Lab 3 (Bhuvnesh Sharma, Weixin Wu)

Bhuvnesh Sharma, Weixin Wu

March 22, 2018

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(sandwich)
library(car)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```r
library(effsize)
```

Introduction

Crime is huge menace in the society, there have been many attempts in past to reduce crime rates within communities in North Carolina. Traditional politicians and conventional approach has assumed that tough on crime is an effective tool to curb crime. Being tough on crime is regularly misunderstood as longer and mandatory prison sentences. This misguided strategy can lead to state's higher investment on prison infrastructure and also make laws which can promote mandatory prison sentences appear as effective crime fighting tool. The goal of this study is to uncover the real facts around the crime rates within North Carolina to develop effective state policy around to reduce crime rates. Key motivation of the report discover the real drivers and instruments which the policy makers can use and have meaningful impact on crime. Study intends to empower the state politicians , key legislative leaders with key facts which have been based on data and not on conventional empirical narratives. Study intends to discover key variables which have major impact on crime rates in North Carolina . This information would be critical for voters to understand so that they can make an informed decision on a important election issue.

Data Cleansing

```r
crimeData <- read.csv("crime_v2.csv")
summary(crimeData)
```

```
##      county           year        crmrte            prbarr      
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277  
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568  
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095  
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492  
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438  
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091  
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6        
##         prbconv      prbpris           avgsen           polpc          
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746  
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231  
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485  
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702  
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877  
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054  
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6         
##     density            taxpc             west           central      
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000  
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736  
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000  
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000  
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6       
##      urban            pctmin80           wcon            wtuc      
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6  
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6  
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5  
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7  
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4  
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2  
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6      
##       wtrd            wfir            wser             wmfg      
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4  
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9  
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2  
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6  
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6  
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9  
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6      
##       wfed            wsta            wloc            mix         
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961  
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074  
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186  
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884  
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175  
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512  
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6        
##     pctymle       
##  Min.   :0.06216  
##  1st Qu.:0.07443  
##  Median :0.07771  
##  Mean   :0.08396  
##  3rd Qu.:0.08350  
```

```
##  Max.   :0.24871
##  NA's   :6
```

As shown in the summary table, there are 6 NA's in every variable. After reviewing the data, we found that all NA's are in 6 rows, so we removed those rows as they did not provide any information.

```r
crimeData2 <- crimeData[complete.cases(crimeData),]
```

Variable 'prbconv' was incorrectly displayed as a text field. We converted it to numermic.

```r
crimeData2 <- transform(crimeData2, prbconv = as.numeric(as.character(prbconv)))
summary(crimeData2$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

Usually the probability variable should be bound between 0 and 1. However, there is one observation with 'prbarr' (probability of arrest) higher than 1, and 10 observations with 'prbconv' (probability of conviction) higher than 1.

```r
nrow(crimeData2[which(crimeData2$prbarr>1),])
```

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

```r
nrow(crimeData2[which(crimeData2$prbconv>1),])
```

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

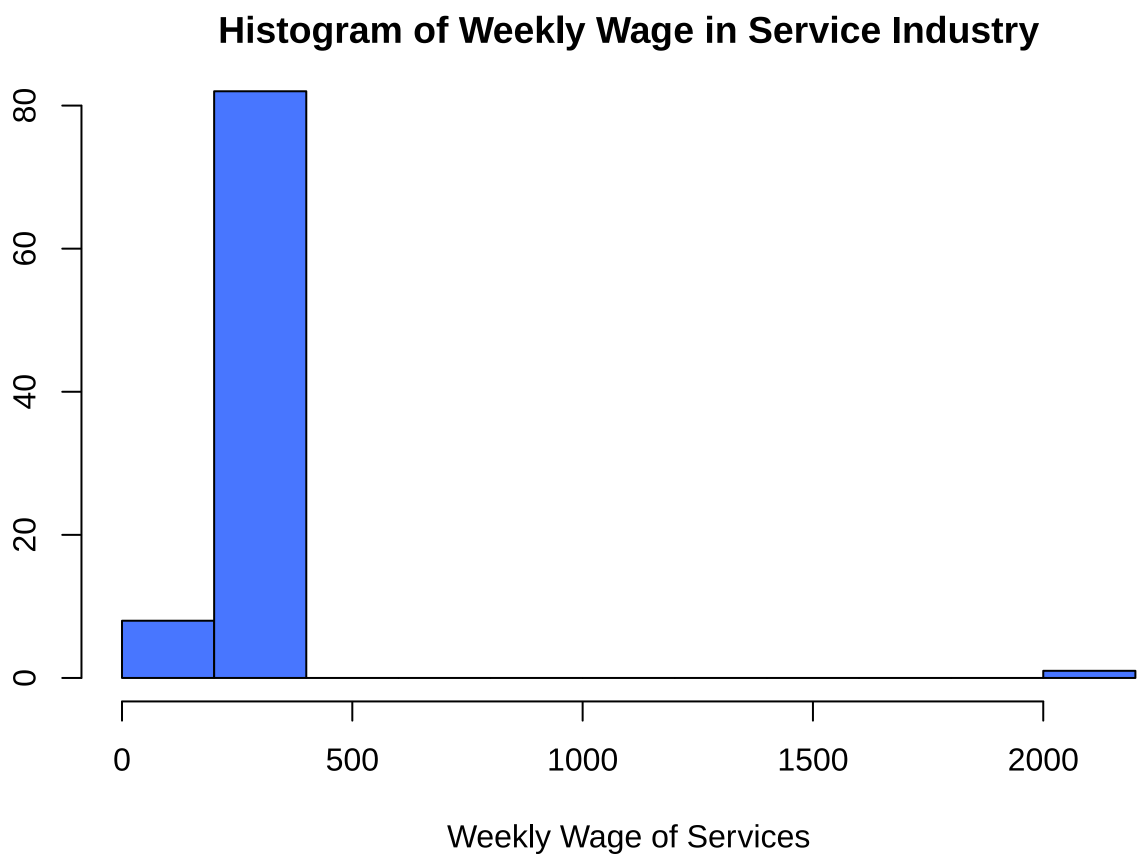
Variable 'prbarr' is defined as the ratio of arrests to offenses. One possible explanation for 'prbarr' being greater than 1 is that multiple people who convicted a single crime together is counted as one conviction but multiple arrests.
Variable 'prbconv' is defined as the ratio of convictions to arrests. One possible explanation for 'prbconv' being greater than 1 is that one person who is convicted of multiple crimes but only arrested once.
Without further information on the variables, we could not conclude whether these values are invalid. So we left those observations in the data.

Variable 'pctmin80' (percent of minority in 1980) is expressed as percentages. We converted it into decimals to be consistent with variable 'pctymle' (percent of young male).

```r
crimeData2$pctmin80_2 <- crimeData2$pctmin80/100
```

The max value of variable 'wser' (weekly wage of service industry) is significantly higher than its third quartile. The histogram below shows that the max value (2177.068) is significantly higher than the rest of values.

```r
par(mar=c(4, 2, 2, 2))
hist(crimeData2$wser, main="Histogram of Weekly Wage in Service Industry",
     xlab = "Weekly Wage of Services" , col = "royalblue1")
```




```
crimeData2[which(crimeData2$wser>2000),]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
##      density    taxpc west central urban pctmin80     wcon    wtuc
## 84 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565
##        wtrd     wfir     wser   wmfg    wfed    wsta    wloc        mix
## 84 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944
##       pctymle pctmin80_2
## 84 0.07008217   0.643482
```

We examined County 185, whose wser is 2177.068. We noticed that most other weekly wage variables for County 185 are below the means. You would expect that a richer county would have weekly wage in multiple industries to be higher than the average. So it's very unlikely for a county to have lower than average weekly wage on constructure, transportation, retail, finance, etc. but extremely high weekly wage on the service industry.

In addition, an average weekly wage of 2177.068 in 1987 is an unreasonable value. So we believed 2177.068 is erroneous. We removed this observation from the data.

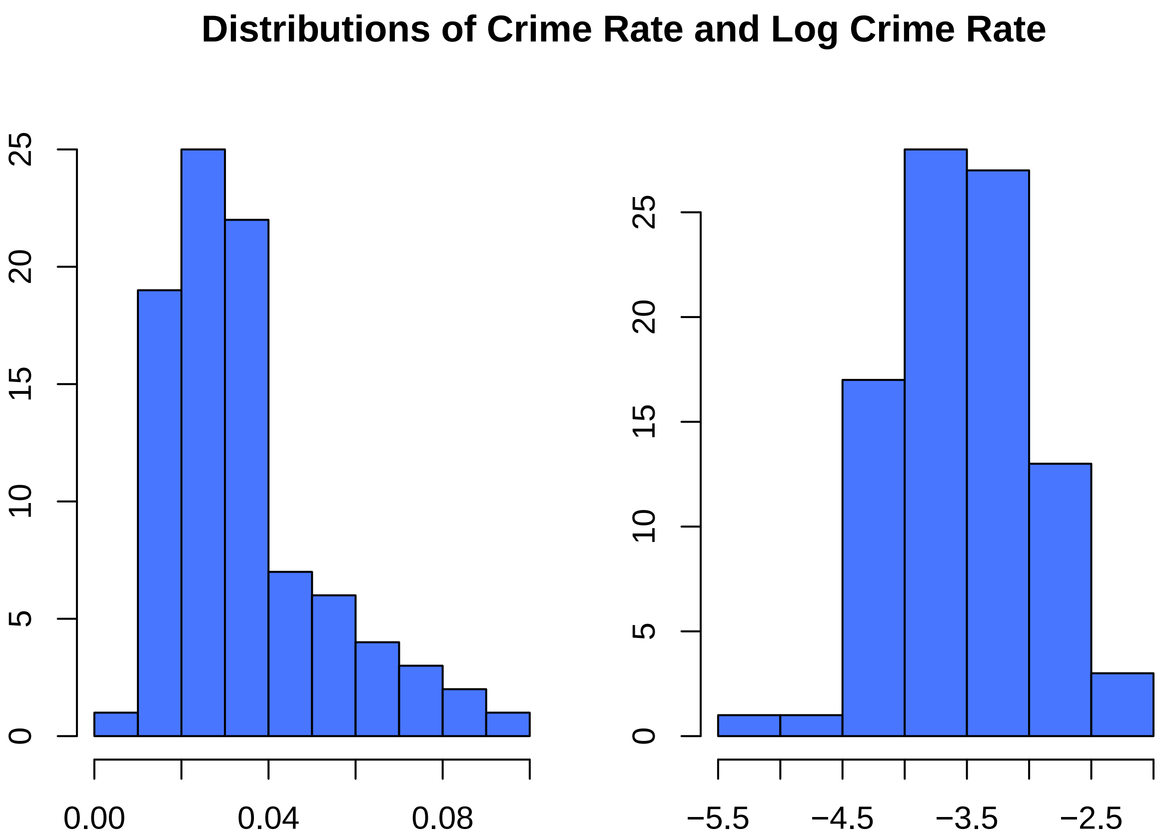
```
crimeData2 <- crimeData2[which(crimeData2$wser<2000),]
```

Exploratory Data Analysis

Crimes committed per person (crmrte)

The distribution of crime rate is skewed to the right, so we considered taking the log of crime rate. After the log transformation, the distribution of crmrte_log is closer to normal. Semilogarithmic form is interpretable later in modeling: it tells us what's the percentage change in crime rate in response to a unit change in explantory variables. Our target variable is crmrte_log.

```
par(mfrow=c(1,2), oma=c(0, 0, 2, 0))
par(mar=c(2, 2, 2, 2))
hist(crimeData2$crmrte, main="",xlab = "Crime Rate" , col = "royalblue1")
crimeData2$crmrte_log = log(crimeData2$crmrte)
hist(crimeData2$crmrte_log, main="", xlab = "Log of Crime Rate", col = "royalblue1")
mtext("Distributions of Crime Rate and Log Crime Rate", outer=TRUE, cex = 1.2, font=2)
```



Probability of arrest (prbarr)

The scatter plot of crmrte vs. prbarr on the left shows an exponential decay trend.

In addition, the variation of crmrte decreases substantially as prbarr increases. We took the log of crime rate, and then re-graph the scatter plot (shown on the right). The scatter plot of crmrte_log vs. prbarr indicates a more linear relationship and the variation of crmrte_log does not vary as much with prbarr. The correlation coefficient further supports the transformation.

- The correlation between crmrte and prbarr is -0.41

- The correlation between crmrte_log and prbarr is -0.50

In addition, we noticed a leveraged data point in the graph, that's County 115. County 115 has significantly higher probability of arrest than all other counties. If we removed County 115 from the data, the correlation coefficient reduced from -0.50 to -0.39. This indicates that County 115 could be an infludential observation. Later when building the model, we will calculate Cook's distance to confirm that County 15 is an influential observation and also address the impact of influential observations to parameter estimates.

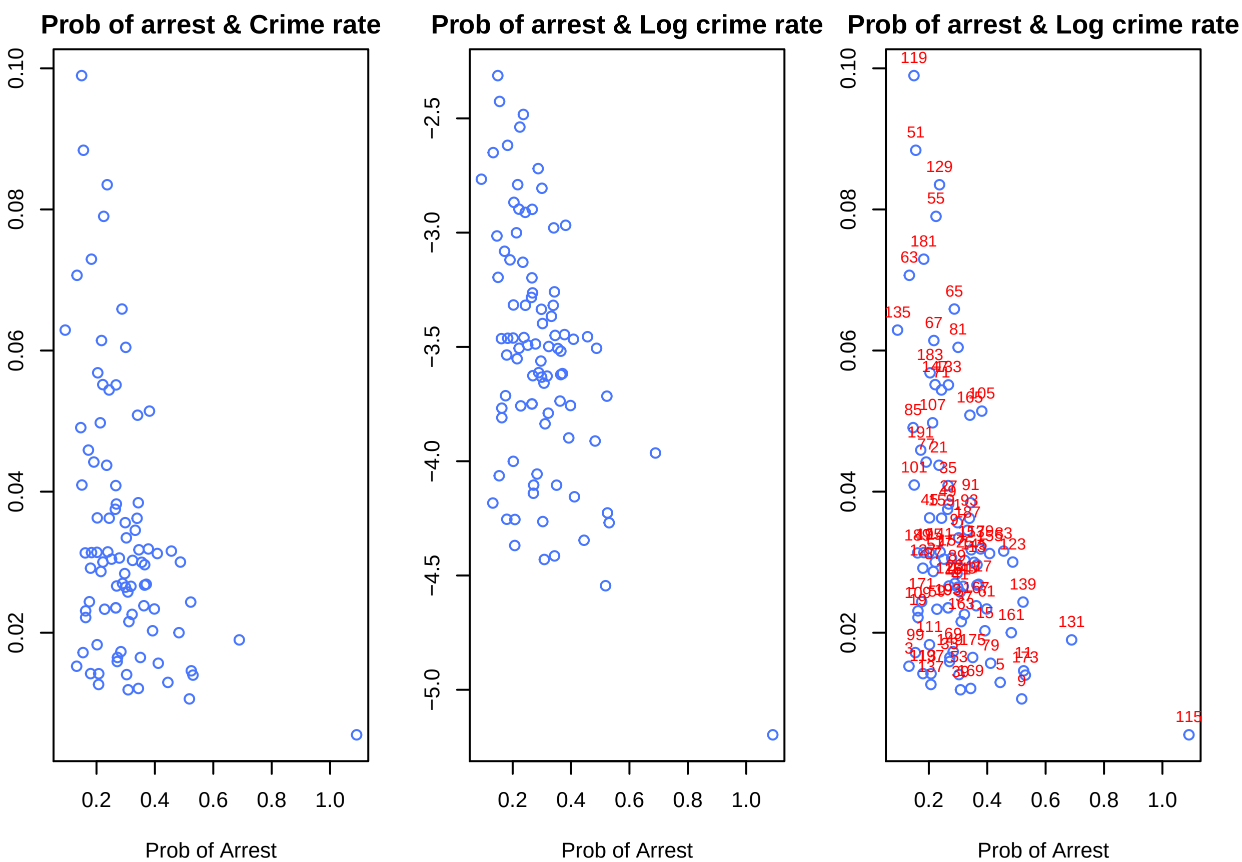
```
par(mfrow=c(1,3))
par(mar=c(4, 2, 2, 2))
plot(crimeData2$prbarr, crimeData2$crmrte , main = "Prob of arrest & Crime rate" ,
     xlab = "Prob of Arrest",ylab = "Crime Rate",col="royalblue1")
plot(crimeData2$prbarr, crimeData2$crmrte_log, main = "Prob of arrest & Log crime rate" ,
     xlab = "Prob of Arrest",ylab = "Log of Crime Rate",col="royalblue1")
cor(crimeData2$prbarr, crimeData2$crmrte)
```

```
## [1] -0.4076239
```

```
cor(crimeData2$prbarr, crimeData2$crmrte_log)
```

```
## [1] -0.4964904
```

```
plot(crimeData2$prbarr, crimeData2$crmrte,main = "Prob of arrest & Log crime rate" ,
     xlab = "Prob of Arrest",ylab = "Log of Crime Rate",col="royalblue1")
text(crimeData2$prbarr, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3,col = "red")
```




```
crimeData3 <- crimeData2[which(crimeData2$county!=115),]
cor(crimeData3$prbarr, crimeData3$crmrte_log)
```

```
## [1] -0.3949839
```

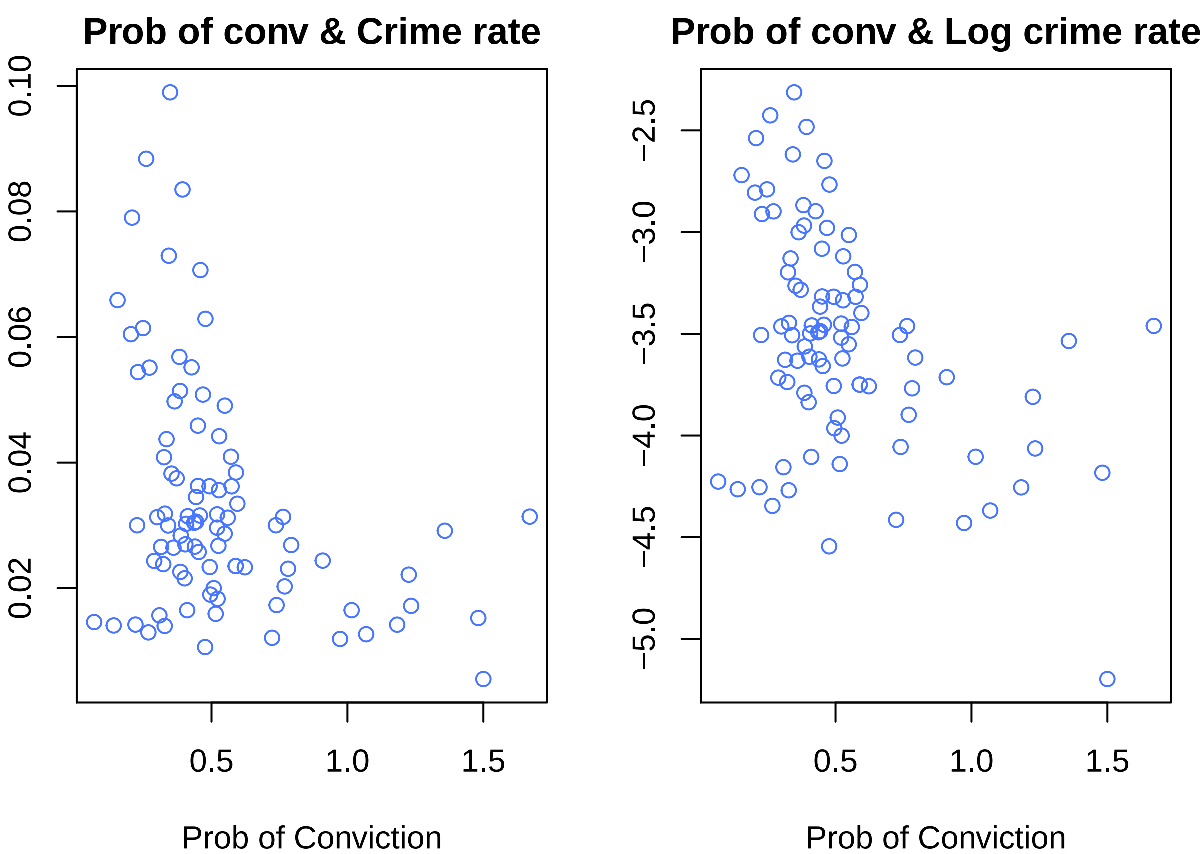
Probability of conviction (prbconv)

Similar to prbarr, the scatter plot of crmrte vs. prbconv on the left shows an exponential decay trend.

In addition, the variation of crmrte decreases substantially as prbconv increases. We took the log of crime rate, and then re-graph the scatter plot (shown on the right). The scatter plot of crmrte_log vs. prbconv indicates a more linear relationship and the variation of crmrte_log does not vary as much with prbconv The correlation coefficient further supports the transformation.

- The correlation between crmrte and prbarr is -0.37
- The correlation between crmrte_log and prbarr is -0.41

```
par(mfrow=c(1,2))
par(mar=c(4, 2, 2, 2))
plot(crimeData2$prbconv, crimeData2$crmrte,main = "Prob of conv & Crime rate" ,
     xlab = "Prob of Conviction",ylab = "Crime Rate",col="royalblue1")
plot(crimeData2$prbconv, crimeData2$crmrte_log,main = "Prob of conv & Log crime rate" ,
     xlab = "Prob of Conviction",ylab = "Log of Crime Rate",col="royalblue1")
```




```
cor(crimeData2$prbconv, crimeData2$crmrte)
```

```
## [1] -0.3728922
```

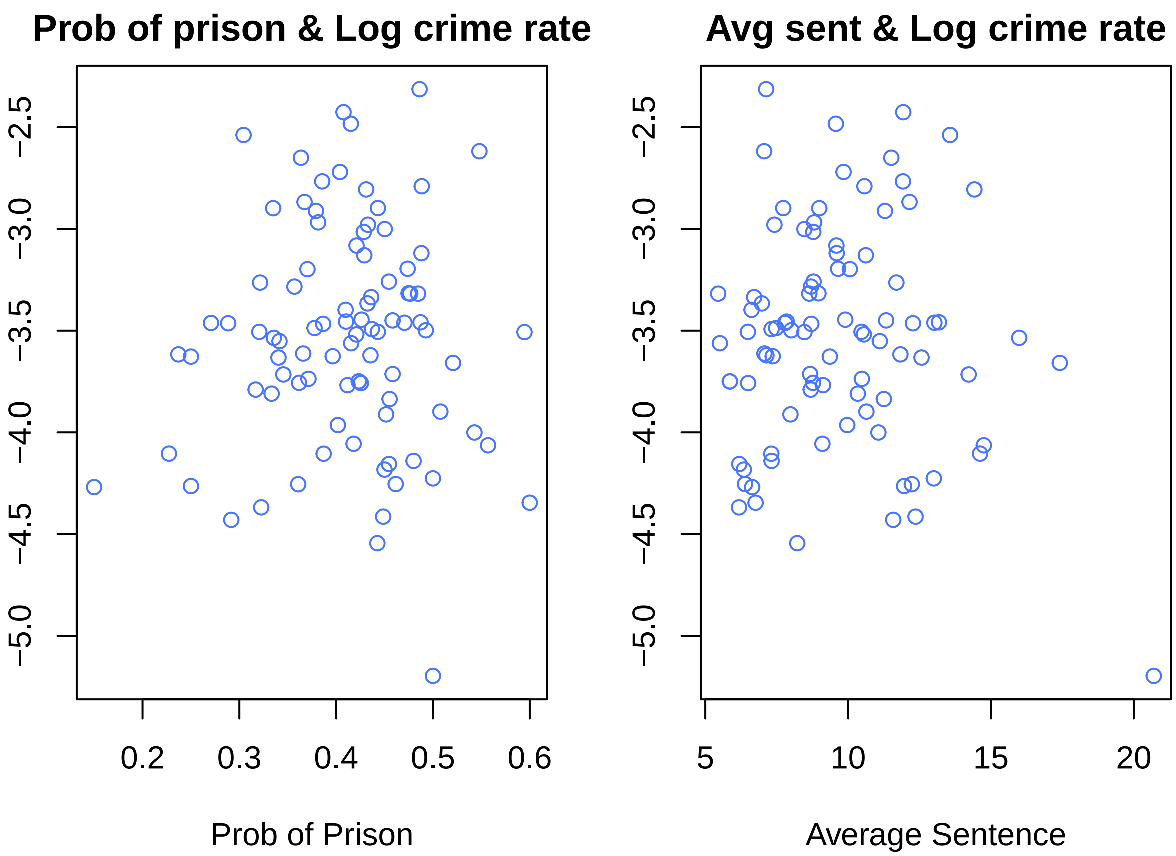
```
cor(crimeData2$prbconv, crimeData2$crmrte_log)
```

```
## [1] -0.4128166
```

Probability of prison (prbpris) | Average sentence days (avgsen)

Neither scatter plots below (prbpris vs. crmrte_log, avgsen vs. crmrte_log) shows obvious relationships. The correlation coefficients are only 0.03 and -0.08 respectively.

```
par(mfrow=c(1,2))
par(mar=c(4, 2, 2, 2))
plot(crimeData2$prbpris, crimeData2$crmrte_log,main = "Prob of prison & Log crime rate" ,
     xlab = "Prob of Prison",ylab = "Log of Crime Rate",col="royalblue1")
plot(crimeData2$avgsen, crimeData2$crmrte_log,main = "Avg sent & Log crime rate" ,
     xlab = "Average Sentence",ylab = "Log of Crime Rate",col="royalblue1")
```




```
cor(crimeData2$prbpris, crimeData2$crmrte_log)
```

```
## [1] 0.02938727
```

```
cor(crimeData2$avgsen, crimeData2$crmrte_log)
```

```
## [1] -0.07567514
```

Police per capita (polpc)

Similar to probabilities of arrest and conviction, we observed a linear relationship between crime rate and policy per capita in the scatter plot below. The scatter plot also shows that County 115 has significantly higher police per capital than any other counties, County 115 is a highly leveraged observation.

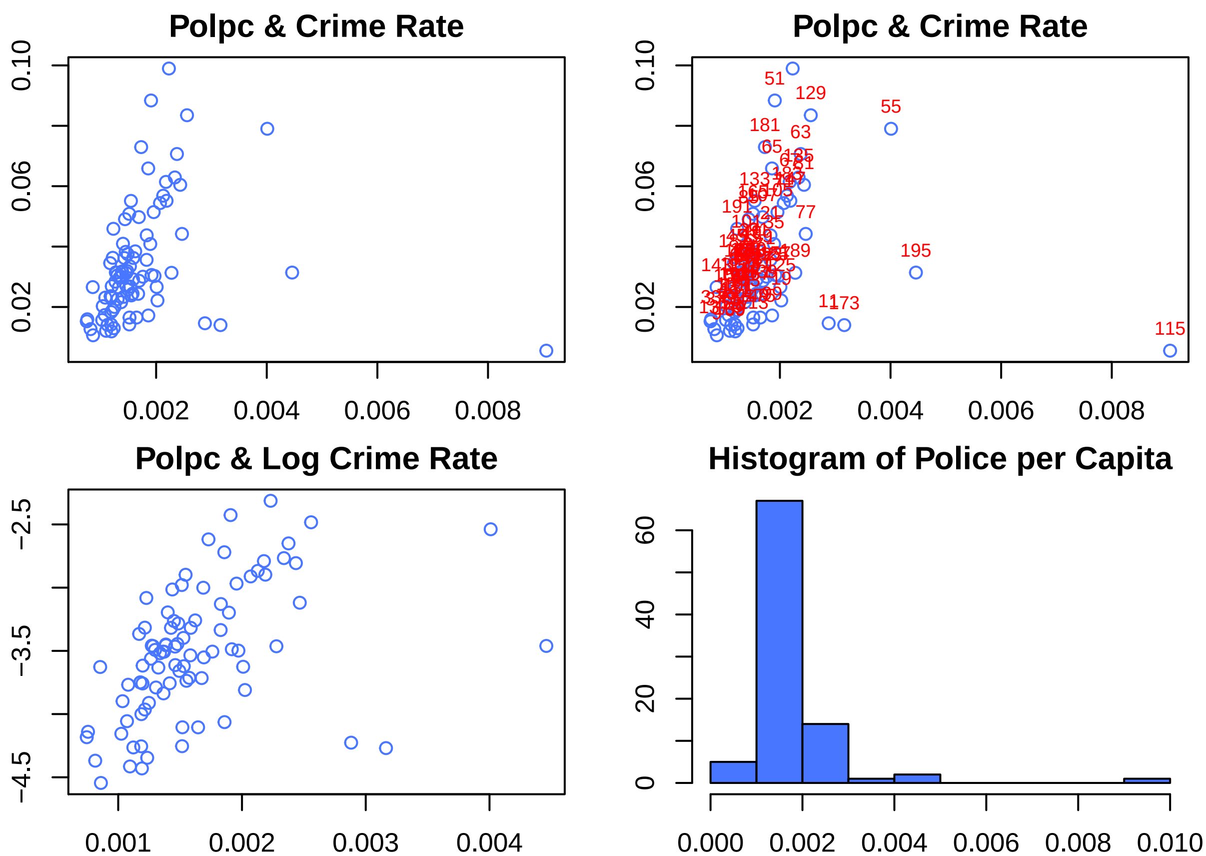
In addition, the variation of crmrte increases as prbconv increases, which justifies taking the log of crmrte. The correlation between crmrte_log and polpc (after removing County 115) is 0.45.

We also noticed that the distribution of polpc is highly skewed to the right, so we took the log of polpc. The correlation coefficient (after removing County 115) increased from 0.45 to 0.54.

```r
par(mfrow=c(2,2))
par(mar=c(2, 2, 2, 2))
plot(crimeData2$polpc, crimeData2$crmrte,main = "Polpc & Crime Rate" ,
     xlab = "Police per capita",ylab = "Crime Rate",col="royalblue1")
plot(crimeData2$polpc, crimeData2$crmrte,main = "Polpc & Crime Rate" ,
     xlab = "Police per capita",ylab = "Crime Rate",col="royalblue1")
text(crimeData2$polpc, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3,col = "red")
plot(crimeData3$polpc, crimeData3$crmrte_log,main = "Polpc & Log Crime Rate" ,
     xlab = "Police per capita",ylab = "Log of Crime Rate",col="royalblue1")
cor(crimeData3$polpc, crimeData3$crmrte_log)
```

```
## [1] 0.453951
```

```r
hist(crimeData2$polpc, main="Histogram of Police per Capita",
     xlab = "Police per capita" , col = "royalblue1")
```




```r
crimeData2$polpc_log <- log(crimeData2$polpc)
crimeData3 <- crimeData2[which(crimeData2$county!=115),]
cor(crimeData3$polpc_log, crimeData3$crmrte_log)
```

```
## [1] 0.541829
```

People per square mile (density) | If in SMSA (urban)

The histogram shows the distribution of density is highly skewed to the right, so we took the log of density. The scatter plot shows County 173 is highly leveraged as it has much lower population density than other counties. Removing County 173 significantly increases correlation coefficient from 0.49 to 0.68. The correlation between density and crmrte_log (without County 173) is 0.63, which is lower than the correlation between density_log and crmrte_log (without County 173) of 0.68. This further confirms that log of density has a stronger linear relationship with log of crime rate than density does.
Urban is a binary variable.

The box plot shows that the the mean and interquartile range of density is significantly different depending on whether county is in urban area or not. Log of density is highly correlated with urban with a correlation coefficient of 0.66. When building the model, we should avoid putting both variables in the model for two reasons:

1. Adding the second variable doesn't explain much additional variation of the response variable

2. High correlation can greatly increase the standard errors of parameter estimates

```
par(mfrow=c(2,2))
par(mar=c(4, 2, 2, 2))
hist(crimeData2$density, main="Histogram of People per Square Mile",
     xlab = "People per Square Mile / Density " , col = "royalblue1")
crimeData2$density_log <- log(crimeData2$density)
plot(crimeData2$density_log, crimeData2$crmrte_log,main = "Log Density & Crime Rate" ,
     xlab = "Log of Density",ylab = "Log of Crime Rate",col="royalblue1")
plot(crimeData2$density_log, crimeData2$crmrte_log,main = "Log Density & Crime Rate" ,
     xlab = "Log of Density",ylab = "Log of Crime Rate",col="royalblue1")
text(crimeData2$density_log, crimeData2$crmrte_log, labels = crimeData2$county, cex=0.7, pos=3 , col =
crimeData4 <- crimeData2[which(crimeData2$county!=173),]
cor(crimeData2$density_log, crimeData2$crmrte_log)
```

```
## [1] 0.4909562
```

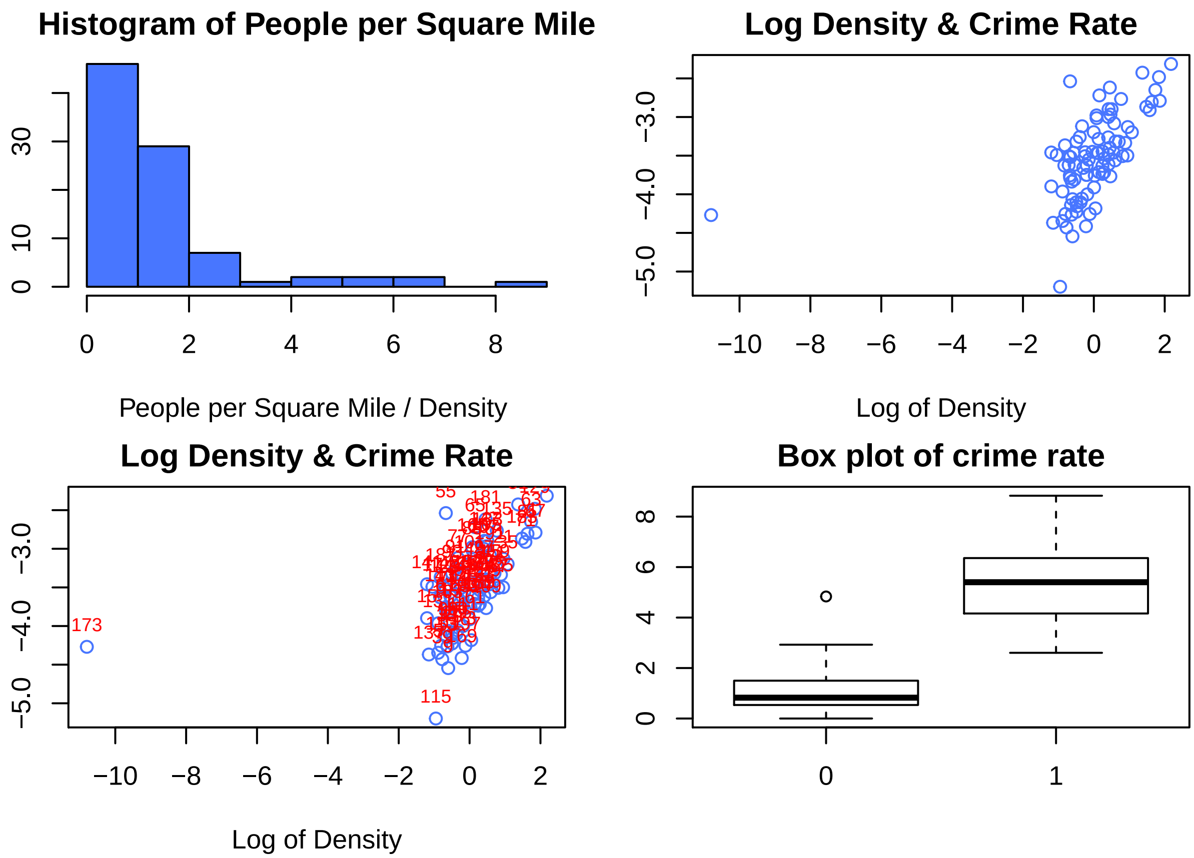
```
cor(crimeData4$density_log, crimeData4$crmrte_log)
```

```
## [1] 0.677355
```

```
cor(crimeData4$density, crimeData4$crmrte_log)
```

```
## [1] 0.6281475
```

```
boxplot(density~urban, data=crimeData2 , main = "Box plot of crime rate")
```



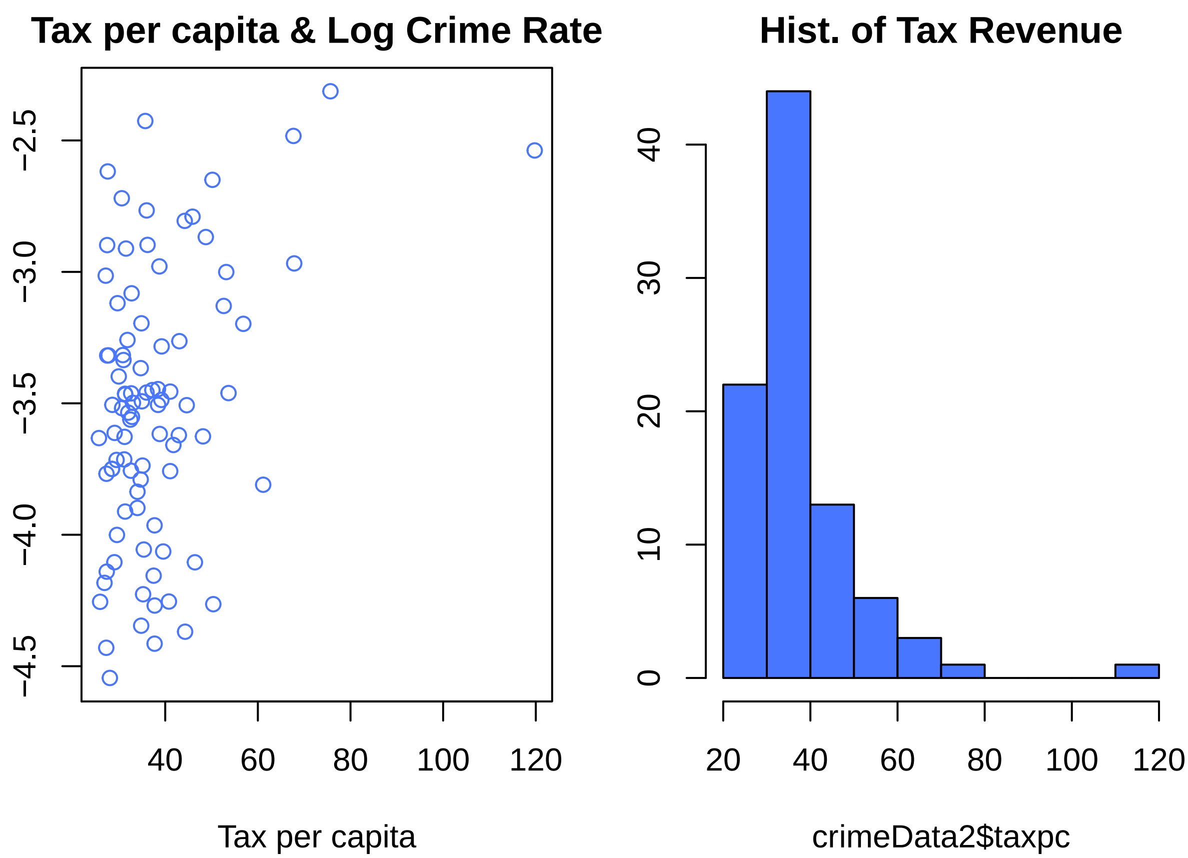

```r
cor(crimeData4$urban, crimeData4$density_log)
```

```
## [1] 0.660531
```

Tax revenue per capita (taxpc)

The scatter plot indicates there may be a weak linear relationship between taxpc and crmrte_log. The histogram of taxpc is skewed to the right, so we considered taking the log of taxpc. However, the correlation between taxpc and crmrte_log (0.37) is slightly higher than the correlation between taxpc_log and crmrte_log (0.36).

```r
par(mfrow=c(1,2))
par(mar=c(4, 2, 2, 2))
plot(crimeData3$taxpc, crimeData3$crmrte_log,main = "Tax per capita & Log Crime Rate" ,
     xlab = "Tax per capita" , ylab = "Log of Crime Rate" , col="royalblue1")
hist(crimeData2$taxpc, main="Hist. of Tax Revenue",col = "royalblue1")
```




```r
crimeData2$taxpc_log <- log(crimeData2$taxpc)
cor(crimeData2$taxpc, crimeData2$crmrte_log)
```

```
## [1] 0.3711452
```

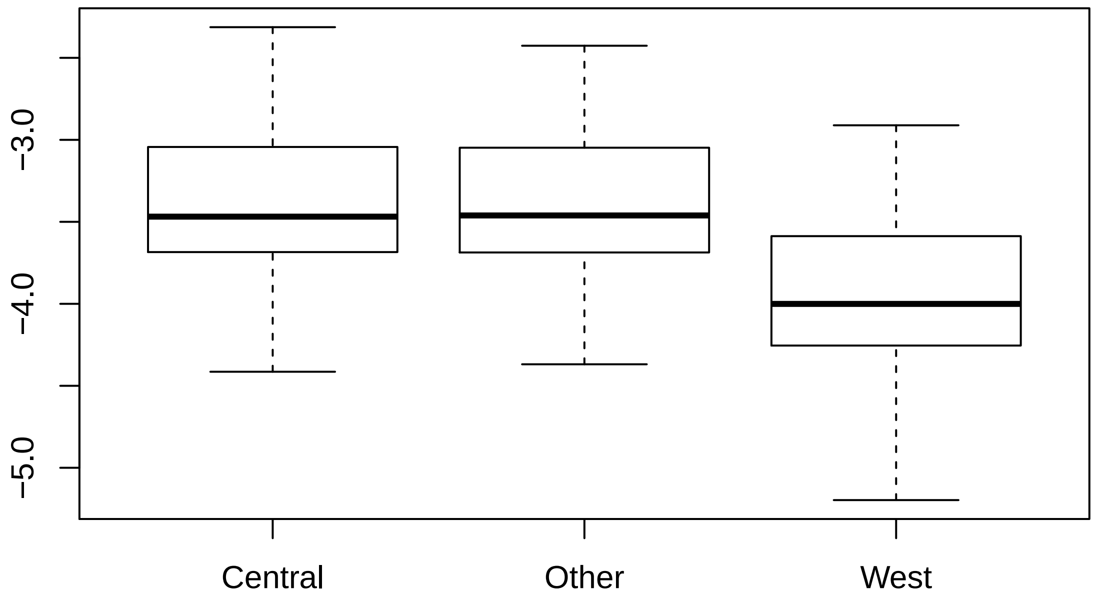
```r
cor(crimeData2$taxpc_log, crimeData2$crmrte_log)
```

```
## [1] 0.3570773
```

If in western/central North Carolina

We created a variable, area, to categorize the area counties reside in. Area takes three values: West, Central, and Other. The box plot shows that the mean and interquartile range of crmrte_log is very similar between Central and Other. The crmrte_log for West area is lower than other areas. In modeling, we can group "Central" and "Other" together and only add the variable "West" to the model.

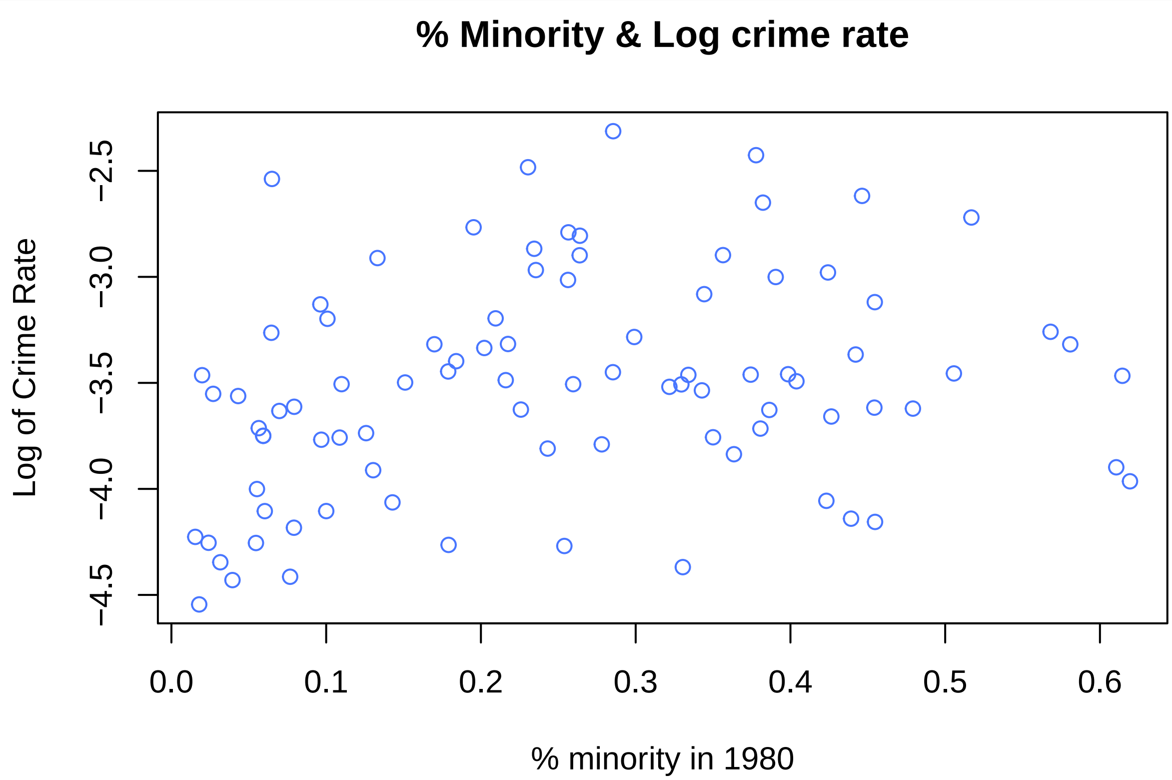
```r
crimeData2$area <- ifelse(crimeData2$west==1, "West", ifelse(crimeData2$central==1, "Central", "Other"))
boxplot(crmrte_log~area, data=crimeData2)
```



Percent of minority in 1980 (pctmin80)

The scatter plot shows a weak linear relationship between log of crime rate and percent of minority. Low correlation coefficient (0.3) also confirms that. Also note that the percent of minority is highly negatively correlated with the indicator "west". This indicates west counties have lower percentage of minority.

```
plot(crimeData3$pctmin80_2, crimeData3$crmrte_log,main = "% Minority & Log crime rate" ,
     xlab = "% minority in 1980" , ylab = "Log of Crime Rate" , col="royalblue1")
```



```
cor(crimeData2$pctmin80_2, crimeData2$crmrte_log)
```

```
## [1] 0.2957882
```

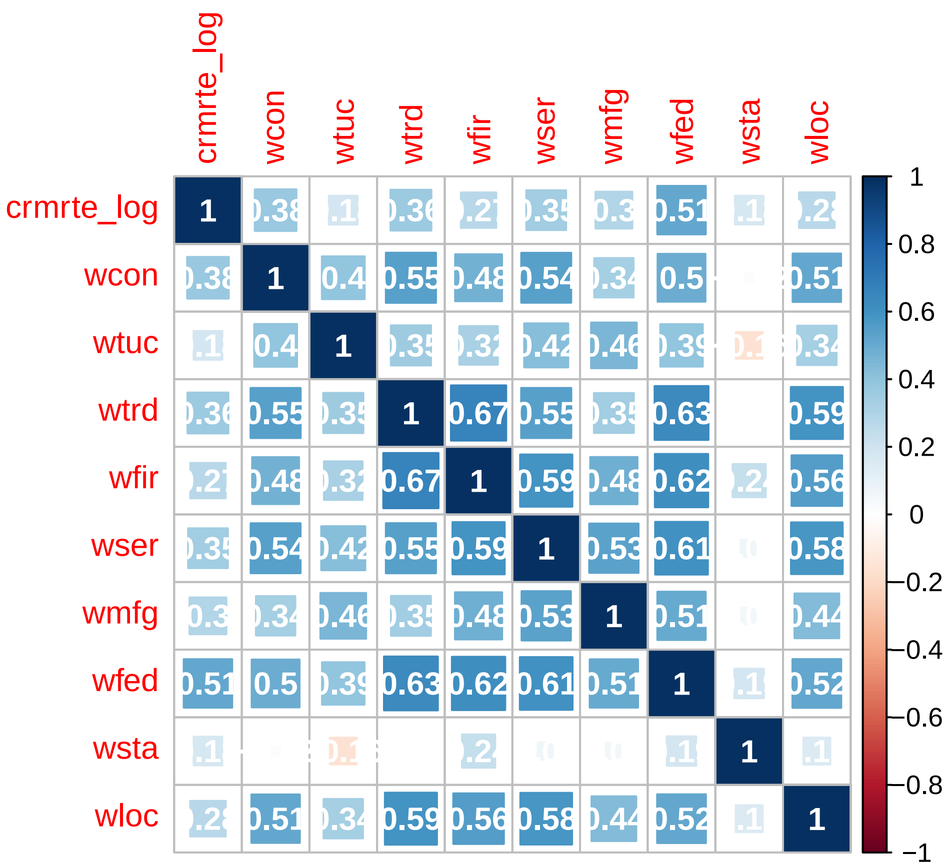
```
cor(crimeData2$pctmin80_2, crimeData2$west)
```

```
## [1] -0.646101
```

Weekly wages

There are nine variables related to weekly wages in the data. They represent weekly wages in different industries. As the correlation matrix shows, most of the weekly wages variables are highly correlated except for wsta. When building the model, we should avoid putting all the correlated variables in the model for the same reason pointed out in the density/urban section of the EDA. We also noticed that log of crime rate has the strongest linear relationship with wfed with correlation coefficeint of 0.51.

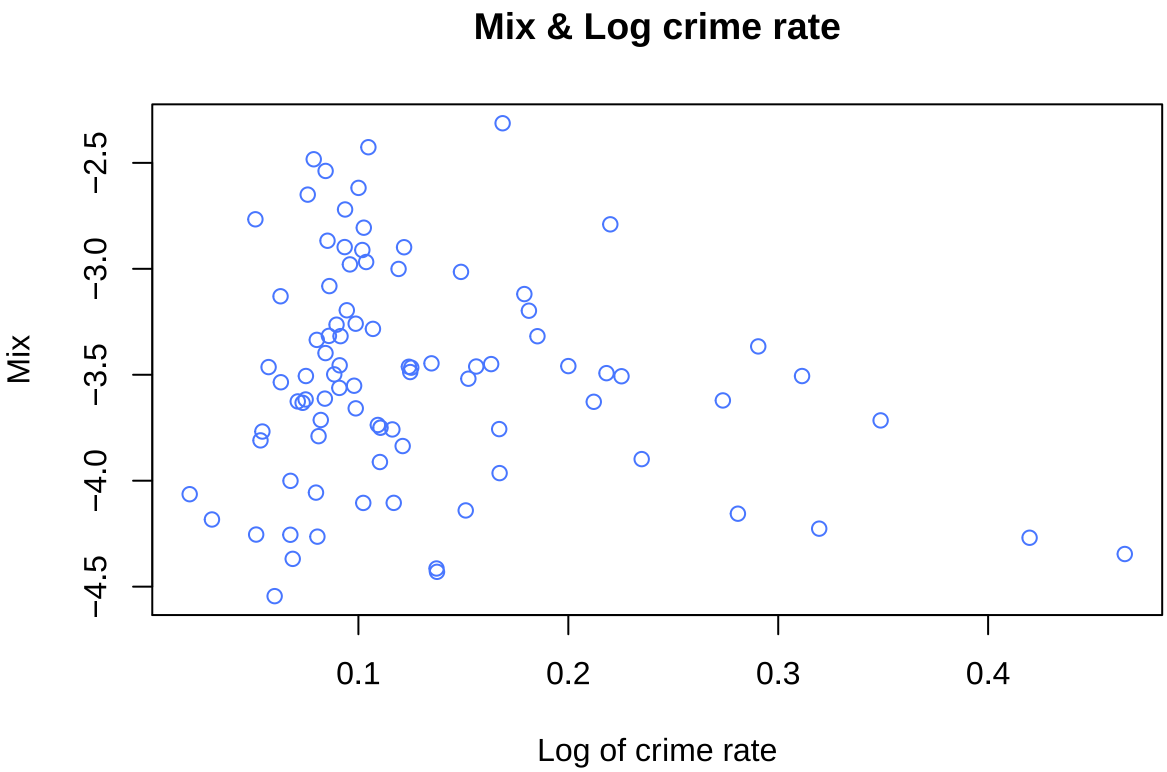
```
crimeData_temp1 <- crimeData2[,c("crmrte_log","wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "w
corr_DenUr <- cor(crimeData_temp1, use="pairwise")
corrplot(corr_DenUr, method="square", addCoef.col="white")
```



Offense mix: face-to-face/other (mix)

The scatter plot doesn't indicate a strong relationship between mix and log of crime rate. The weak correlation coefficient (-0.15) also confirms that.

```r
plot(crimeData3$mix, crimeData3$crmrte_log,main = "Mix & Log crime rate",
     xlab="Log of crime rate",ylab="Mix", col="royalblue1")
```



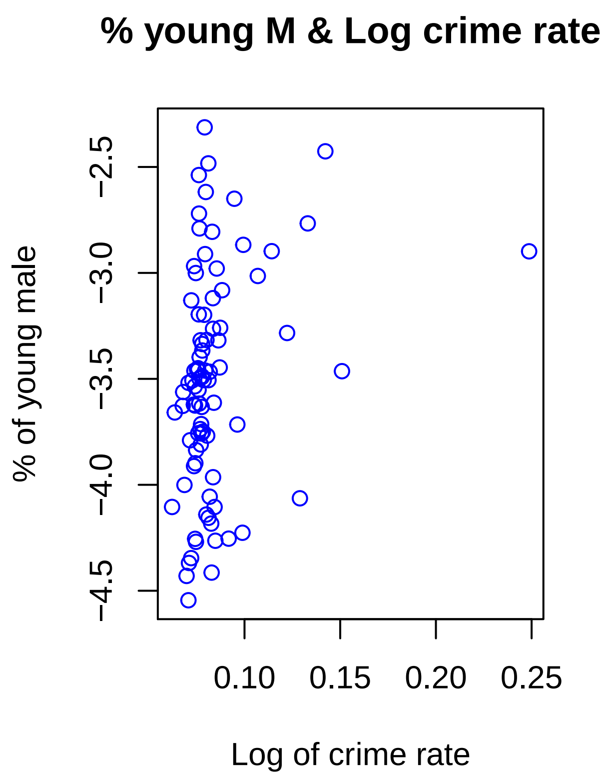

```r
cor(crimeData2$mix, crimeData2$crmrte_log)
```

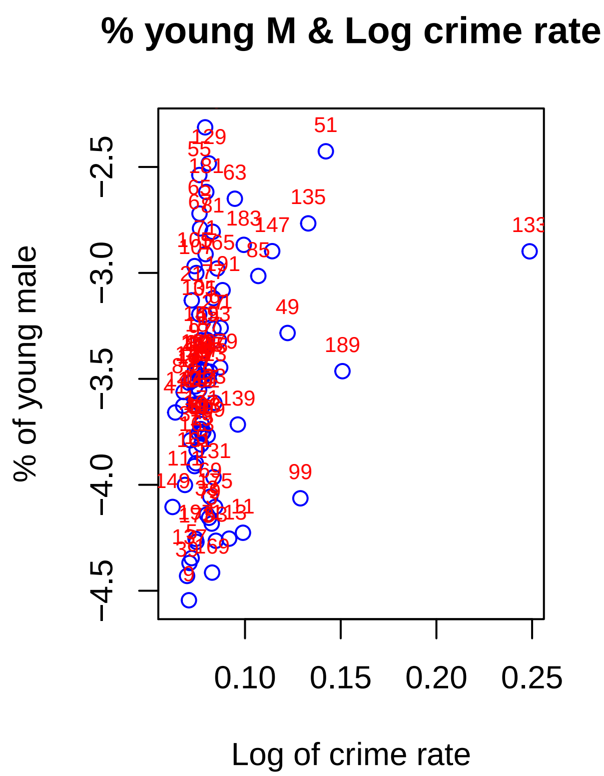
```
## [1] -0.1466527
```

Percent young male (pctymle)

The scatter plot shows that the majority of counties have 5%-10% of young male. County 133 has significantly higher male percentage than the rest of counties. Log of crime rate doesn't seem to vary by the percent of young male based on the scatter plot, which is also evidented by 0.27 correlation coefficient.

```r
par(mfrow=c(1,2))
plot(crimeData3$pctymle, crimeData3$crmrte_log,main = "% young M & Log crime rate",
     xlab="Log of crime rate",ylab="% of young male", col="blue")
plot(crimeData3$pctymle, crimeData3$crmrte_log,main = "% young M & Log crime rate",
     xlab="Log of crime rate",ylab="% of young male", col="blue")
text(crimeData2$pctymle, crimeData2$crmrte_log, labels = crimeData2$county, cex=0.7, pos=3,col = "red")
```






```r
cor(crimeData2$pctymle, crimeData2$crmrte_log)
```

```
## [1] 0.2723973
```

Model Building 1

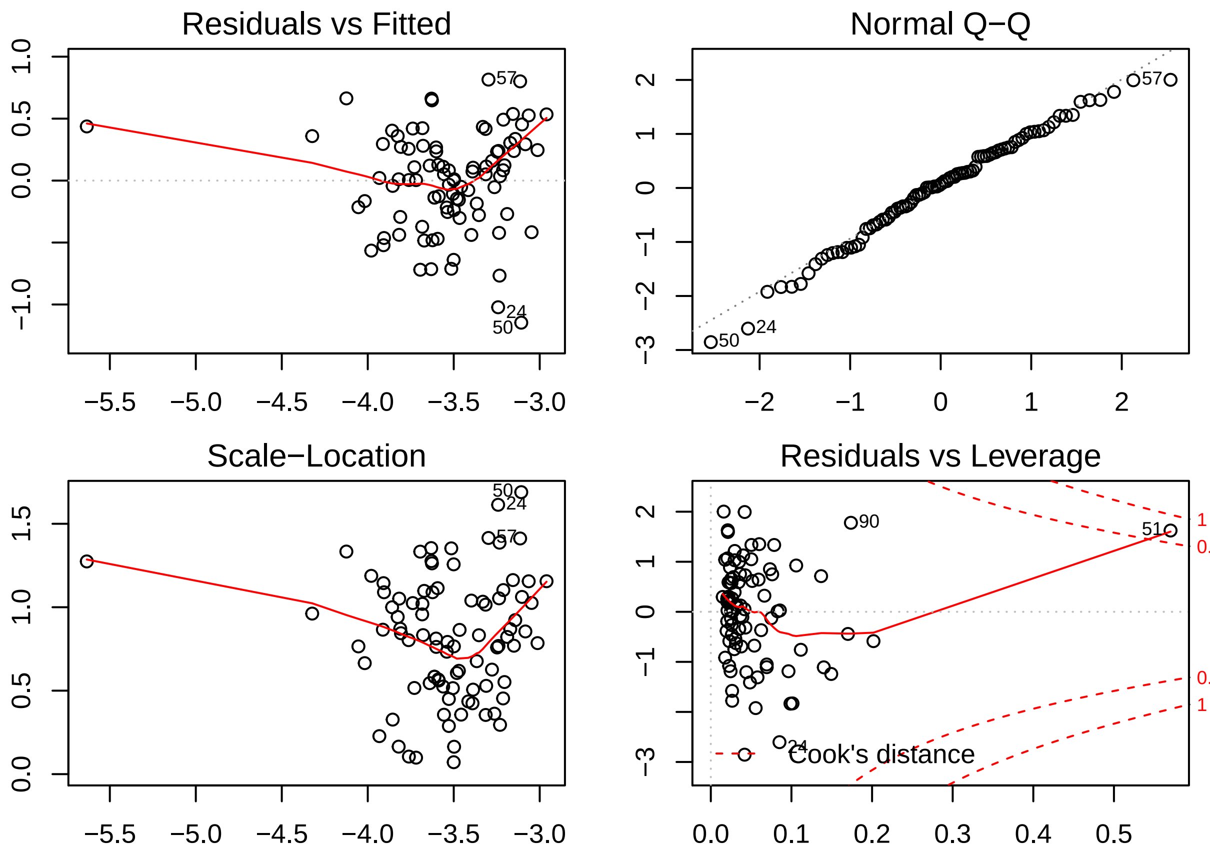
In the first model, we only include the four key variables we are interested in. They are probability of arrest, probability of conviction, probability of prison, average sentences. Based on the EDA above, we take the log of crime rate as our response variable and didn't find any transformation to be necessary for the four explanatory variables.

$$\log\left(\text{Crime Rate}\right) = \beta_0 + \beta_1 \cdot (\text{prbarr}) + \beta_2 \cdot (\text{prbconv}) + \beta_3 \cdot (\text{prbpris}) + \beta_4 \cdot (\text{avgsen})$$

```r
model1 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen, data=crimeData2)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.924260   0.441535 -6.6229 3.024e-09 ***
## prbarr      -2.097400   0.531077 -3.9493 0.0001611 ***
## prbconv     -0.796546   0.198873 -4.0053 0.0001322 ***
## prbpris      0.426283   0.652947  0.6529 0.5156084
## avgsen       0.027054   0.019351  1.3980 0.1657380
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(2,2))
par(mar=c(2, 2, 2, 2))
plot(model1)
```




```r
vif(model1)
```

```
##   prbarr  prbconv  prbpris   avgsen
## 1.039667 1.079254 1.012988 1.122232
```

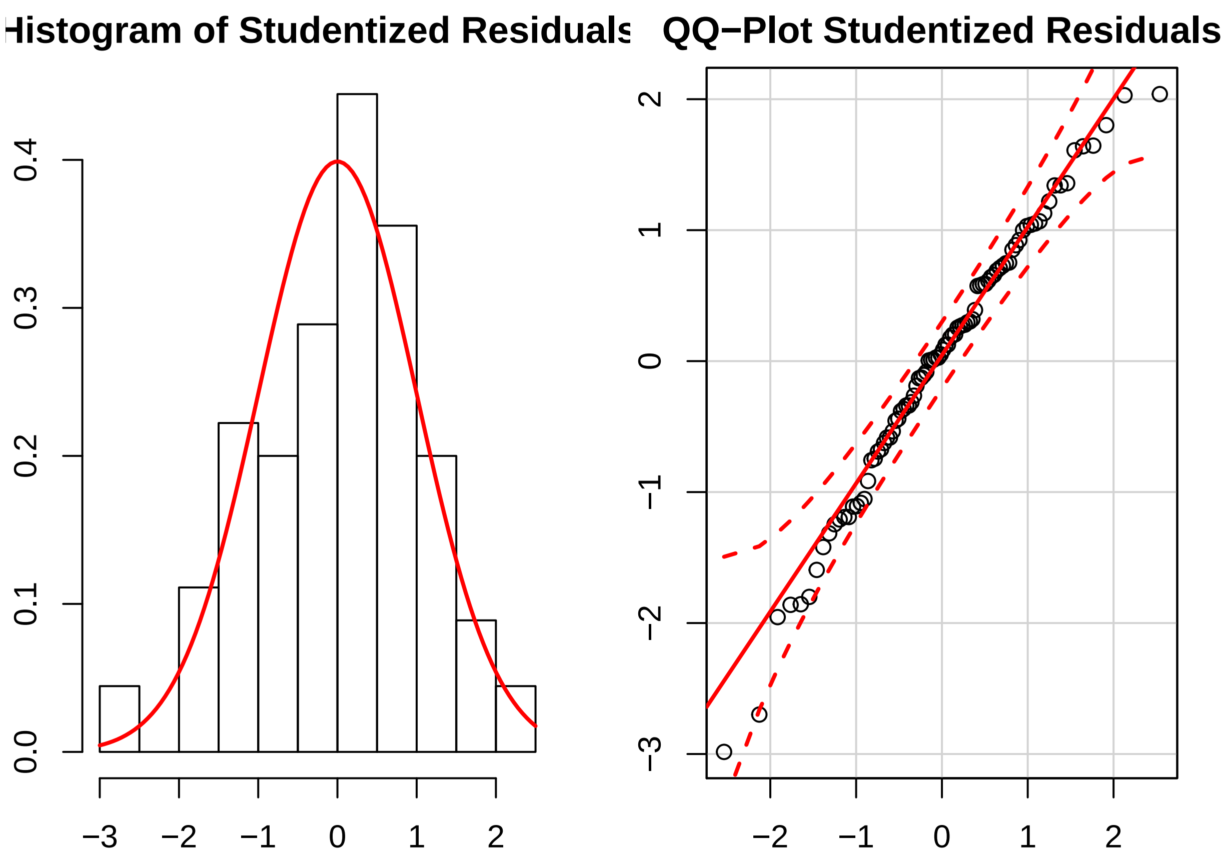
```r
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 3.6802, df = 4, p-value = 0.451
```

```r
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.139468    Df = 1     p = 0.2857648
```

```r
par(mfrow=c(1,2))
hist(rstudent(model1), main="Histogram of Studentized Residuals", breaks=10, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
qqPlot(rstudent(model1), main="QQ-Plot Studentized Residuals")
```




```r
step(model1)
```

```
## Start:  AIC=-155.45
## crmrte_log ~ prbarr + prbconv + prbpris + avgsen
##
##           Df Sum of Sq    RSS     AIC
## - prbpris  1    0.1037 14.421 -156.80
## <none>                 14.317 -155.45
## - avgsen   1    0.4637 14.781 -154.58
## - prbconv  1    5.0998 19.417 -130.03
## - prbarr   1    7.0987 21.416 -121.21
##
## Step:  AIC=-156.8
## crmrte_log ~ prbarr + prbconv + avgsen
##
##           Df Sum of Sq    RSS     AIC
## <none>                 14.421 -156.80
## - avgsen   1    0.4245 14.846 -156.19
## - prbconv  1    5.0741 19.495 -131.67
## - prbarr   1    7.0159 21.437 -123.12

##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + avgsen, data = crimeData2)
##
## Coefficients:
## (Intercept)       prbarr      prbconv       avgsen
##    -2.74283     -2.08048     -0.79439      0.02575
```

```r
AIC(model1)
```

```
## [1] 101.9591
```

- CLM 1: A linear model
  - The model is specified such that the dependent variable is a linear function of the explanatory variables.This assumption is satisfied.

- CLM 2: Random sampling
  - Each observation represents a county in North Carolina. Only a selection of counties are included in the data.Since we do not have knowledge how counties are selected, we cannot confirm the selection is random.

- CLM 3: No perfect multicollinearity
  - R would alert us if there is perfect multicollinearity among explanatory variables.

  - The VIFs for all variables in the model are all less than 2, which suggests there are no high correlation among them. This assumption is satisfied.

- CLM 4: Zero-conditional mean
  - The red spline curve in the Residuals vs. Fitted plot is mostly flat except for the end points where the number of observations is small.

  - In the Residuals vs. Leverage plot, no observations have Cook's distance greater than 1.This assumption is satistified.

- CLM 5: Homoscedasticity
  - In both Residuals vs. Fitted plot and Scale-Location plot, the variance of residuals is curve.
  - Breusch-Pagan test shows p-value is 0.451 implies homoscedascity and we fail to reject the null hypothesis.This assumption is likely satisfied.

- CLM 6: Normality of residuals
  - The histogram of studentized residuals is fairly normal ditributed albeit a bit light in the right tail.

  - Q-Q plot shows most data points in the right tail are below the diagonal line, which also confirms a light right tail. Overall, Q-Q plot doesn't deviate significantly from normality and most data points are within the confidence interval. This assumption is satisfied.

Evaluation of statistical and pratical significance

Model 3 shows:

- A 1 percent point increase in the probability of arrest decreases crime rate by 1.77 %, given that all other variables remain constant.

- A 1 percent point increase in the probability of conviction decreases crime rate by 0.64 %, given that all other variables remain constant.
  Adding the these control variables improves the model. AIC decreases from 102 (Model 1) to 45 (Model 2) to 31 (Model 3)

Cook's distance for Observation 51, which is County 115, is close to 1.

Model Building 2

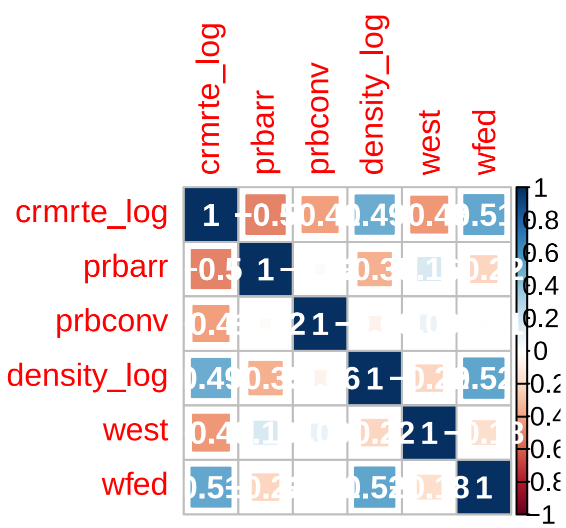
Based on EDA, we concludes that the following variables are highly correlated with the response variable, with correlation coefficient greater than 0.3:
prbarr, prbconv, density_log, west, and wfed.
The variable "urban" and a few other weekly wages variables are excluded from the list above due to their higher correlation with other explanatory variables.
The pairwise correlation table shows that log of density is highly correlated with federal employee wages. Since the proportion of federal employees in a county is small, this variable may not represent the wage characteristics of a county. So we remove federal employee wages from the model.

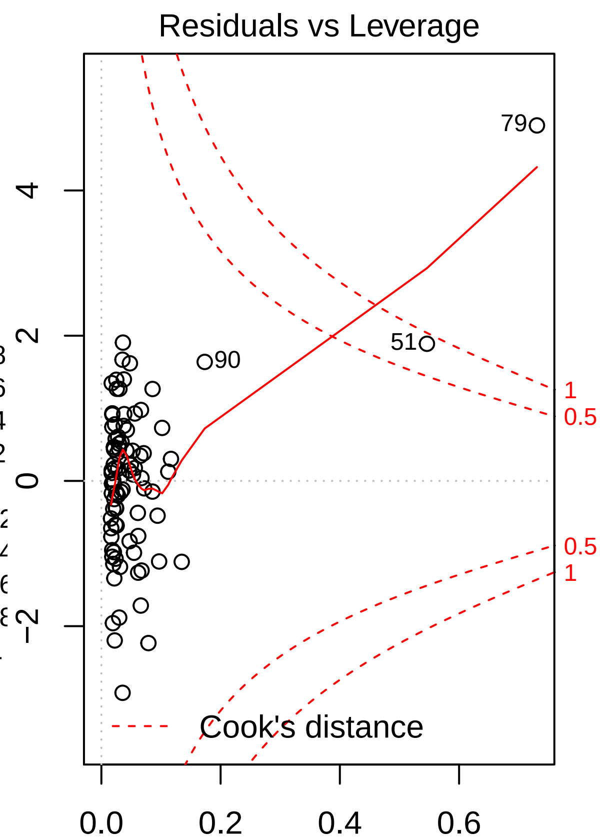
$$\log(\text{Crime Rate}) = \beta_0 + \beta_1 \cdot (\text{prbarr}) + \beta_2 \cdot (\text{prbconv}) + \beta_3 \cdot (\log \text{density}) + \beta_4 \cdot (\text{west})$$

Since the 79th observation (County 173) has Cook's distance greater than 1, we removed this observation from the model to eliminate undue influence.

```
par(mfrow=c(1,2))
par(mar=c(2, 2, 2, 2))
crimeData_temp2 <- crimeData2[,c("crmrte_log","prbarr","prbconv", "density_log", "west", "wfed")]
corr_Model2 <- cor(crimeData_temp2, use="pairwise")
corrplot(corr_Model2, method="square", addCoef.col="white")

model2 <- lm(crmrte_log ~ prbarr+prbconv+density_log+west, data=crimeData2)
plot(model2, which=5)
```



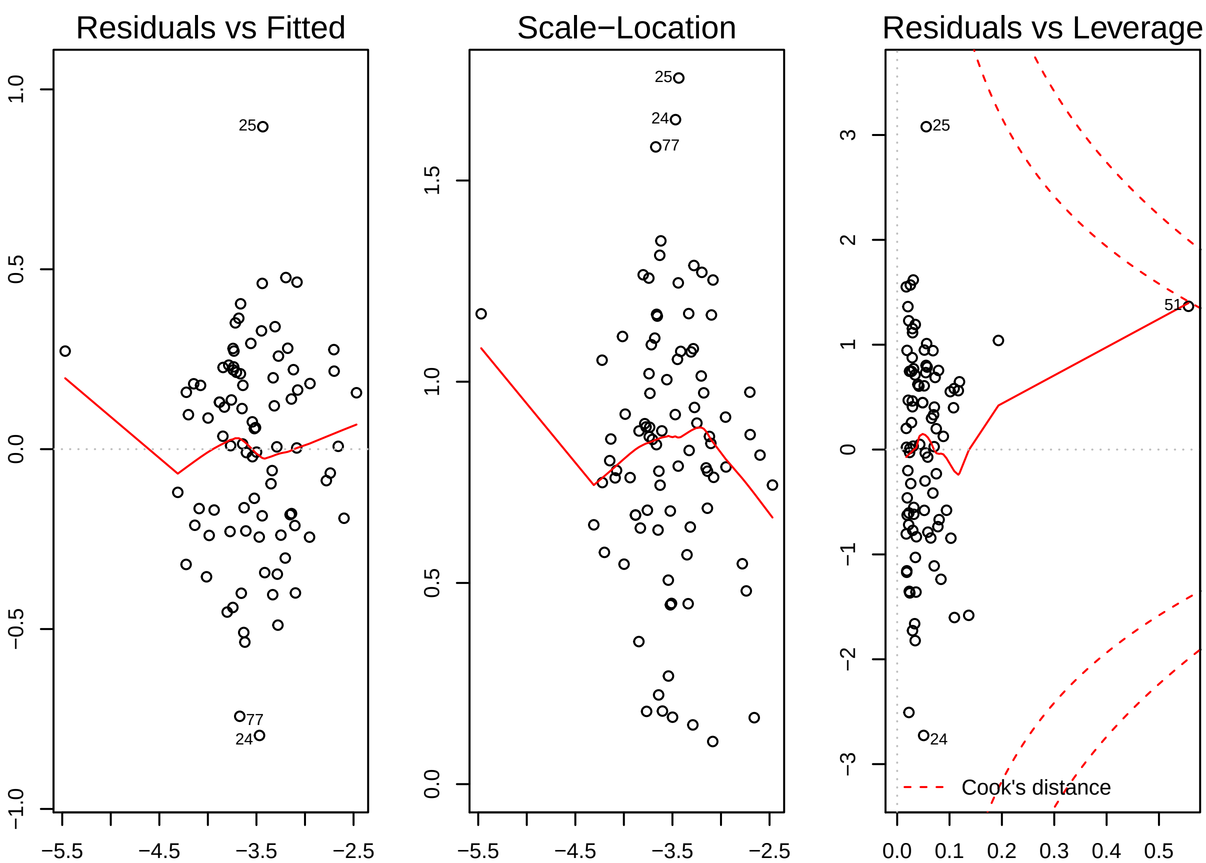



```
crimeData5 <- crimeData2[which(crimeData2$county!=173),]
```

```r
model2 <- lm(crmrte_log ~ prbarr+prbconv+density_log+west, data=crimeData5)
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -2.865305   0.176534 -16.2309 < 2.2e-16 ***
## prbarr      -1.062075   0.392088  -2.7088 0.0081828 **
## prbconv     -0.508004   0.145606  -3.4889 0.0007748 ***
## density_log  0.335447   0.051834   6.4716 6.154e-09 ***
## west        -0.364090   0.063973  -5.6913 1.800e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(1,3))
par(mar=c(2, 2, 2, 2))
plot(model2, which=1)
plot(model2, which=3)
plot(model2, which=5)
```




```r
vif(model2)
```

```
##      prbarr     prbconv density_log        west
##    1.187549    1.070095    1.238060    1.033034
```

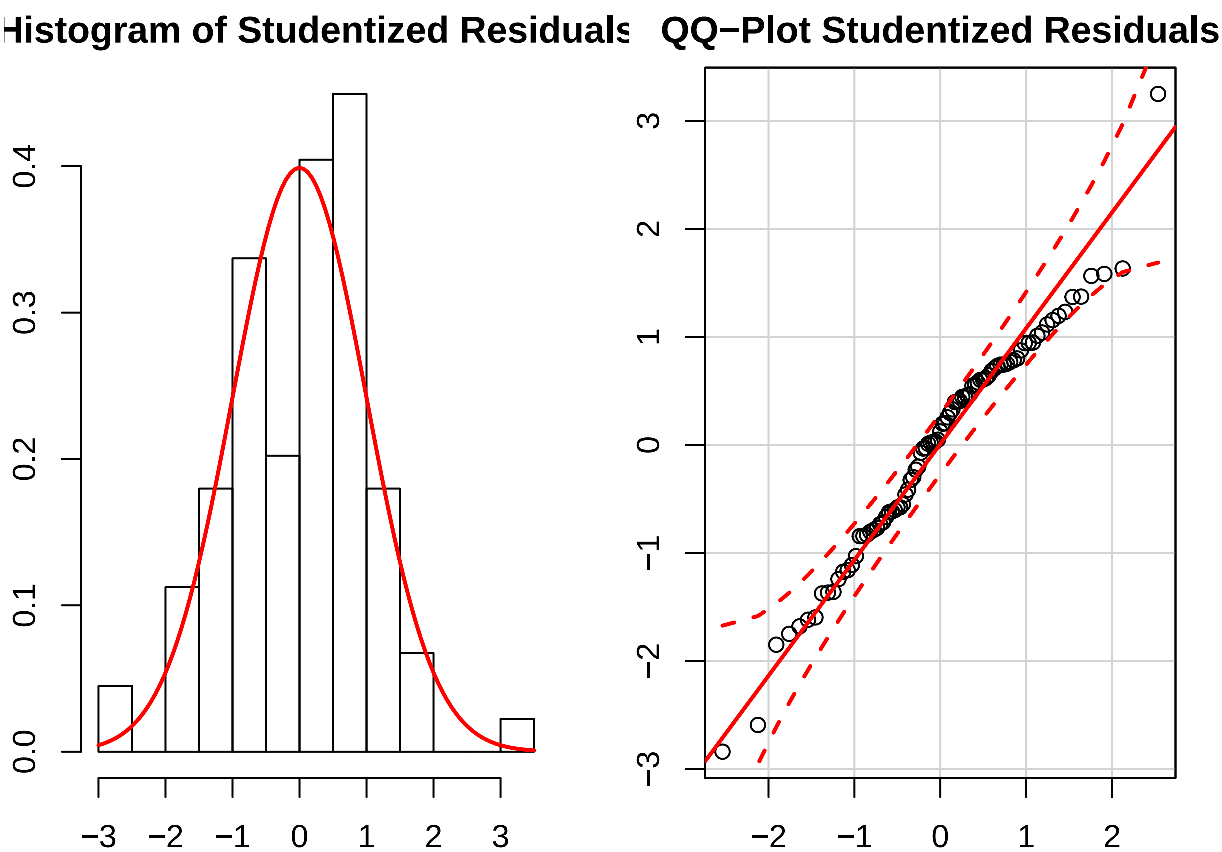
```r
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 9.6299, df = 4, p-value = 0.04714
```

```r
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.006666746    Df = 1     p = 0.9349249
```

```r
par(mfrow=c(1,2))
hist(rstudent(model2), main="Histogram of Studentized Residuals", breaks=10, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
qqPlot(rstudent(model2), main="QQ-Plot Studentized Residuals")
```




```r
AIC(model2)
```

```
## [1] 44.8252
```

Assessment of the CLM assumptions

- CLM 1: A linear model
  - The model is specified such that the dependent variable is a linear function of the explanatory variables.

- This assumption is satisfied.

- CLM 2: Random sampling
  - Each observation represents a county in North Carolina. Only a selection of counties are included in the data. Since we do not have knowledge how counties are selected, we cannot confirm the selection is random.

- CLM 3: No perfect multicollinearity
  - R would alert us if there is perfect multicollinearity among explanatory variables.

  - The VIFs for the four variables in the model are all less than 2, which suggests there are no high correlation among them.
  - This assumption is satisfied.

- CLM 4: Zero-conditional mean
  - The red spline curve in the Residuals vs. Fitted plot is mostly flat except for the end points where the number of observations is small.

  - In the Residuals vs. Leverage plot, no observations have Cook's distance greater than 1.

  - This assumption is satistified.

- CLM 5: Homoscedasticity
  - In both Residuals vs. Fitted plot and Scale-Location plot, the variance of residuals slightly increase then decrease when fitted values are between -4.5 and -2.5.

  - Breusch-Pagan test shows significant p-value while the Score-test shows insignificant p-value. These tests are producing mixed evidence of homoscedasticity.

  - This assumption is likely satisfied.

  - If this assumption is not satisfied, the usual formulas for standard errors are inaccurate. Heteroskedasticity-robust standard errors should be used to test the significance of the parameter estimates.

- CLM 6: Normality of residuals
  - The histogram of studentized residuals is fairly normal ditributed albeit a bit light in the right tail.

  - Q-Q plot shows most data points in the right tail are below the diagonal line, which also confirms a light right tail. Overall, Q-Q plot doesn't deviate significantly from normality and all data points are within the confidence interval.

  - This assumption is satisfied.

Evaluation of statistical and pratical significance

All four explanatory variables are statistically significant.
Model 2 shows:

- A 1 percent point increase in the probability of arrest decreases crime rate by 1.062%, given that all other variables remain constant.

- A 1 percent point increase in the probability of conviction decreases crime rate by 0.58%, given that all other variables remain constant.

- Density is a geographic control variable. A 1% increase in the density increases crime rate by 0.335%, given that all other variables remain constant.

- Indicator "west" is also a geographic control variable. Western counties have 36.4% less crime rate, given that all other variables remain constant.

All the parameter estimates are practically significant.

Adding the two control variables improves the model. AIC decreases from 102 (Model 1) to 45 (Model 2).

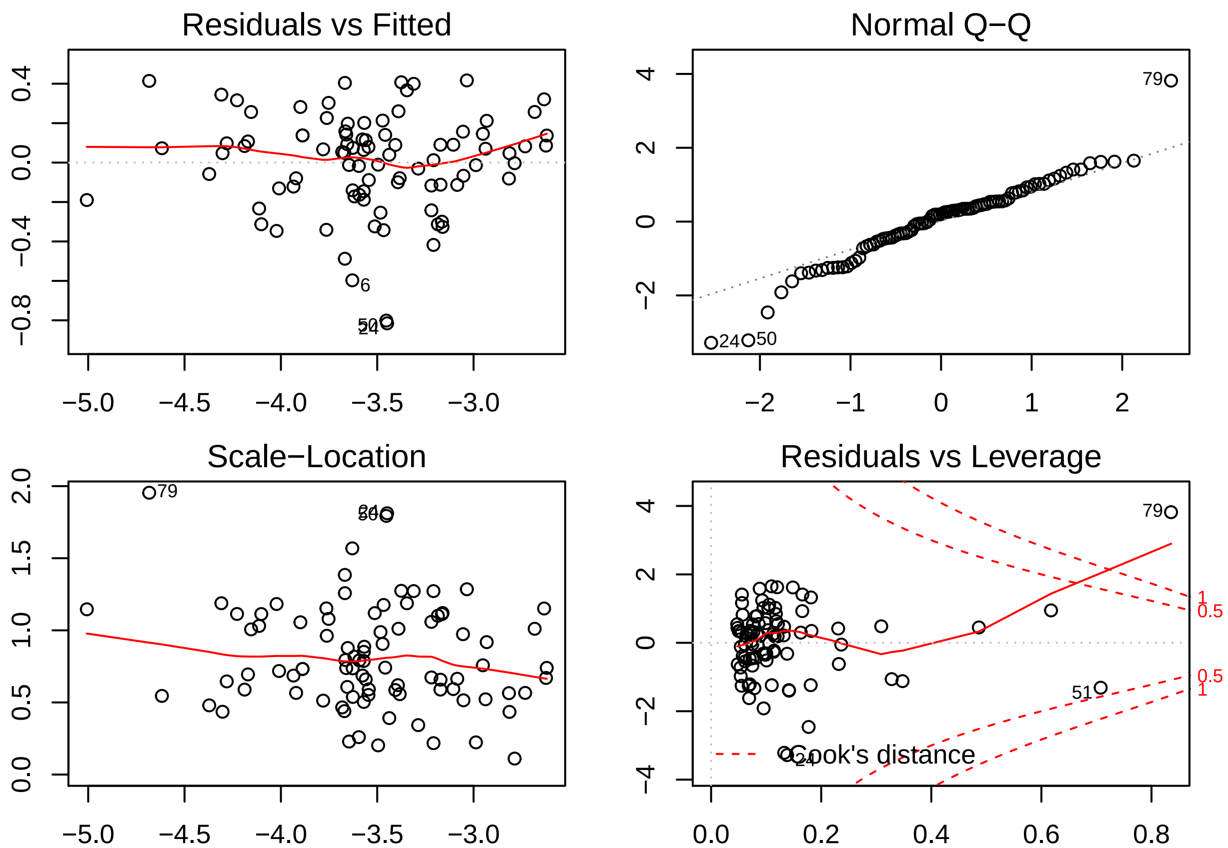
Model Building 3

For this model, we include all the variables that are not highly correlated.

```
model3 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen+polpc_log+density_log+taxpc
             +west+central+mix+pctymle, data=crimeData2)
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.7011891  2.0779724  1.2999 0.1974582
## prbarr      -1.7790882  0.5325906 -3.3404 0.0012855 **
## prbconv     -0.6410498  0.1731862 -3.7015 0.0003977 ***
## prbpris     -0.3964859  0.5802484 -0.6833 0.4964397
## avgsen      -0.0281447  0.0144016 -1.9543 0.0542534 .
## polpc_log    0.7216939  0.2585875  2.7909 0.0066053 **
## density_log  0.1239070  0.1737687  0.7131 0.4779382
## taxpc       -0.0017524  0.0032767 -0.5348 0.5943056
## west        -0.4514779  0.0934789 -4.8297 6.689e-06 ***
## central     -0.1796883  0.1113237 -1.6141 0.1105428
## mix          0.1168549  0.6355153  0.1839 0.8545892
## pctymle     -0.3116119  1.7470917 -0.1784 0.8589024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
par(mar=c(2, 2, 2, 2))
plot(model3)
```




```r
vif(model3)
```

```
##      prbarr     prbconv     prbpris      avgsen   polpc_log density_log
##    1.563087    1.391537    1.305307    1.506613    1.859822    1.902031
##       taxpc        west     central         mix     pctymle
##    1.608725    1.370656    1.424896    1.790530    1.303687
```

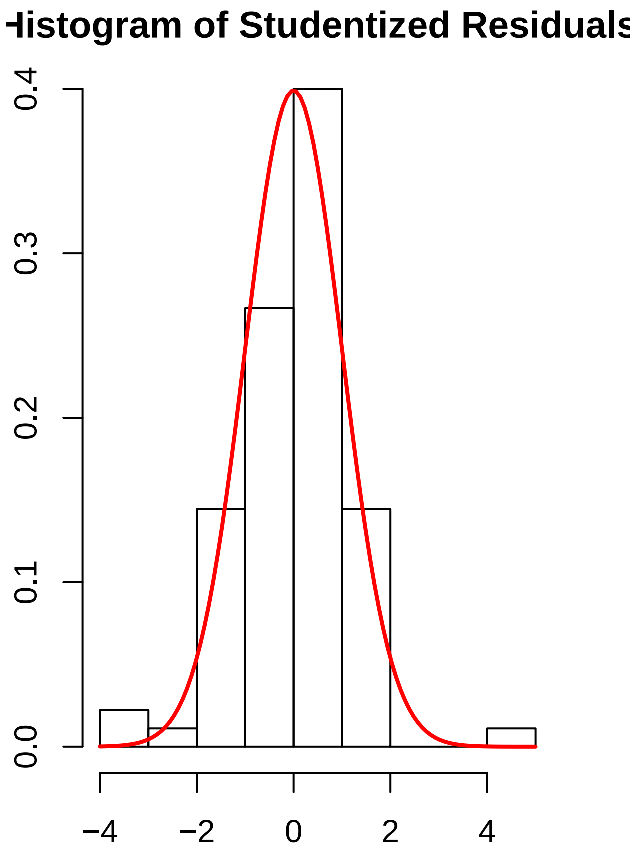
```r
bptest(model3)
```

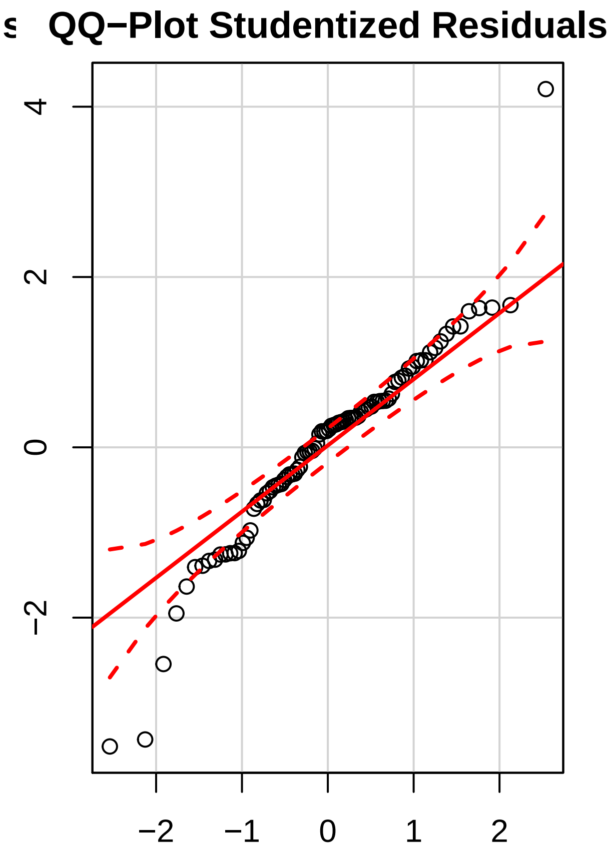
```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 14.241, df = 11, p-value = 0.22
```

```r
ncvTest(model3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.06838443    Df = 1     p = 0.7937037
```

```r
par(mfrow=c(1,2))
hist(rstudent(model3), main="Histogram of Studentized Residuals", breaks=10, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
qqPlot(rstudent(model3), main="QQ-Plot Studentized Residuals")
```






```r
step(model3)
```

```
## Start:  AIC=-226.21
## crmrte_log ~ prbarr + prbconv + prbpris + avgsen + polpc_log +
##     density_log + taxpc + west + central + mix + pctymle
##
##                Df Sum of Sq    RSS     AIC
## - pctymle       1    0.0036 5.5862 -228.16
## - mix           1    0.0045 5.5871 -228.14
## - taxpc         1    0.0294 5.6120 -227.74
## - prbpris       1    0.0697 5.6523 -227.10
## <none>                      5.5826 -226.21
## - avgsen        1    0.3738 5.9564 -222.38
## - central       1    0.4736 6.0562 -220.89
## - density_log   1    1.3493 6.9319 -208.73
## - west          1    2.5463 8.1289 -194.40
## - prbconv       1    2.5618 8.1444 -194.22
## - prbarr        1    3.3972 8.9798 -185.44
## - polpc_log     1    3.4962 9.0788 -184.45
##
## Step:  AIC=-228.16
## crmrte_log ~ prbarr + prbconv + prbpris + avgsen + polpc_log +
##     density_log + taxpc + west + central + mix
##
##                Df Sum of Sq    RSS     AIC
## - mix           1    0.0048 5.5911 -230.08
## - taxpc         1    0.0258 5.6120 -229.74
```

```
## - prbpris      1    0.0671 5.6534 -229.08
## <none>                    5.5862 -228.16
## - avgsen       1    0.3738 5.9600 -224.33
## - central      1    0.4819 6.0681 -222.71
## - density_log  1    1.3481 6.9343 -210.70
## - west         1    2.5829 8.1692 -195.95
## - prbconv      1    2.6716 8.2578 -194.98
## - prbarr       1    3.5553 9.1415 -185.83
## - polpc_log    1    3.8006 9.3868 -183.44
##
## Step:  AIC=-230.08
## crmrte_log ~ prbarr + prbconv + prbpris + avgsen + polpc_log +
##     density_log + taxpc + west + central
##
##               Df Sum of Sq    RSS     AIC
## - taxpc        1    0.0262 5.6173 -231.66
## - prbpris      1    0.0623 5.6534 -231.08
## <none>                    5.5911 -230.08
## - avgsen       1    0.3804 5.9715 -226.15
## - central      1    0.4870 6.0780 -224.56
## - density_log  1    1.6081 7.1992 -209.33
## - west         1    2.6927 8.2838 -196.70
## - prbconv      1    3.1200 8.7111 -192.17
## - prbarr       1    3.7492 9.3403 -185.89
## - polpc_log    1    3.8200 9.4111 -185.21
##
## Step:  AIC=-231.66
## crmrte_log ~ prbarr + prbconv + prbpris + avgsen + polpc_log +
##     density_log + west + central
##
##               Df Sum of Sq     RSS     AIC
## - prbpris      1    0.0544  5.6717 -232.79
## <none>                      5.6173 -231.66
## - avgsen       1    0.3703  5.9876 -227.91
## - central      1    0.4660  6.0833 -226.48
## - density_log  1    1.5936  7.2109 -211.18
## - west         1    2.7266  8.3439 -198.04
## - prbconv      1    3.1145  8.7318 -193.96
## - prbarr       1    3.7771  9.3944 -187.37
## - polpc_log    1    4.5125 10.1298 -180.59
##
## Step:  AIC=-232.79
## crmrte_log ~ prbarr + prbconv + avgsen + polpc_log + density_log +
##     west + central
##
##               Df Sum of Sq     RSS     AIC
## <none>                      5.6717 -232.79
## - avgsen       1    0.3319  6.0036 -229.67
## - central      1    0.4929  6.1646 -227.29
## - density_log  1    1.5824  7.2541 -212.64
## - west         1    2.7667  8.4384 -199.03
## - prbconv      1    3.1990  8.8708 -194.53
## - prbarr       1    4.1422  9.8139 -185.44
## - polpc_log    1    4.4825 10.1542 -182.37
```

```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + avgsen + polpc_log +
##     density_log + west + central, data = crimeData2)
##
## Coefficients:
## (Intercept)       prbarr      prbconv       avgsen    polpc_log
##      2.1939      -1.7607      -0.6409      -0.0260       0.6851
## density_log         west      central
##      0.1118      -0.4438      -0.1757
```

```
AIC(model3)
```

```
## [1] 31.195
```

```
fit <- aov(prbarr ~ taxpc, data=crimeData2)
fit <- aov(prbarr ~ taxpc, data=crimeData2)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## taxpc        1  0.030 0.03000   1.602  0.209
## Residuals   88  1.648 0.01872
```

```
cohen.d(crimeData2$prbconv , crimeData2$prbarr, pooled=TRUE, paired=FALSE, na.rm=FALSE,
    hedges.correction = FALSE, conf.level = 0.95, noncentral = FALSE)
```

```
##
## Cohen's d
##
## d estimate: 0.9860997 (large)
## 95 percent confidence interval:
##       inf       sup
## 0.6745597 1.2976397
```

- CLM 1: A linear model
  - The model is specified such that the dependent variable is a linear function of the explanatory variables. This assumption is satisfied.

- CLM 2: Random sampling
  - Each observation represents a county in North Carolina. Only a selection of counties are included in the data.Since we do not have knowledge how counties are selected, we cannot confirm the selection is random.

- CLM 3: No perfect multicollinearity
  - R would alert us if there is perfect multicollinearity among explanatory variables.

  - The VIFs for all variables in the model are all less than 2, which suggests there are no high correlation among them. This assumption is satisfied.

- CLM 4: Zero-conditional mean
  - The red spline curve in the Residuals vs. Fitted plot is mostly flat except for the end points where the number of observations is small.

  - In the Residuals vs. Leverage plot, no observations have Cook's distance greater than 1. This assumption is satistified.

- CLM 5: Homoscedasticity

- In both Residuals vs. Fitted plot and Scale-Location plot, the variance of residuals remain flat with minor decrease and eventually increasing. Most points are when fitted values are between -5.0 and -2.5.

- Breusch-Pagan test shows p-value is 0.22 implies homoscedascity and we fail to reject the null hypothesis. These tests are producing mixed evidence of homoscedasticity. This assumption is likely satisfied.

- CLM 6: Normality of residuals
  - The histogram of studentized residuals is fairly normal ditributed albeit a bit light in the right tail.

  - Q-Q plot shows most data points in the right tail are below the diagonal line, which also confirms a light right tail. Overall, Q-Q plot doesn't deviate significantly from normality and most data points are within the confidence interval. This assumption is satisfied.

Evaluation of statistical and pratical significance

Model 3 shows:

- A 1 percent point increase in the probability of arrest decreases crime rate by 1.77 %, given that all other variables remain constant.

- A 1 percent point increase in the probability of conviction decreases crime rate by 0.64 %, given that all other variables remain constant.
  Adding the these control variables improves the model. AIC decreases from 102 (Model 1) to 45 (Model 2) to 31 (Model 3)

Model Display

```r
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))
stargazer(model1, model2, model3, type = "latex",
          title = "Linear Models Predicting Log of Crime Rate",
          omit.stat="f",
          se=list(se.model1, se.model2, se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          float=FALSE)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 15, 2018 - 20:09:00

	Dependent variable:		
	crmrte_log		
	(1)	(2)	(3)
prbarr	−2.097***	−1.062**	−1.779***
	(0.531)	(0.392)	(0.533)
prbconv	−0.797***	−0.508***	−0.641***
	(0.199)	(0.146)	(0.173)
prbpris	0.426		−0.396
	(0.653)		(0.580)
avgsen	0.027		−0.028
	(0.019)		(0.014)
polpc_log			0.722**
			(0.259)
density_log		0.335***	0.124
		(0.052)	(0.174)
taxpc			−0.002
			(0.003)
west		−0.364***	−0.451***
		(0.064)	(0.093)
central			−0.180
			(0.111)
mix			0.117
			(0.636)
pctymle			−0.312
			(1.747)
Constant	−2.924***	−2.865***	2.701
	(0.442)	(0.177)	(2.078)
Observations	90	89	90
R^2	0.447	0.702	0.784
Adjusted R^2	0.421	0.688	0.754
Residual Std. Error	0.410 (df = 85)	0.300 (df = 84)	0.268 (df = 78)
Note:			*p<0.05; **p<0.01; ***p<0.001

The parameter estimates for the two key variables of interest are relatively robust.

- In all three models, the parameter estimate for the probability of arrest are negative and statistically significant. The parameter estimates range from -2.10 to -1.06. The difference is mostly caused by the geographic control variables in the model.

- In all three models, the parameter estimate for the probability of conviction are negative and statistically significant. The parameter estimates range from -0.80 to -0.51. The difference is mostly caused by the geographic control variables in the model.

Omitted Variables

We are interested in the relationship between prbarr and prbconv and our "omitted" variables mentioned below. 1.) Education 2.) Percentage of people using drug 3.) Percentage of people are married with kids 4.) Percentage of people who own guns 5.) Percentage of smart mobile penetration

Suppose our estimation model is represented by below equation , where A is prbarr and B is prbconv. $\beta_0, \beta_1, \beta_2$ is the intercepts and coefficients of A (prbarr) and B (prbconv) respectively. We can represent the Population regression as

$$Y(Population) = \beta_0 + \beta_1 A + \beta_2 B$$

$$Y(Estimated) = \alpha_0 + \alpha_1 A + \alpha_2 B$$

$$Y(EstModel1) = -2.92426013 - 2.09740014 * prbarr - 0.79654556 * prbconv$$

Omitted variable 1: Education

Crime rate and Education are negatively correlated (NEGATIVE).

We estimate a negative correlation between Crime Rate and Education based on following logic and reasoning. We estimate that higher the education within the population then they are less likely to commit a crime. Another reason could be that educated people tend to understand the law better and are more likely to respect the law and less likely to commit crime.

prbarr and Education are negatively correlated (NEGATIVE)

We estimate that educated people are more likely to be aware of their rights and they are more likely to avoid being wrongfully arrested , hence reduced arrest rates. Also police are less likely to believe an educated person committed the crime. Finally more educated people are more familiar with investigation methods and hence are less likely to leave less clues if they had committed the crime , leading to less likelihood of arrests. Hence we estimate that education and prbarr are negatively correlated.

$$\alpha_1 < 0$$

this is positive bias , alpha1 (estimated) = -2.1 = beta1 (true) + positive

$$\alpha(estimated) = -2.1 = \beta_1(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in statistical significance.

prbconv and Education are negatively correlated (NEGATIVE).

We estimate that higher education people are likely to have have higher income.People with higher income are more likely to afford better lawyers and legal help and are less likely to be convicted. Overall positive bias.

$$\alpha_1 < 0$$

This is positive bias , alpha (estimated) = -0.8 = beta (true) + positive

$$\alpha(estimated) = -0.8 = \beta(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in statistical significance.

Omitted variable 2: % of people using drug

Crime rate and % of people using drug are positively correlated (POSITIVE).

We estimate a positive correlation between Crime Rate and % of people using drug based on following logic and reasoning. We estimate that people under the influence of drugs are more likely to commit a crime.

prbarr and % of people using drug are positively correlated (POSITIVE).

We estimate that drug addicts are easier to be found if they committed the crime leading to likelihood of higher arrests. Another reason could be that police are more likely to believe that drug addicts committed the crime.Finally the people consuming drugs are likely to consume drugs in groups and when police perform arrests they are likely to find other drug consuming people leading to higher arrests. Hence we estimate that % of people using drugs and prbarr are positively correlated.

$$\alpha_1 < 0$$

this is positive bias , alpha1 (estimated) = -2.1 = beta1 (true) + positive

$$\alpha(estimated) = -2.1 = \beta_1(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in STATISTICAL significance.

prbconv and % of people using drug are positively correlated (POSITIVE).

We estimate that % of people using drugs are likely to be spend all their money on drugs leading to poverty. People in poverty are less likely to get best legal help and are likely to have higher convition rates. Overall positive bias.

$$\alpha_1 < 0$$

This is positive bias , alpha (estimated) = -0.8 = beta (true) + positive

$$\alpha(estimated) = -0.8 = \beta(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in STATISTICAL significance.

Omitted variable 3: % of people are married with kids

Crime rate and % of people are married with kids are negatively correlated (NEGATIVE).

We estimate a negative correlation between Crime Rate and % of people are married with kids based on following logic and reasoning. We estimate that married couples with kids to have more considerations for family and thus less likely to commit a crime.

prbarr and % of people are married with kids are negatively correlated (NEGATIVE).

We estimate that people married with kids are unlikely to come forward as a witness of a crime (in worry of wellbeing of their family) with any information leading to reduced probability of arrest.

$$\alpha_1 < 0$$

this is positive bias , alpha1 (estimated) = -2.1 = beta1 (true) + positive

$$\alpha(estimated) = -2.1 = \beta_1(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in STATISTICAL significance.

prbconv and % of people are married with kids are negatively correlated (NEGATIVE).

We estimate that % of people are married with kids are expected to have less convitions since the judges are likely to keep in consideration impact on the family and kids of the conviction.

$$\alpha_1 < 0$$

This is positive bias , alpha (estimated) = -0.8 = beta (true) + positive

$$\alpha(estimated) = -0.8 = \beta(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in STATISTICAL significance.

Omitted variable 4: % of people who own guns

Crime rate and % of people who own guns are positively correlated (POSITIVE).

We estimate a positive correlation between Crime Rate and % of people who own guns based on following logic and reasoning. With more people owning guns , small alterations / conflicts can lead to gun fight leading to higher rates of crime.

prbarr and % of people who own guns are positively correlated (POSITIVE).

We estimate that people who are gun owners are more likely to involved in shooting related crimes. We estimate that since the guns and gun owners are more likley to easily tracked this can lead to higher probability of arrests.

$$\alpha_1 < 0$$

this is positive bias , alpha1 (estimated) = -2.1 = beta1 (true) + positive

$$\alpha(estimated) = -2.1 = \beta_1(true) + (positive)$$

True Coefficent is even more negative what is estimated and hence increase in STATISTICAL significance.

prbconv and % of people who own guns are negatively correlated (NEGATIVE).

We estimate that % of people who own guns are expected to richer than others and hence more likely to afford better lawyers and legal help. this can lead to less conviction rates.

$$\alpha_1 < 0$$

This is positive bias , alpha (estimated) = -0.8 = beta (true) + negative

$$\alpha(estimated) = -0.8 = \beta(true) + (negative)$$

True Coefficent is less negative what is estimated and hence loose STATISTICAL significance.

Omitted variable 5: % of smart mobile penetration

Crime rate and % of smart mobile penetration are positively correlated (POSITIVE).

We estimate a positive correlation between Crime Rate and % of smart mobile penetration based on following logic and reasoning. With higher smart phone users are more likely to use better communication methods via encrypted apps for committing the crime , hence increasing the likelihood of crime rates. Another possible reason can be that smart phones themselves are expensive devices and with higher smart phones in a county can lead to higher theft cases of smart phones itself , hence increasing crime rates.

prbarr and % of smart mobile penetration are negatively correlated (NEGATIVE).

We estimate that people using smart phone are more likely to use encrypted apps on the smart phones for communication for committing the crime. Since these encrypted application are extremely difficult to track and hence leading to reduced arrest rates.

$$\alpha_1 < 0$$

this is less negative bias , alpha1 (estimated) = -2.1 = beta1 (true) + negative

$$\alpha(estimated) = -2.1 = \beta_1(true) + (negative)$$

prbarr has less impact on the log of crime rate (lose statistical significance)

prbconv and % of smart mobile penetration are negatively correlated (NEGATIVE).

We estimate that smart phone has higher protection (such as iPhone), which are harder to crack by police , hence reducing the conviction rates. Also we estimate that not all police departments are good at dealing with digital evidence and hence leading to reduced conviction rates.

$$\alpha_1 < 0$$

this is less negative bias , alpha (estimated) = -0.8 = beta (true) + negative

$$\alpha(estimated) = -0.8 = \beta(true) + (negative)$$

prbconv has less impact on the log of crime rate (lose statistical significance).

Conclusion

Our models show that an increase in the probability of arrest and/or an increase in the proability of conviction reduces the crime rate. On the other hand, there are no statistically significance relationships between the severity of punishment (the probability of prison and average sentences) and crime rate. This conclusion is robust and is not sensitive to modeling specifications. We believe the key to reduce crime rates is to be effective in apprehending criminals. We recommend community service as part of the punishment to help rehabilitate criminals rather than increasing their sentences.