistent correlation with crime rate. We will *not* consider them in our regression model.

- Probability of prison sentence ($R = 0.0537$)
- Offense mix ($R = 0.0136$)
- Average sentence duration ($R = 0.0438$)

In addition, we will use geographic regions to come up with separate models, later in this analysis.

4. 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")
```
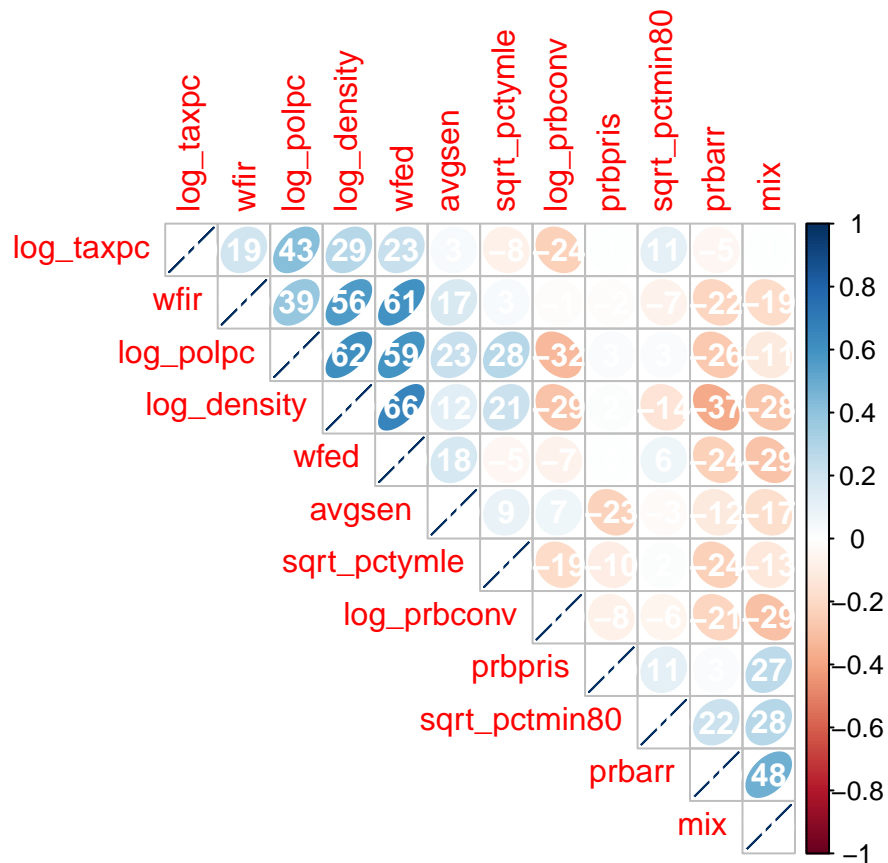

The following correlations are worth a remark:

- 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.

5. 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.
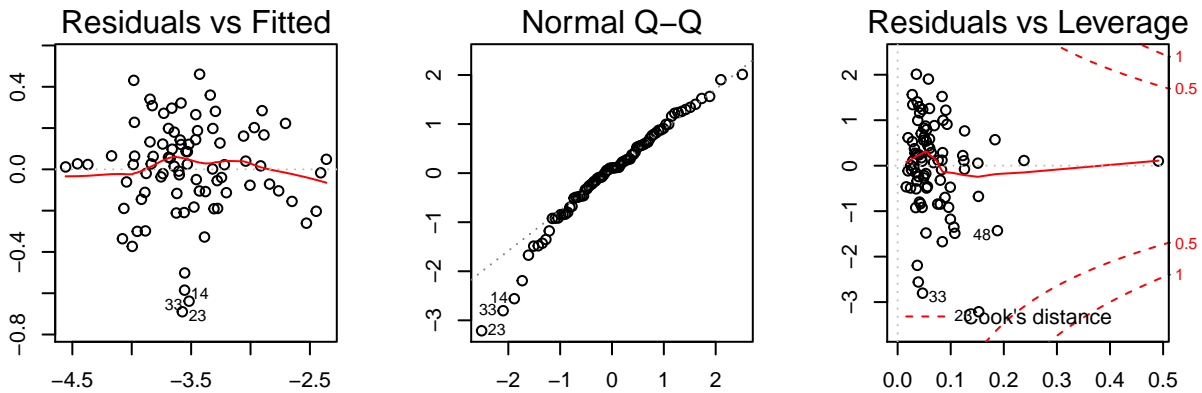
$$\log\left(crmrte\right) = \beta_0 + \beta_1 \ prbarr + \beta_2 \ \log\left(prbconv\right) + \beta_3 \ \log\left(density\right) + \beta_4 \ \sqrt{pctmin80} + \beta_6 \ \sqrt{pctymle}$$

Let us fit the model:

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

```
## [1] 0.7848028
```

```
par(mfrow=c(2,3), mai=c(0.35,0.35,0.35,0.35))
plot(model1, which=c(1,2,5))
```

The model shows a fit of about 0.785 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 that we could not correct via transformations. This is for *density*, *pctymle* variables, which were quite skewed to begin with.

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)$adj.r.squared
```

```
## [1] 0.8237172
```

```
par(mfrow=c(2,3), mai=c(0.35,0.35,0.35,0.35))
plot(model2, which=c(1,2,5))
```
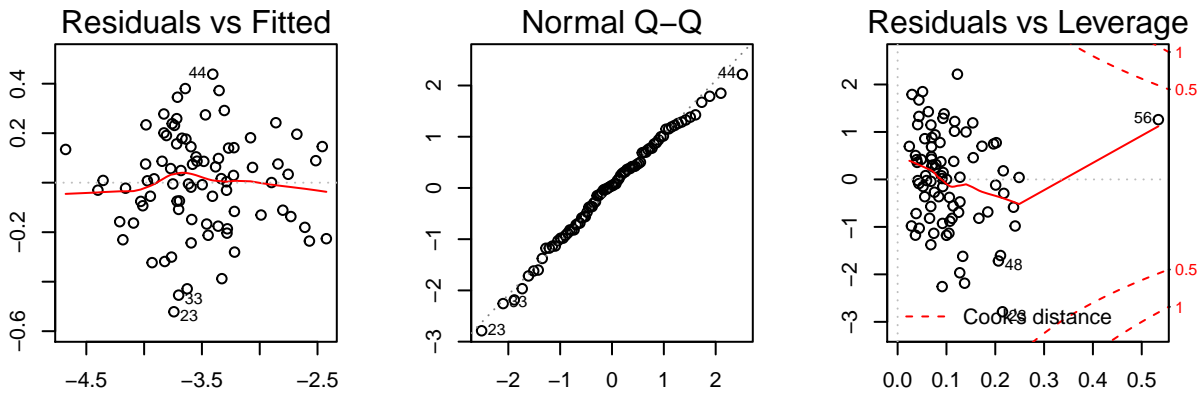
The fit improves to 0.824 (adjusted R^2). This is because some of the newly added variables are more normal in distribution (=less skew).

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(model2)$adj.r.squared
```

```
## [1] 0.8237172
```

```
par(mfrow=c(2,3), mai=c(0.35,0.35,0.35,0.35))
plot(model3, which=c(1,2,5))
```
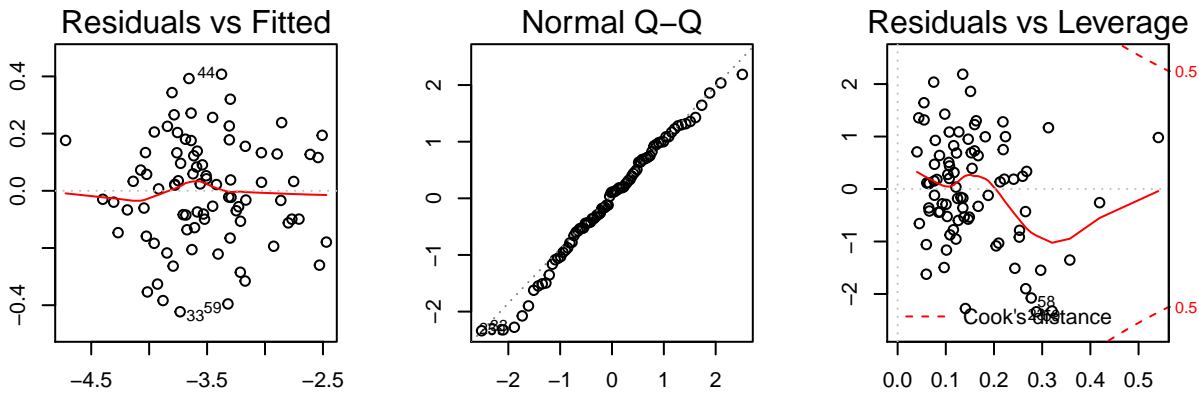
Residuals vs Fitted   **Normal Q–Q**   **Residuals vs Leverage**

This tops fit at about 0.841 (adjusted R^2). However, it is worth noting that the benefit is not much: all those extra variables improved the fit by less than 2%.

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

```
##          df        AIC
## model1  7    1.606574
## model2 10  -12.443044
## model3 14  -17.806321
```

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

Model 4

We will now attempt to develop a custom model based on geographical region.

```
crime_west = crime_data %>% filter(west==1)
crime_central = crime_data %>% filter(central==1)
crime_other_region = crime_data %>% filter(central==0 & west==0)

formula = log_crmrte ~ log_density + sqrt_pctymle + sqrt_pctmin80 +
  prbarr + log_prbconv + log_taxpc

logcrmrte.west.lm4a = lm(formula, data=crime_west)
summary(logcrmrte.west.lm4a)$r.squared
```

```
## [1] 0.8927545
```

```
logcrmrte.central.lm4b = lm(formula, data=crime_central)
summary(logcrmrte.central.lm4b)$r.squared
```

[1] 0.8351015

```
logcrmrte.other_rgn.lm4c = lm(formula, data=crime_other_region)
summary(logcrmrte.other_rgn.lm4c)$r.squared
```

[1] 0.7555855

The models for West and Central regions are better than the general model 2 we have in terms of adjusted R^2.

However, for other regions, our model has lower adjusted R^2 than West or Central. This shows that the observations for other regions need to be analyzed with a model different from the generic one. It could also be that our model is influenced heavily by the observations from the urban areas in West and Central.

	Dependent variable:		
	log_crmrte		
	(1)	(2)	(3)
prbarr	−1.243***	−1.133***	−1.191***
	(0.295)	(0.269)	(0.268)
log_prbconv	−0.252***	−0.228***	−0.181***
	(0.063)	(0.060)	(0.060)
log_density	0.396***	0.286***	0.317***
	(0.039)	(0.050)	(0.049)
sqrt_pctmin80	0.135***	0.126***	0.120***
	(0.015)	(0.014)	(0.014)
sqrt_pctymle	1.027	0.584	0.814
	(0.799)	(0.805)	(0.782)
log_polpc		0.481***	0.499***
		(0.132)	(0.129)
log_taxpc		−0.142	−0.127
		(0.115)	(0.110)
prbpris			−0.198
			(0.313)
avgsen			−0.019**
			(0.009)
wfir			−0.001**
			(0.001)
wfed		0.001	0.001*
		(0.001)	(0.001)
mix			0.380
			(0.389)
Constant	−4.314***	−0.747	−0.335
	(0.256)	(1.328)	(1.329)
Observations	84	84	84
R^2	0.798	0.841	0.864
Adjusted R^2	0.785	0.824	0.841
Residual Std. Error	0.233 (df = 78)	0.211 (df = 75)	0.200 (df = 71)
F Statistic	61.539*** (df = 5; 78)	49.479*** (df = 8; 75)	37.632*** (df = 12; 71)

Note: *p<0.1; **p<0.05; ***p<0.01

CLM on Model 2

CLM 1.0 - Linear in parameters

Our model is represented as:

$$\log\left(crmrte\right) = \beta_0 + \beta_1\,prbarr + \beta_2\,\log\left(prbconv\right) + \beta_3\,\log\left(density\right) + \beta_4\,\sqrt{pctmin80} + \beta_6\,\sqrt{pctymle} + \beta_7\,\log\left(polpc\right) + \beta_8\,\log\left(taxp\right)$$

Any population distribution could be represented as a linear model plus some error (error might be poorly behaved). We chose the above model such that the dependent variable is a linear function of the explanatory variables. Therefore the CLM.1 assumption is met for our all of our models we created

CLM 2.0 - Random Sampling

For this assumption, the data needs to be a random sample drawn from the population.

We first note that the US state of North Carolina is divided in 100 counties (information from Wikipedia). Our dataset contained information from 90 different counties. Though we didn't use data from all counties, there was no indication of non-random sampling during our analysis. We therefore assume that the counties sampled are indeed random and thus our assumption is met for all the models we created

CLM 3.0 - MultiCollinearity

From our EDA it was apparent that none of our variables had constant values. In addition, inspection of the correlation plot we generated indicates there are no perfectly correlated variable pairs.

```
vif(model2)
```

```
##        prbarr   log_prbconv   log_density sqrt_pctmin80   sqrt_pctymle
##      1.475865      1.457300      2.712759      1.160857      1.416303
##     log_polpc      log_taxpc          wfed
##      2.438924      1.370976      2.533486
```

The variance inflation factor does not provide any evidence of multicollinearity as well. We therefore assume CLM 3.0 condition is satisfied

CLM 4.0 - Zero Conditional Mean

By examining the residuals verus fitted values plot for model, we conclude that the assumption of zero conditional mean is met. The red spline curve does not deviate much from zero
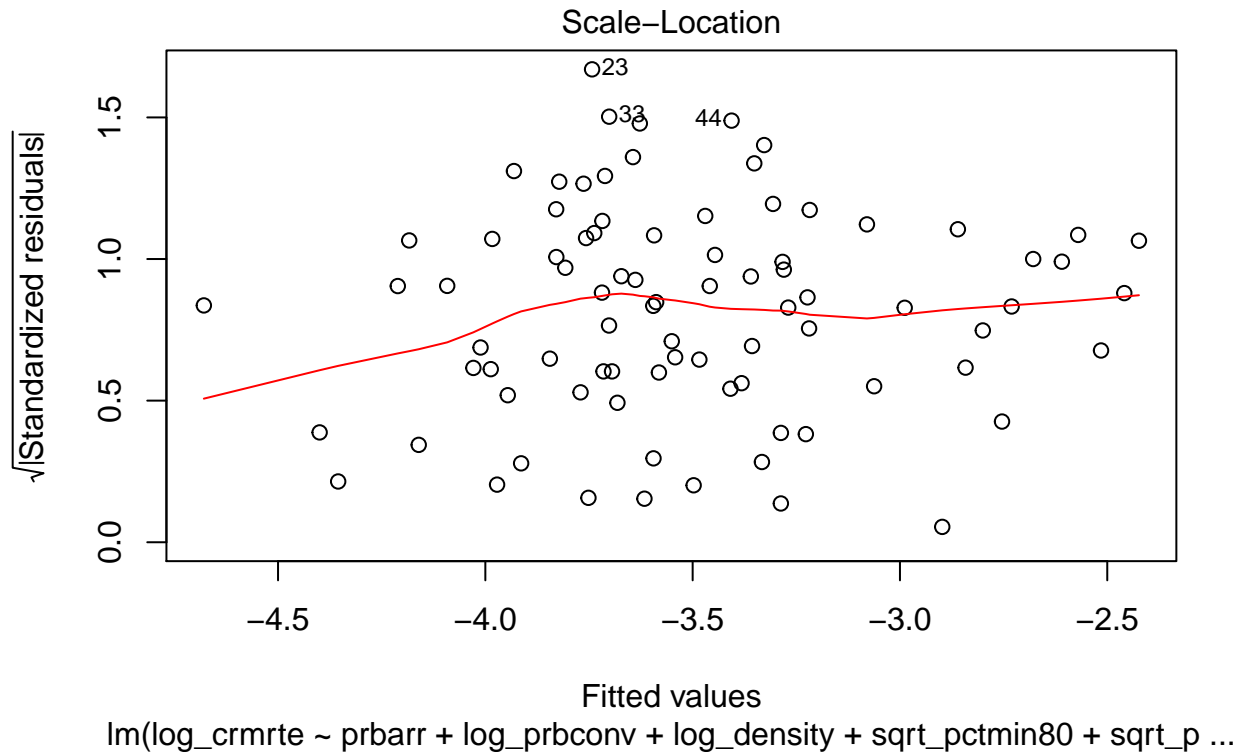
```
plot(model2, which = 1)
```

Residuals vs Fitted

Fitted values
lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...
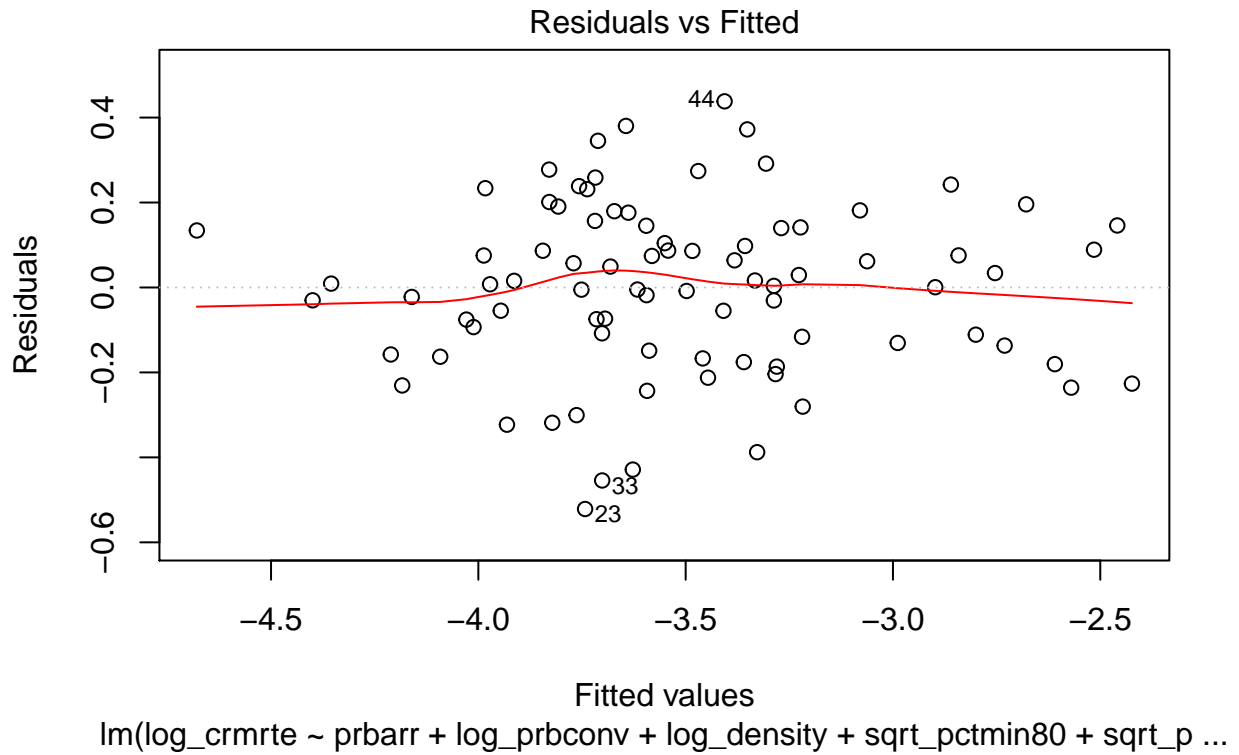
CLM 5.0 - Homoskedasticity

When we examined the residuals versus fitted values plot, it was apparent that the variance of errors to the right of the plot is smaller than the variance of errors in the middle of the plot. This suggested heteroskedasticity, so we examined the scale-location plot. The spline curve on the scale-location plot is curved rather than flat, indicating heteroskedasticity. Despite this violation of CLM, we are able to proceed with our OLS model by using heteroskedasticity-robust standard errors.

```
plot(model2, which = 3)
```

Scale–Location

lm(log_crmrte ~ prbarr + log_prbconv + log_density + sqrt_pctmin80 + sqrt_p ...

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 14.698, df = 8, p-value = 0.06529
```

We also conducted the studentized Breusch-Pagan test where the null hypothesis states that the model supports homoskedasticity. The p-value obtained indicates marginal statistical significance.

CLM 6.0 - Normality of error terms

Looking at the histogram of the residual term indicates a more or less normal distribution. This can further be confimed by the Shapiro test. The null hypothesis of this test is that residuals have a normal distribution. The p-value indicates we do not reject this null hypothesis.

```
hist(model2$residuals, breaks=50)
```

Histogram of model2$residuals
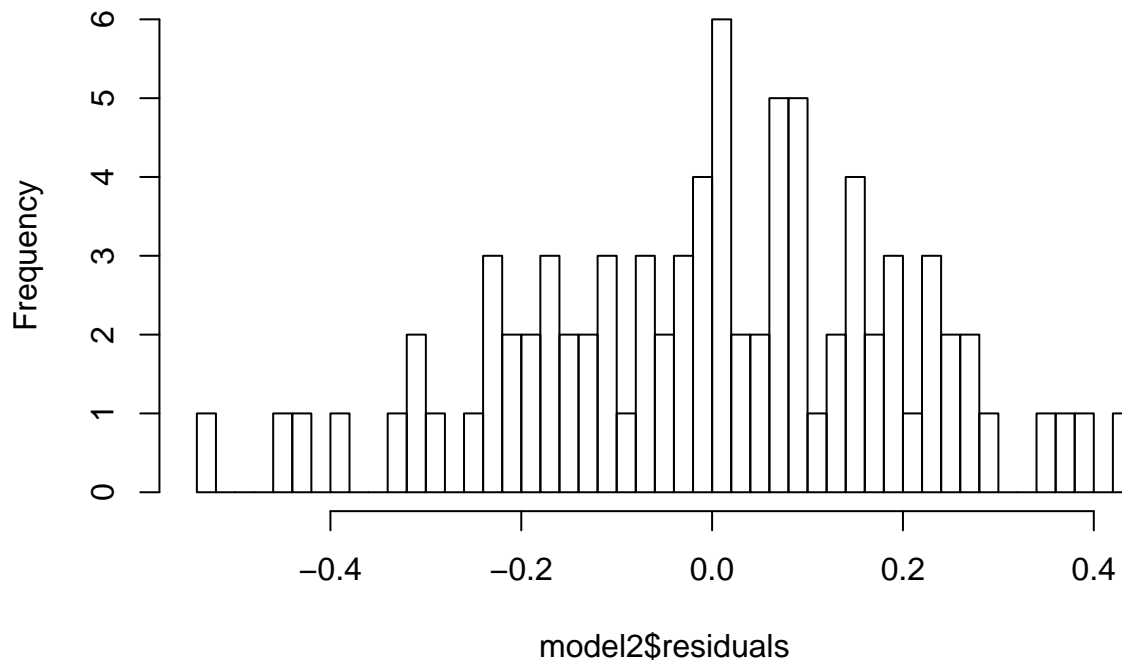

```r
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.99174, p-value = 0.8764
```

6. 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 income to size ratio: 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.

- Social factors and family cohesiveness: 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.

- Repeat crimes: The data does not have information about serial criminals as against first-time offenders. If most of the crime is due to repeat offenders, then we will have to accommodate our policies accordingly.

- Underreporting and unreported crimes: It is also worth checking via other sources how much crime goes unreported in North Carolina, and whether data is often "corrected" as time goes by.

- Neighborhood crime watch: Do neighborhoods have crime-watch communities? Having such vigil can reduce crime in areas of high density.

We can talk about biases for a lot of the above omitted variables:

Education, neighborhood watch, family cohesiveness:

- Positive bias on: taxpc
- Negative bias on: prbarr, prbconv, prbpris, avgsen

Unemployment, poverty, family-income-to-size ratio, substance abuse:

- Positive bias on: prbarr, prbconv, prbpris, avgsen
- Negative bias on: taxpc

7. Conclusion

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

- 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.