

Lab 3 (Bhuvnesh Sharma, Weixin Wu)

Bhuvnesh Sharma, Weixin Wu

March 22, 2018

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

```
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      avgsgen      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      wmfgr
## 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.

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

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

```
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.

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

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

```
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.

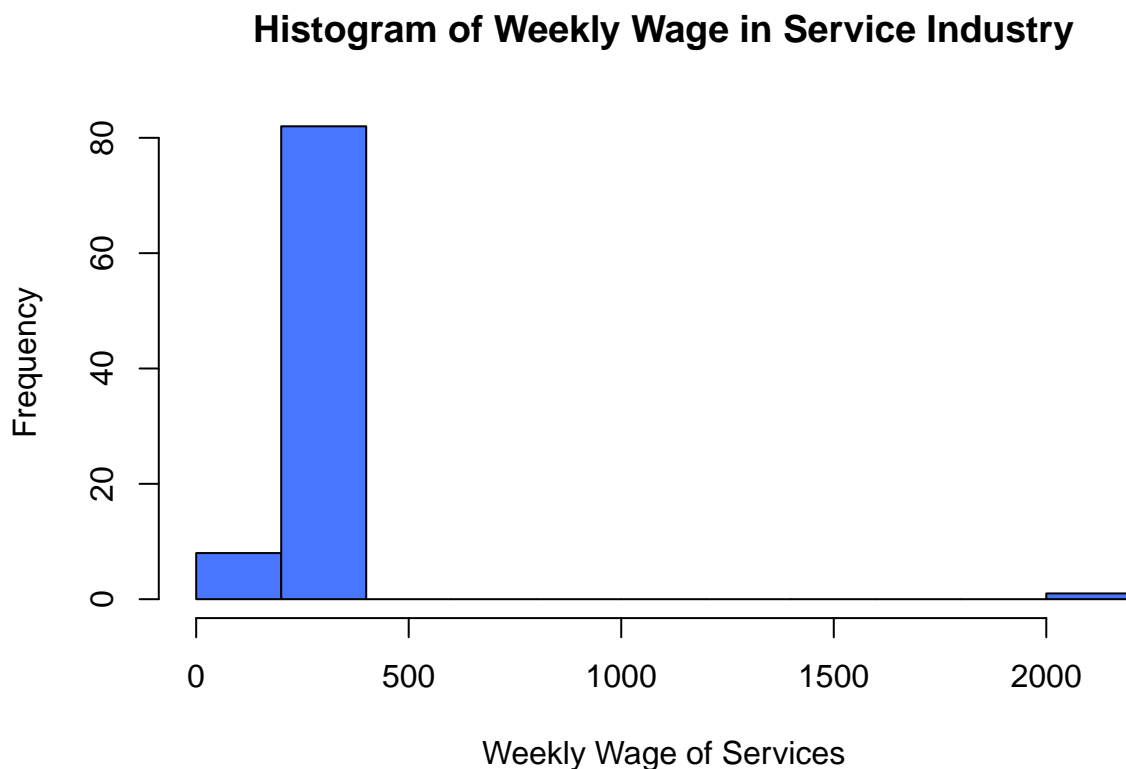
Without further information on the variables, we could not conclude whether these values are invalid. So we left those observations in the data.

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).

```
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.

```
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 construction, 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.

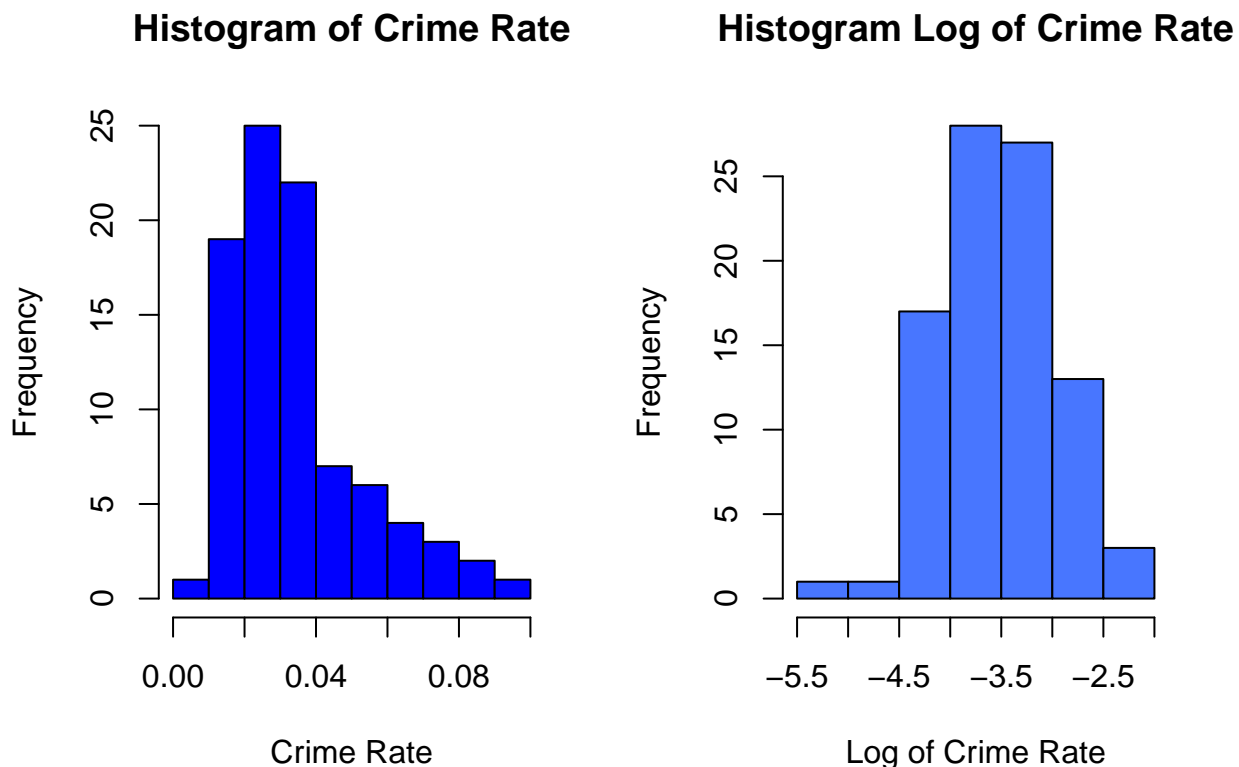
```
crimeData2 <- crimeData2[which(crimeData2$wuser<2000),]
```

Exploratory Data Analysis

Crimes committed per person (crrmrte)

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 crrmrte_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 explanatory variables. Our target variable is crrmrte_log.

```
par(mfrow=c(1,2))
hist(crimeData2$crrmrte, main="Histogram of Crime Rate",xlab = "Crime Rate" , col = "blue")
crimeData2$crrmrte_log = log(crimeData2$crrmrte)
hist(crimeData2$crrmrte_log, main="Histogram Log of Crime Rate",
     xlab = "Log of Crime Rate" ,
     col = "royalblue1")
```



Probability of arrest (prbarr)

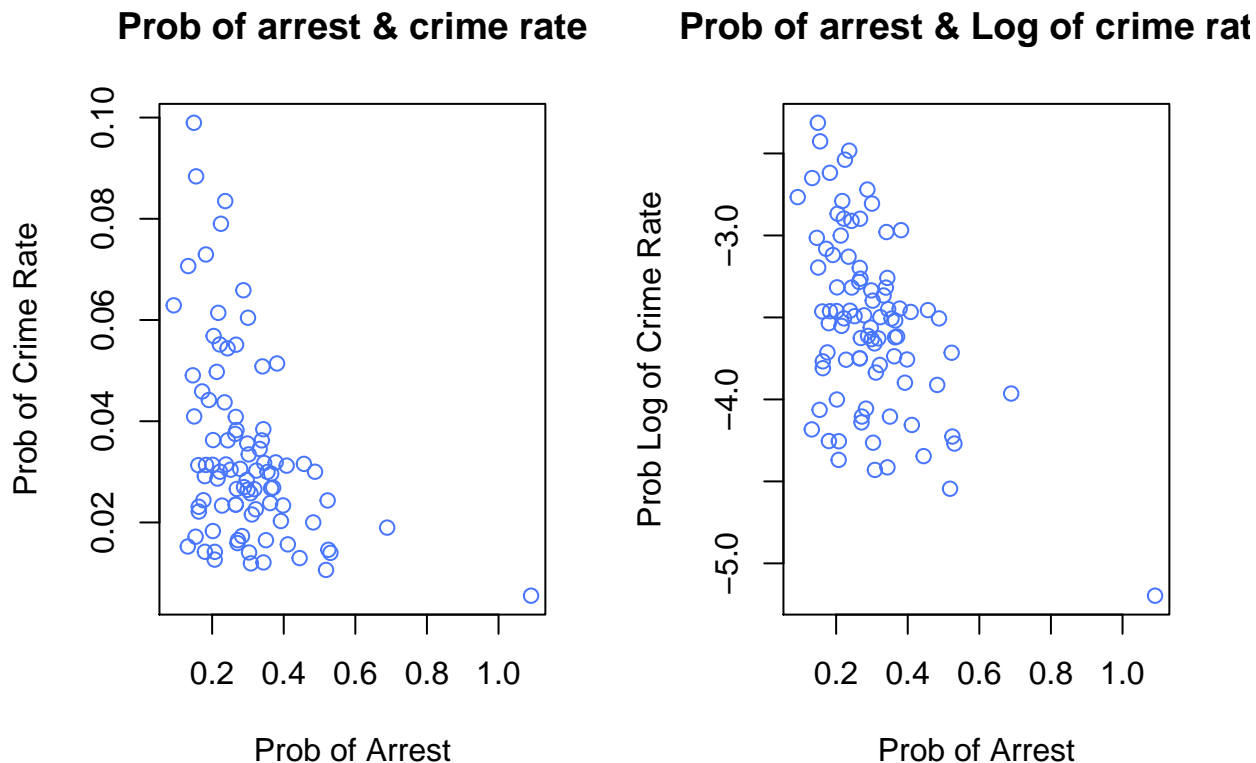
The scatter plot of crrmrte vs. prbarr on the left shows an exponential decay trend. In addition, the variation of crrmrte decreases substantially as prbarr increases. We took the log of crime rate, and then re-graph

the scatter plot (shown on the right). The scatter plot of `crmte_log` vs. `prbarr` indicates a more linear relationship and the variation of `crmte_log` does not vary as much with `prbarr`. The correlation coefficient further supports the transformation.

* The correlation between `crmte` and `prbarr` is -0.41

* The correlation between `crmte_log` and `prbarr` is -0.50

```
par(mfrow=c(1,2))
plot(crimeData2$prbarr, crimeData2$crmte, main = "Prob of arrest & crime rate",
     xlab = "Prob of Arrest", ylab = "Prob of Crime Rate", col="royalblue1")
plot(crimeData2$prbarr, crimeData2$crmte_log, main = "Prob of arrest & Log of crime rate",
     xlab = "Prob of Arrest", ylab = "Prob Log of Crime Rate", col="royalblue1")
```



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

## [1] -0.4076239

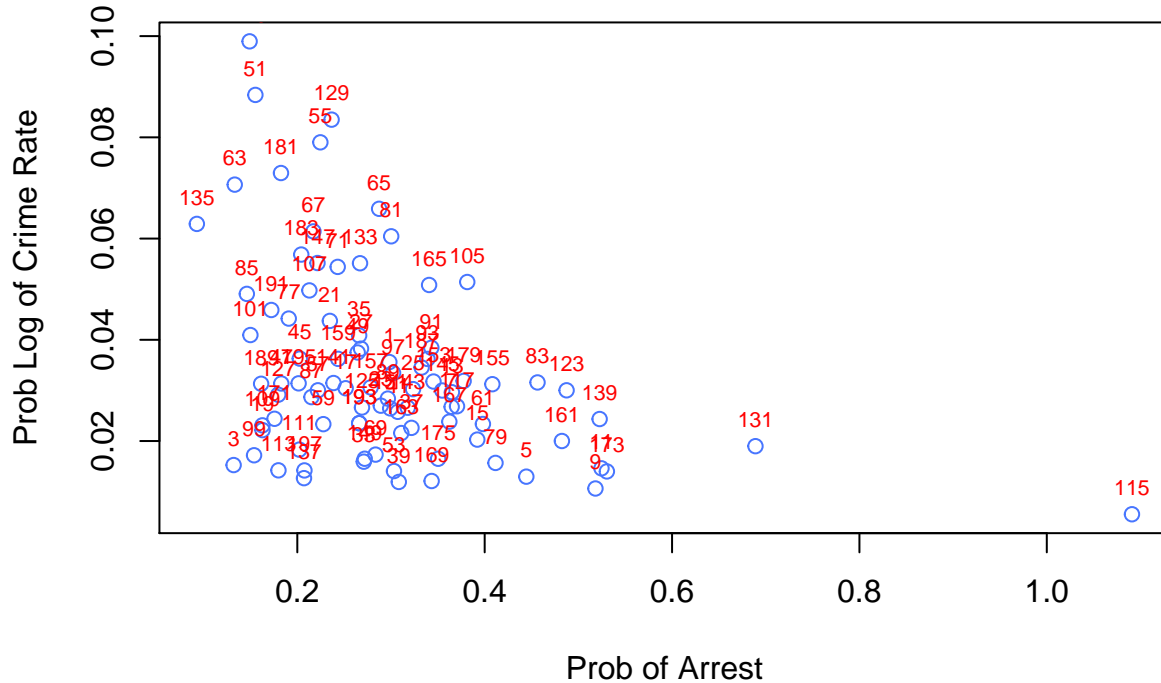
cor(crimeData2$prbarr, crimeData2$crmte_log)

## [1] -0.4964904
```

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 influential 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.

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

Prob of arrest & Log crime rate



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

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

Probability of conviction (prbconv)

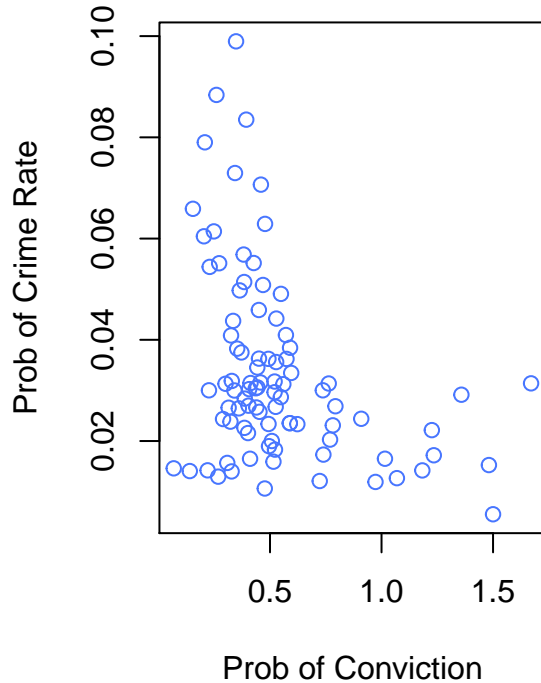
Similar to prbarr, the scatter plot of crmrte vs. prbconv on the left shows an exponential decay trend. In addition, the variation of crmrte decreases substantially as prbconv increases. We took the log of crime rate, and then re-graph the scatter plot (shown on the right). The scatter plot of crmrte_log 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

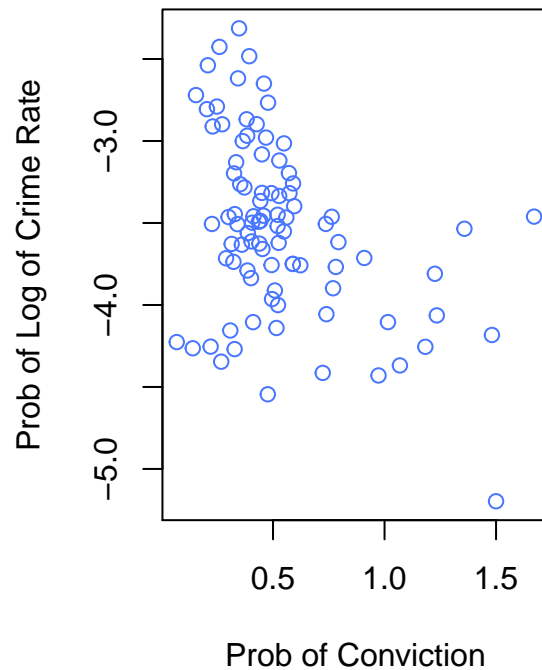
* The correlation between crmrte_log and prbarr is -0.41

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

Prob of conv & crime rate



Prob of conv & log crime rate



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

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

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

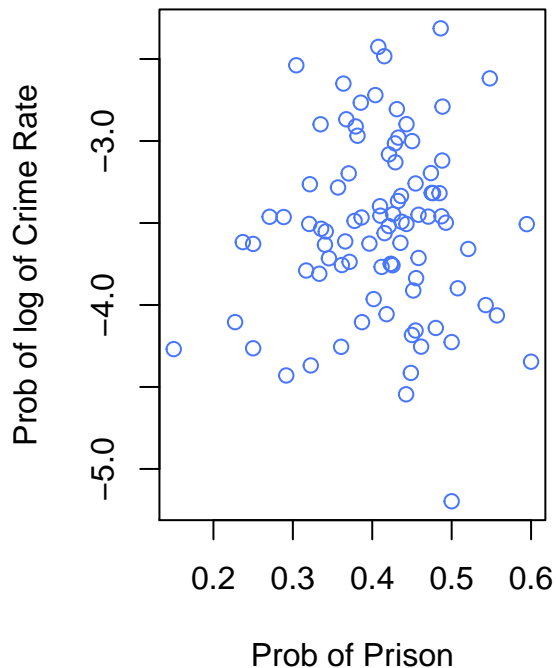
```
## [1] -0.4128166
```

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

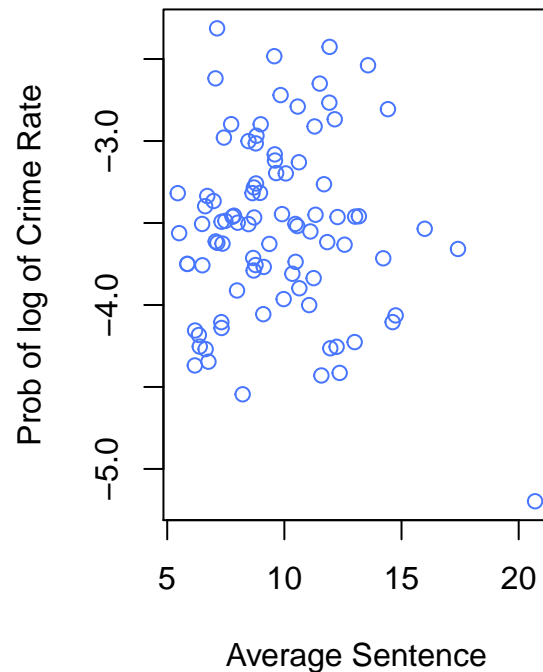
Neither scatter plots below (prbpris vs. crmrte_log, avgsen vs. crmrte_log) shows obvious relationships. The correlation coefficients are only 0.03 and -0.08 respectively.

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

Prob of prison & log of crime rat



Avg sent & log of crime rate



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

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

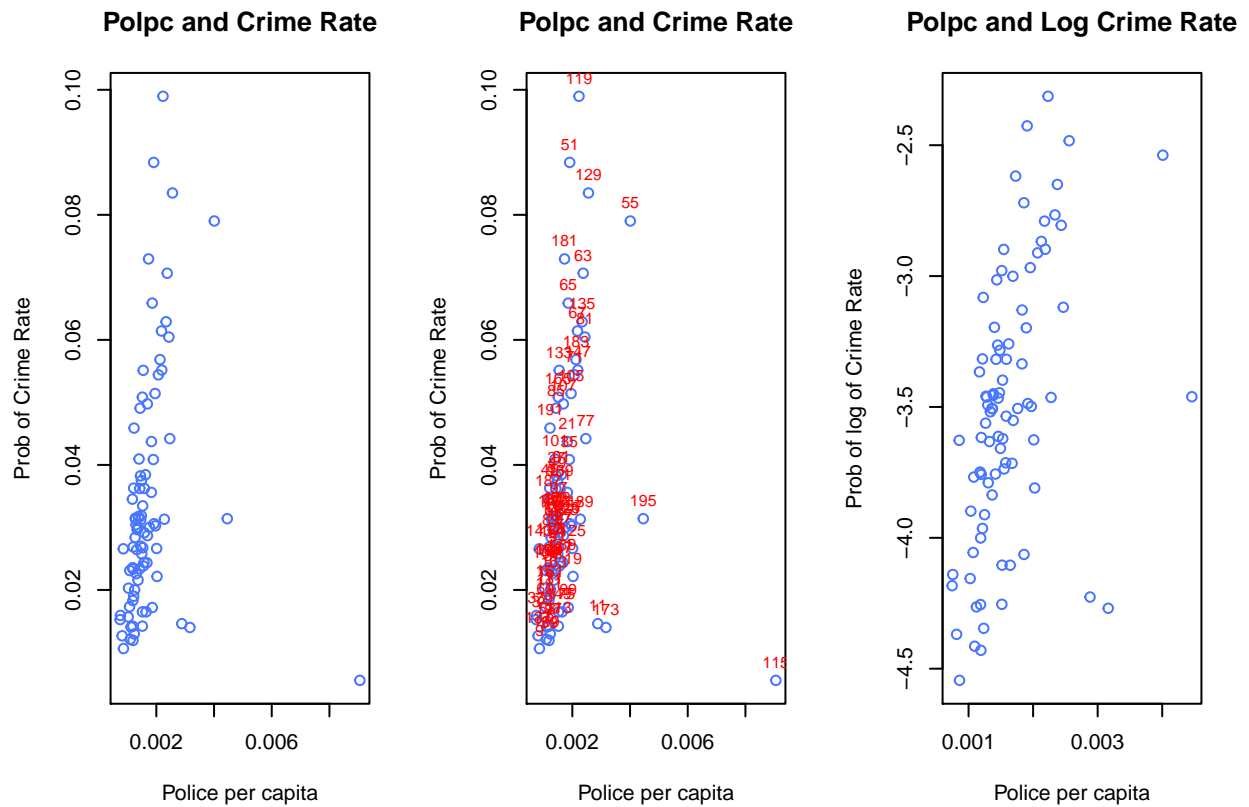
```
cor(crimeData2$avgsgen, crimeData2$crmrtelog)
```

```
## [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 15 has significantly higher police per capital than any other counties, County 15 is a highly leveraged observation. In addition, the variation of `crmrtelog` increases as `prbconv` increases, which justifies taking the log of `crmrtelog`. The correlation between `crmrtelog` and `polpc` (after removing County 115) is 0.45.

```
par(mfrow=c(1,3))
plot(crimeData2$polpc, crimeData2$crmrtelog, main = "Polpc and Crime Rate" ,
     xlab = "Police per capita", ylab = "Prob of Crime Rate", col="royalblue1")
plot(crimeData2$polpc, crimeData2$crmrtelog, main = "Polpc and Crime Rate" ,
     xlab = "Police per capita", ylab = "Prob of Crime Rate", col="royalblue1")
text(crimeData2$polpc, crimeData2$crmrtelog, labels = crimeData2$county, cex=0.7, pos=3, col = "red")
plot(crimeData3$polpc, crimeData3$crmrtelog, main = "Polpc and Log Crime Rate" ,
     xlab = "Police per capita", ylab = "Prob of log of Crime Rate", col="royalblue1")
```

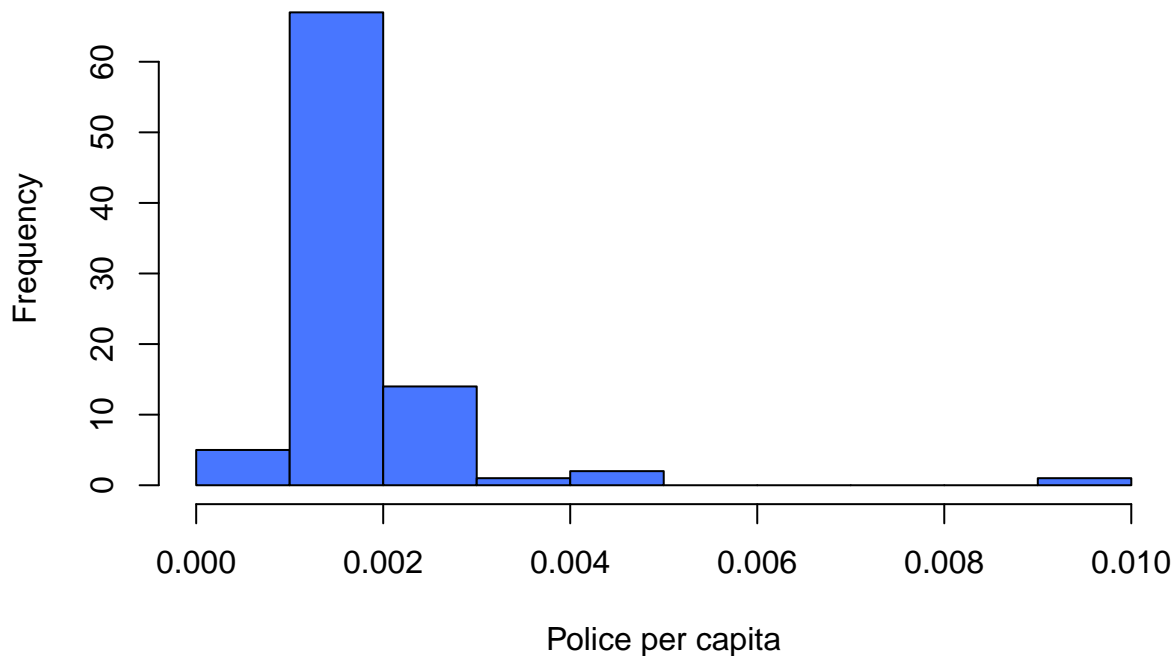
```
cor(crimeData3$polpc, crimeData3$crmrte_log)
```

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

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.

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

Histogram of Police per Capita



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

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

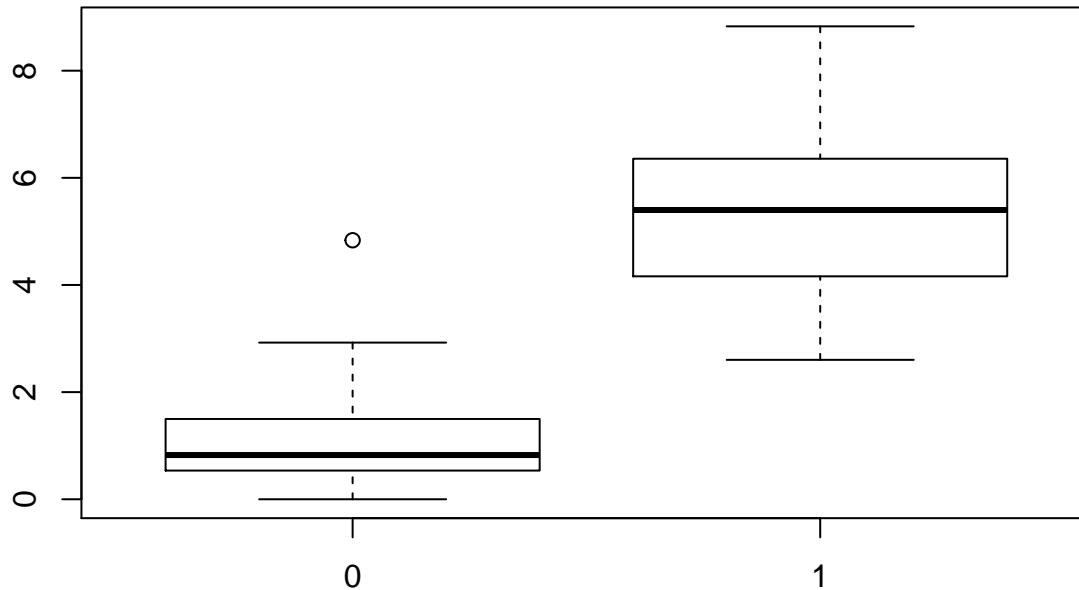
People per square mile (density) | If in SMSA (urban)

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.

```
par(mfrow=c(1,3))
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$crmrate_log, main = "Log of Density & Crime Rate" ,
     xlab = "Log of Density", ylab = "Prob of Log of Crime Rate", col="royalblue1")
plot(crimeData2$density_log, crimeData2$crmrate_log, main = "Log of Density & Crime Rate" ,
     xlab = "Log of Density", ylab = "Prob of Log of Crime Rate", col="royalblue1")
text(crimeData2$density_log, crimeData2$crmrate_log, labels = crimeData2$county, cex=0.7, pos=3 , col =
```

Log of Density & Crime Rate

Box plot of crime rate



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

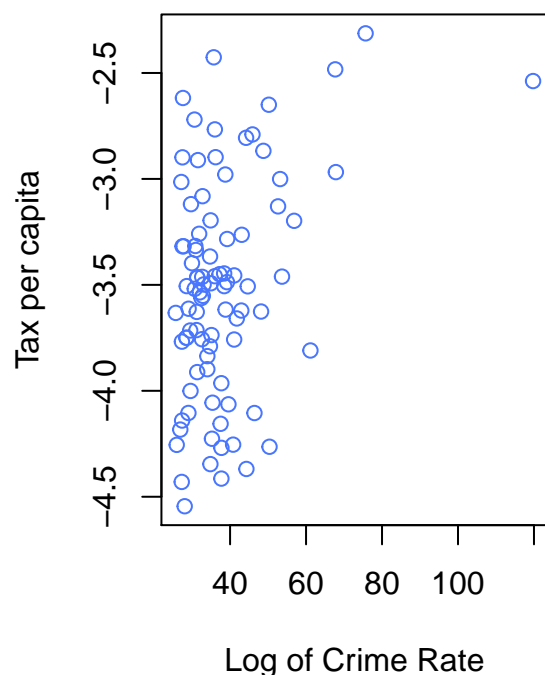
```
## [1] 0.660531
```

Tax revenue per capita (taxpc)

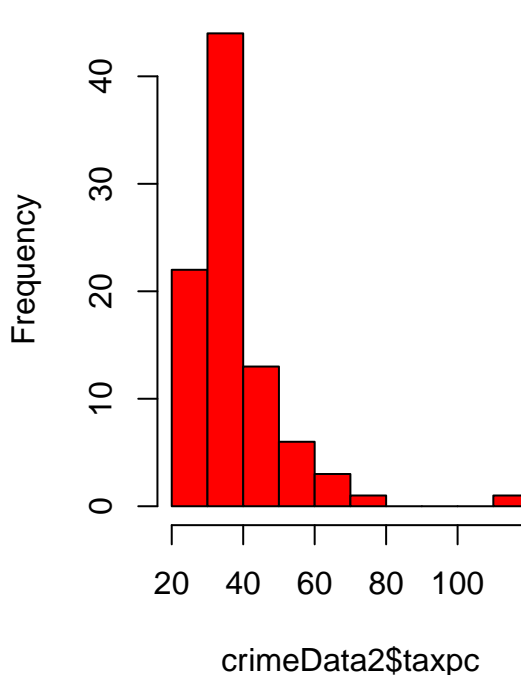
The scatter plot indicates there may be a weak linear relationship between taxpc and crmrte_log. The histogram of taxpc is skewed to the right, so we considered taking the log of taxpc. However, the correlation between taxpc and crmrte_log (0.37) is slightly higher than the correlation between taxpc_log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

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

Crime Rate and tax per capita



Hist. of Tax Revenue



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

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

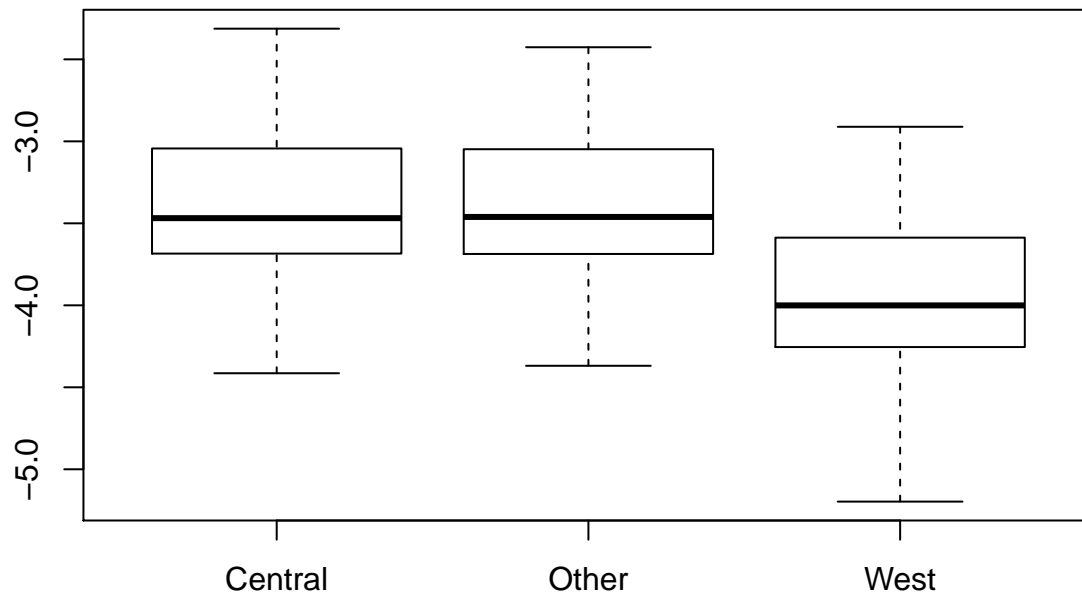
```
cor(crimeData2$taxpc_log, crimeData2$crmrte_log)
```

```
## [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.

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

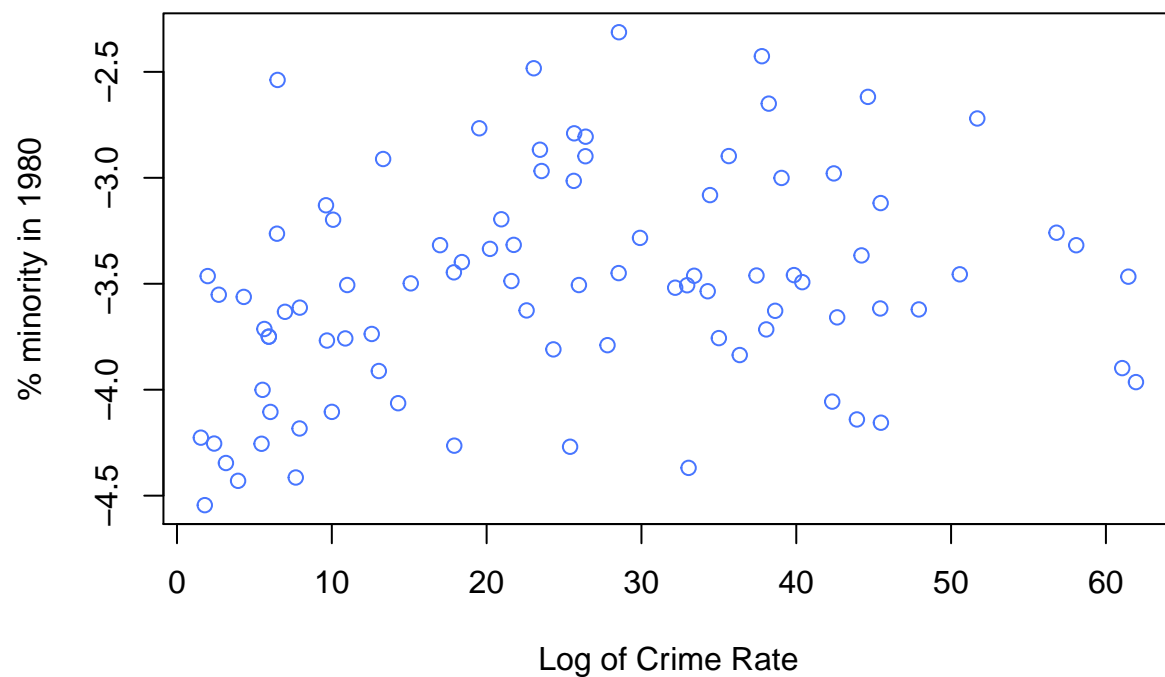


Percent of minority in 1980 (pctmin80)

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.

```
plot(crimeData3$pctmin80, crimeData3$crmrte_log, main = "% Minority and Log crime rate" ,
      xlab = "Log of Crime Rate" , ylab = "% minority in 1980" , col="royalblue1")
```

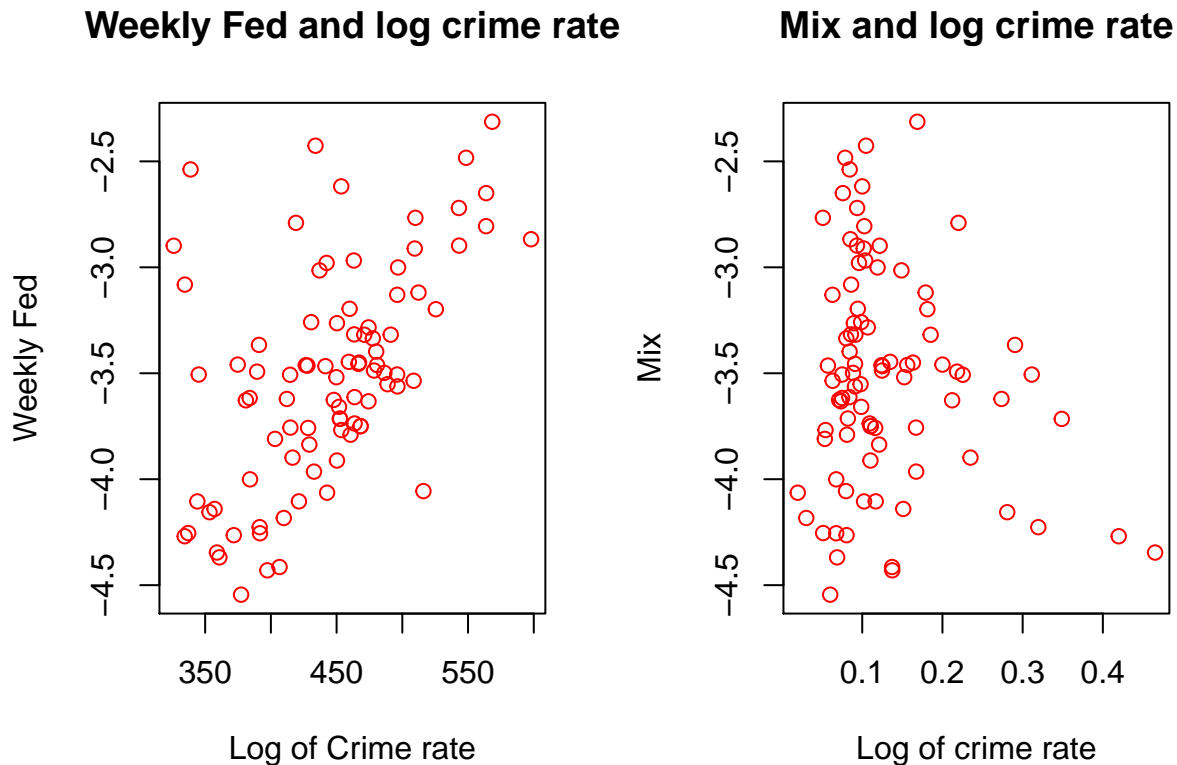
% Minority and Log crime rate



```
cor(crimeData2$pctmin80, crimeData2$crmrte_log)
```



```
plot(crimeData3$mix, crimeData3$crmrte_log, main = "Mix and log crime rate",
     xlab="Log of crime rate", ylab="Mix", col="red")
```



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

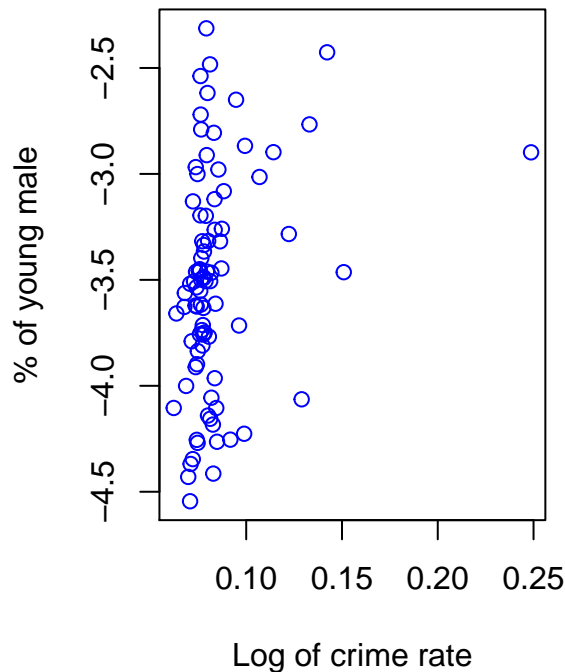
```
## [1] -0.1466527
```

Percent young male (pctymle)

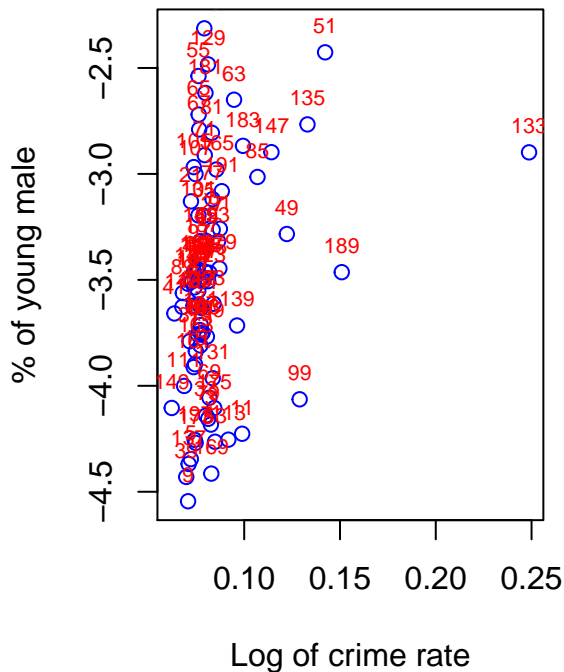
The scatter plot shows that the majority of counties have 5%-10% of young male. County 133 has significantly higher male percentage than the rest of counties. Log of crime rate doesn't seem to vary by the percent of young male based on the scatter plot, which is also evidenced by 0.27 correlation coefficient.

```
par(mfrow=c(1,2))
plot(crimeData3$pctymle, crimeData3$crmrte_log, main = "% young M and Log crime rate",
     xlab="Log of crime rate", ylab="% of young male", col="blue")
plot(crimeData3$pctymle, crimeData3$crmrte_log, main = "% young M and 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")
```


% young M and Log crime rate



% young M and Log crime rate



```
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.

```
model1 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen, data=crimeData2)
model1$coefficients
```

```
## (Intercept)      prbarr      prbconv      prbpris      avgsen
## -2.92426013 -2.09740014 -0.79654556  0.42628304  0.02705387
```

```
#library(lmtest)#coefTest(model1, vcov = vcovHC)#plot(model1)#AIC(model1)
```

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

Model Building 2

We put all the variables with correlation higher than 0.25 in the model except for those highly correlated with each other. Also We took the log of police per capita and people per square mile because EDA showed that they have a stronger linear relationship with log of crime rate compared to their original form.

```
model2 <- lm(crmrte_log ~ prbarr+prbconv+polpc_log+density_log+west+pctmin80+wfed+
  pctymle, data=crimeData2)
model2$coefficients

## (Intercept)      prbarr      prbconv      polpc_log      density_log
## -0.002959757 -1.876715831 -0.665762798  0.546997370  0.083469653
##          west      pctmin80          wfed      pctymle
## -0.101386251  0.010011686  0.001356532  1.111404818

#coeftest(model2, vcov = vcovHC)#plot(model2)#AIC(model2)
```

Model Building 3

For this model, we add all the variables to the model.

```
model3 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen+polpc_log+density_log+taxpc+west+
  central+urban+pctmin80_2+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc+mix+
  pctymle, data=crimeData2)
model3$coefficients

## (Intercept)      prbarr      prbconv      prbpris      avgsen
## 0.6276444944 -1.8190558437 -0.5659797104 -0.6806825821 -0.0270524579
##      polpc_log      density_log          taxpc          west      central
## 0.6021669897  0.1027476586  0.0019230455 -0.0683115182 -0.0712727704
##      urban      pctmin80_2          wcon          wtuc          wtrd
## 0.1513316157  0.9980385344  0.0003784841  0.0002289608  0.0017024817
##      wfir          wser          wmfg          wfed          wsta
## -0.0009627851 -0.0020813035 -0.0003396119  0.0017579658 -0.0006933810
##      wloc          mix      pctymle
## 0.0016560744 -0.0504661995  2.0567468321

#coeftest(model3, vcov = vcovHC)#plot(model3)#AIC(model3)
```

Model Display

```
#se.model1 = sqrt(diag(vcovHC(model1)))#se.model2 = sqrt(diag(vcovHC(model2)))#se.model3 = sqrt(diag(vcovHC(model3)))
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

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: Mon, Apr 02, 2018 - 17:48:00

	<i>Dependent variable:</i>		
	crmrte_log		
	(1)	(2)	(3)
prbarr	−2.097*** (0.323)	−1.877*** (0.215)	−1.819*** (0.232)
prbconv	−0.797*** (0.145)	−0.666*** (0.082)	−0.566*** (0.098)
prbpris	0.426 (0.543)		−0.681 (0.362)
avgsen	0.027 (0.016)		−0.027* (0.011)
polpc_log		0.547*** (0.079)	0.602*** (0.106)
density_log		0.083*** (0.024)	0.103*** (0.030)
taxpc			0.002 (0.003)
west		−0.101 (0.082)	−0.068 (0.109)
pctmin80		0.010*** (0.002)	
central			−0.071 (0.074)
urban			0.151 (0.116)
pctmin80_2			0.998*** (0.258)
wcon			0.0004 (0.001)
wtuc			0.0002 (0.0004)
wtrd			0.002 (0.001)
wfir			−0.001 (0.001)
wser			−0.002* (0.001)
wmfg		19	−0.0003 (0.0004)
wfed		0.001* (0.001)	0.002* (0.001)

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 , $\alpha_1(\text{estimated}) = -2.1 = \beta_1(\text{true}) + \text{positive}$

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

True Coefficient 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 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 , $\alpha_1(\text{estimated}) = -0.8 = \beta_1(\text{true}) + \text{positive}$

$$\alpha(\text{estimated}) = -0.8 = \beta_1(\text{true}) + (\text{positive})$$

True Coefficient 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 , $\alpha_1(\text{estimated}) = -2.1 = \beta_1(\text{true}) + \text{positive}$

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

True Coefficient 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 conviction rates. Overall positive bias.

$$\alpha_1 < 0$$

This is positive bias , $\alpha_1(\text{estimated}) = -0.8 = \beta_1(\text{true}) + \text{positive}$

$$\alpha(\text{estimated}) = -0.8 = \beta_1(\text{true}) + (\text{positive})$$

True Coefficient 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 , $\alpha_1(\text{estimated}) = -2.1 = \beta_1(\text{true}) + \text{positive}$

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

True Coefficient 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 convictions 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 , $\alpha_1(\text{estimated}) = -0.8 = \beta_1(\text{true}) + \text{positive}$

$$\alpha(\text{estimated}) = -0.8 = \beta_1(\text{true}) + (\text{positive})$$

True Coefficient 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 , $\alpha_1(\text{estimated}) = -2.1 = \beta_1(\text{true}) + \text{positive}$

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

True Coefficient 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 , $\alpha_1(\text{estimated}) = -0.8 = \beta_1(\text{true}) + \text{negative}$

$$\alpha(\text{estimated}) = -0.8 = \beta_1(\text{true}) + (\text{negative})$$

True Coefficient 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 , α_1 (estimated) = -2.1 = β_1 (true) + negative

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

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 , α (estimated) = -0.8 = β (true) + negative

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

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