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           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.

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

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

```
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
```

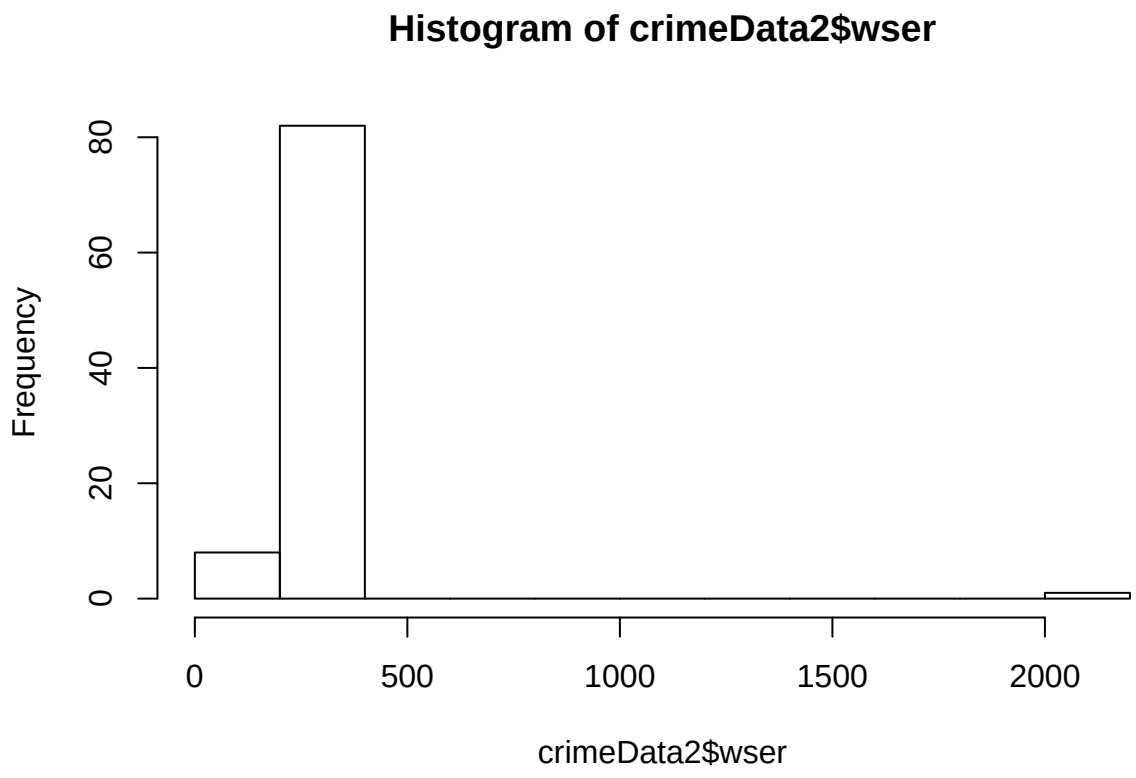
Variable 'prbarr' is defined as the ratio of arrests to offenses. One possible explanation for 'prbarr' being greater than 1 is that multiple people who convicted a single crime together is counted as one conviction but multiple arrests.
Variable 'prbconv' is defined as the ratio of convictions to arrests. 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)
```




```
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),]
```

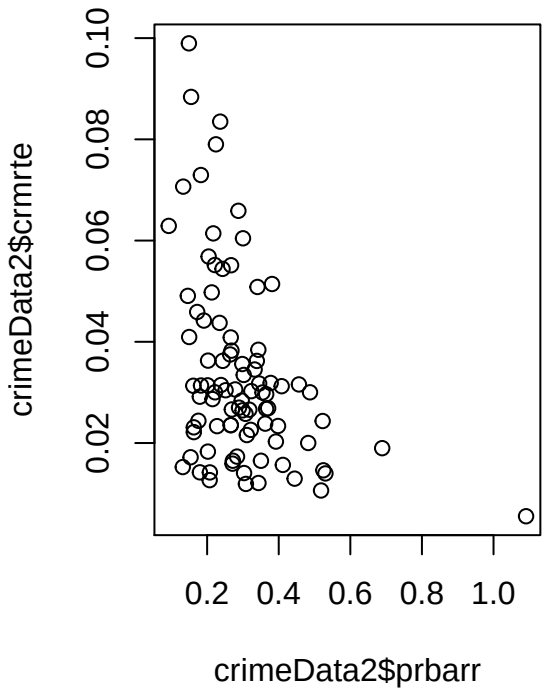
Exploratory Data Analysis

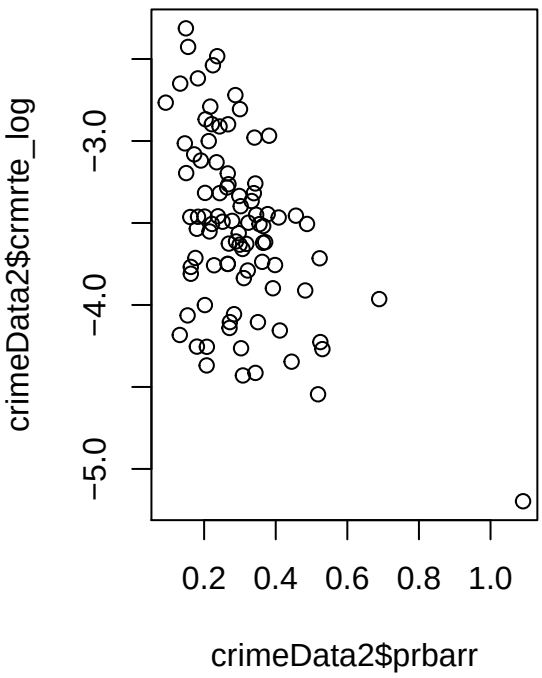
We wanted to first investigate the relationship between our target variable (crime rate) and the variables of key interest.

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. This suggests that the relationship between crmrte and prbarr is not linear. So 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

```
par(mfrow=c(1,2))
plot(crimeData2$prbarr, crimeData2$crmrte)
crimeData2$crmrte_log = log(crimeData2$crmrte)
plot(crimeData2$prbarr, crimeData2$crmrte_log)
```






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

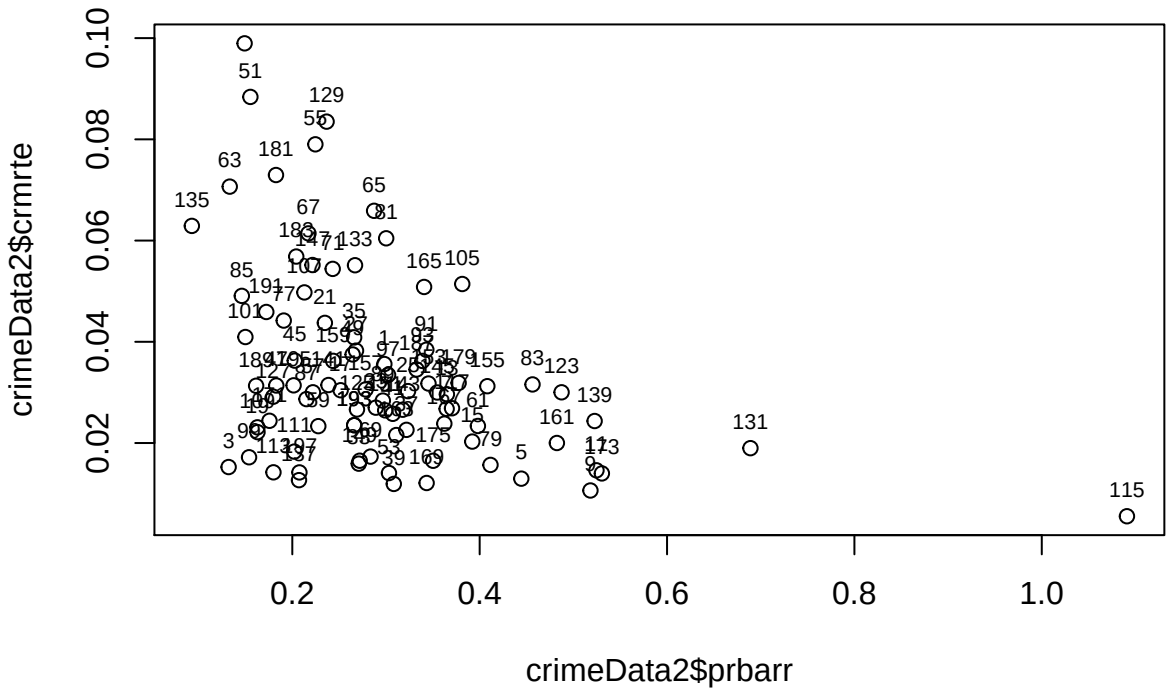
```
## [1] -0.4076239
```

```
cor(crimeData2$prbarr, crimeData2$crmrte_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.

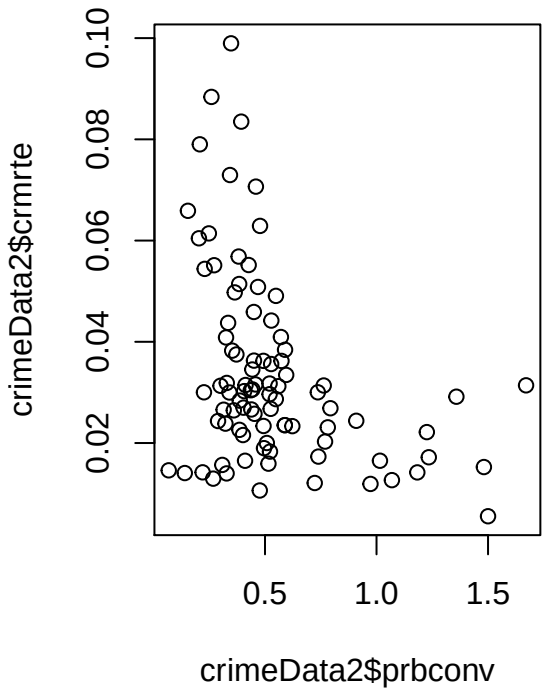
```
plot(crimeData2$prbarr, crimeData2$crmrte)
text(crimeData2$prbarr, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3)
```

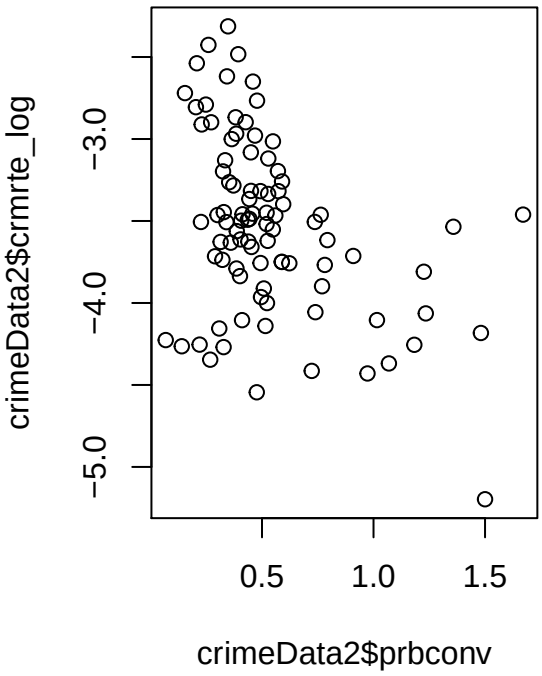


Probability of conviction

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. This suggests that the relationship between crmrte and prbconv is not linear. So 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$prbconv, crimeData2$crmrte)
plot(crimeData2$prbconv, crimeData2$crmrte_log)
```





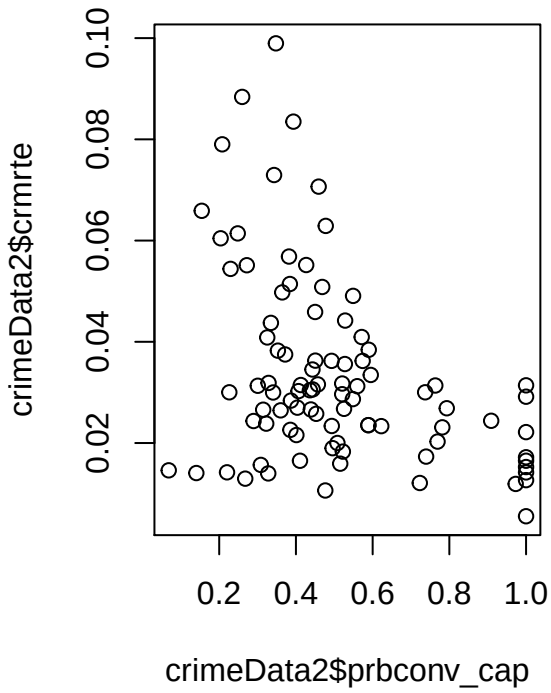

```
cor(crimeData2$prbconv, crimeData2$crmrte)
```

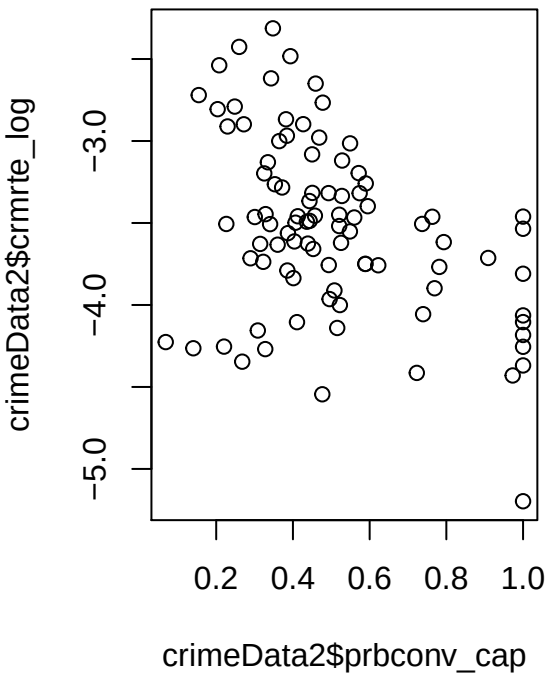
```
## [1] -0.3728922
```

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

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

```
########################## ? ##################################
crimeData2$prbconv_cap = replace(crimeData2$prbconv, crimeData2$prbconv > 1, 1)
plot(crimeData2$prbconv_cap, crimeData2$crmrte)
plot(crimeData2$prbconv_cap, crimeData2$crmrte_log)
```






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

```
## [1] -0.4024425
```

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

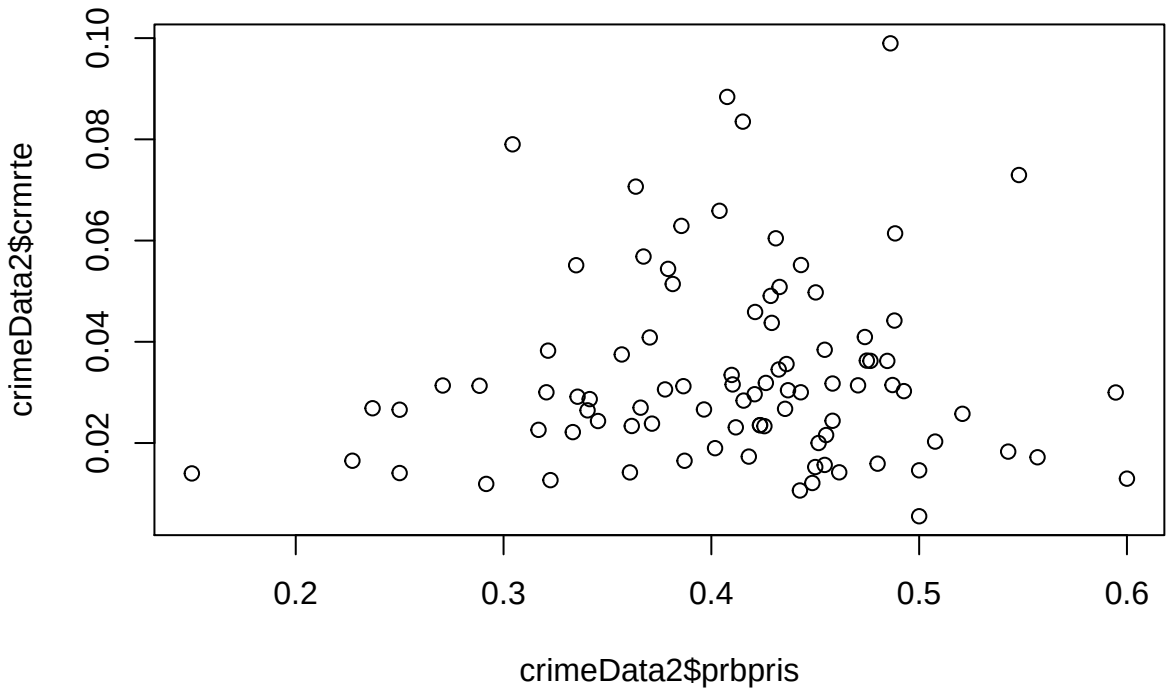
```
## [1] -0.4204057
```

```
#####################################################################
```

Probability of prison

The scatter plot of prbpris vs. crmrte doesn't show an obvious relationship. The correlation coefficient is only 0.05.

```
plot(crimeData2$prbpris, crimeData2$crmrte)
```



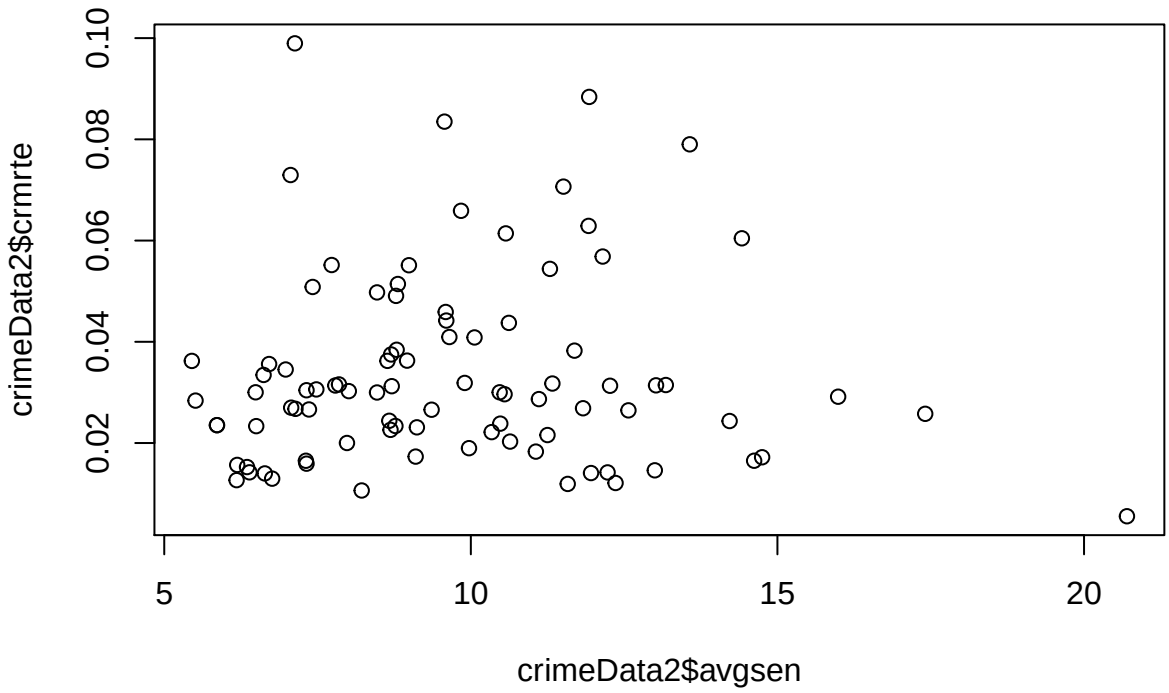

```
cor(crimeData2$prbpris, crimeData2$crmrte)
```

```
## [1] 0.05284061
```

Average sentence days

The scatter plot of avgsen vs. crmrte doesn't show an obvious relationship. The correlation coefficient is only 0.01.

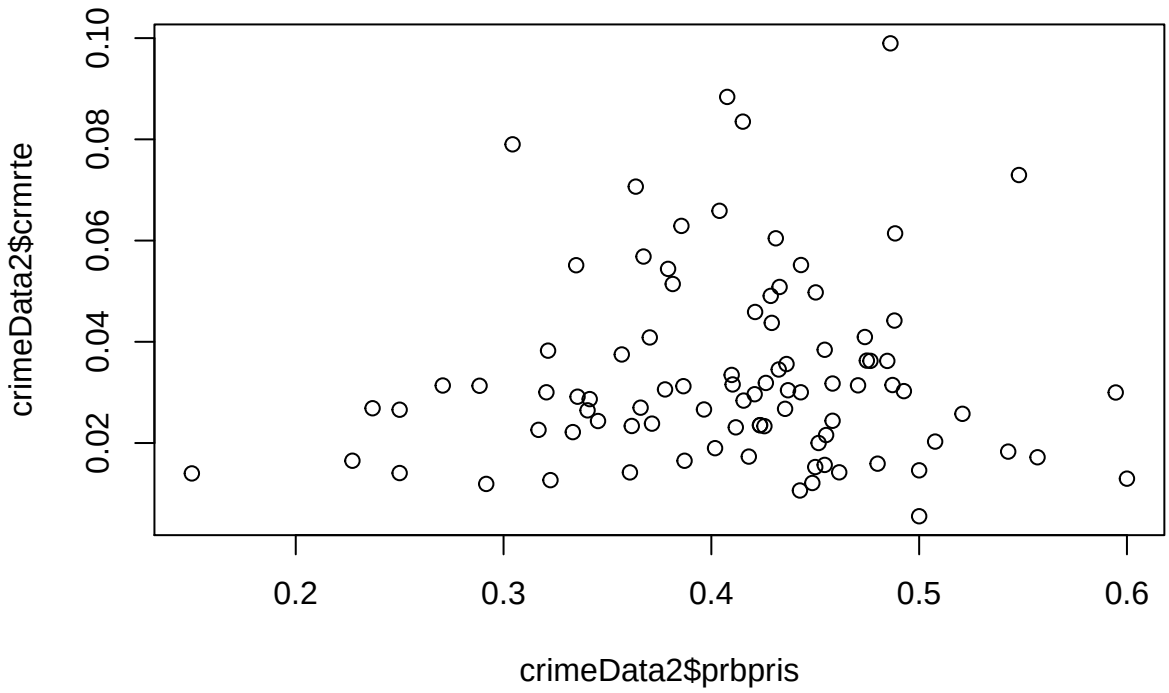
```
plot(crimeData2$avgsen, crimeData2$crmrte)
```



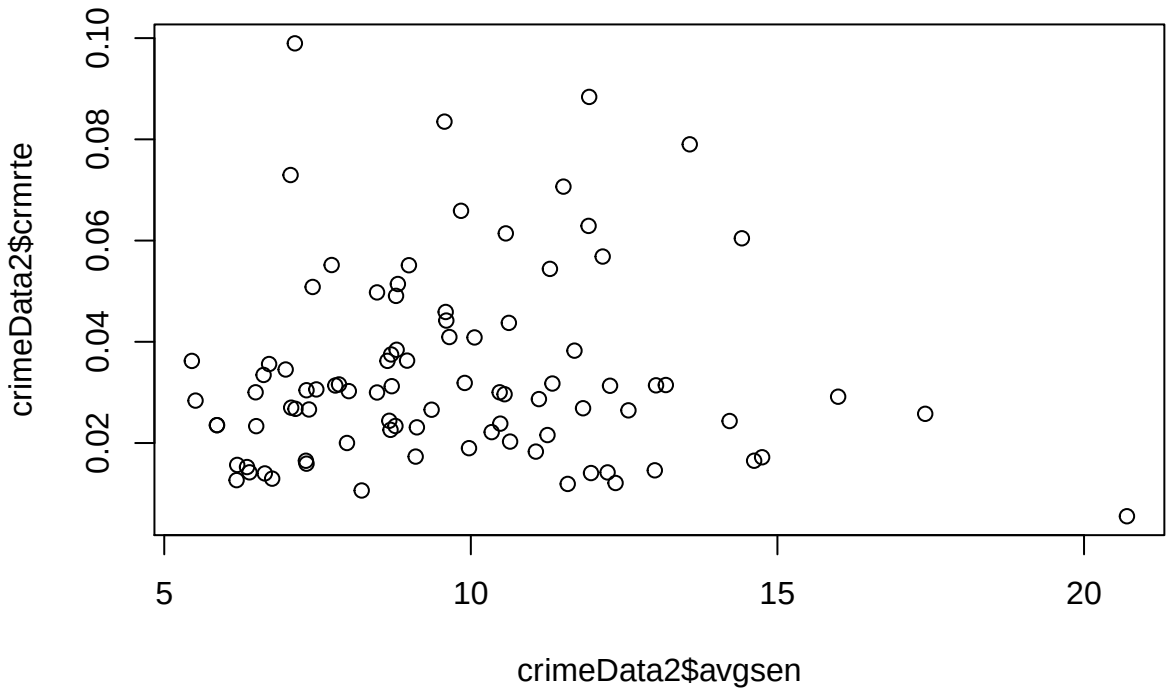

```
cor(crimeData2$avgsen, crimeData2$crmrte)
```

```
## [1] 0.007397583
```

```
plot(crimeData2$prbpris, crimeData2$crmrte)
```




```
plot(crimeData2$avgsen, crimeData2$crmrte)
```




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

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

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

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

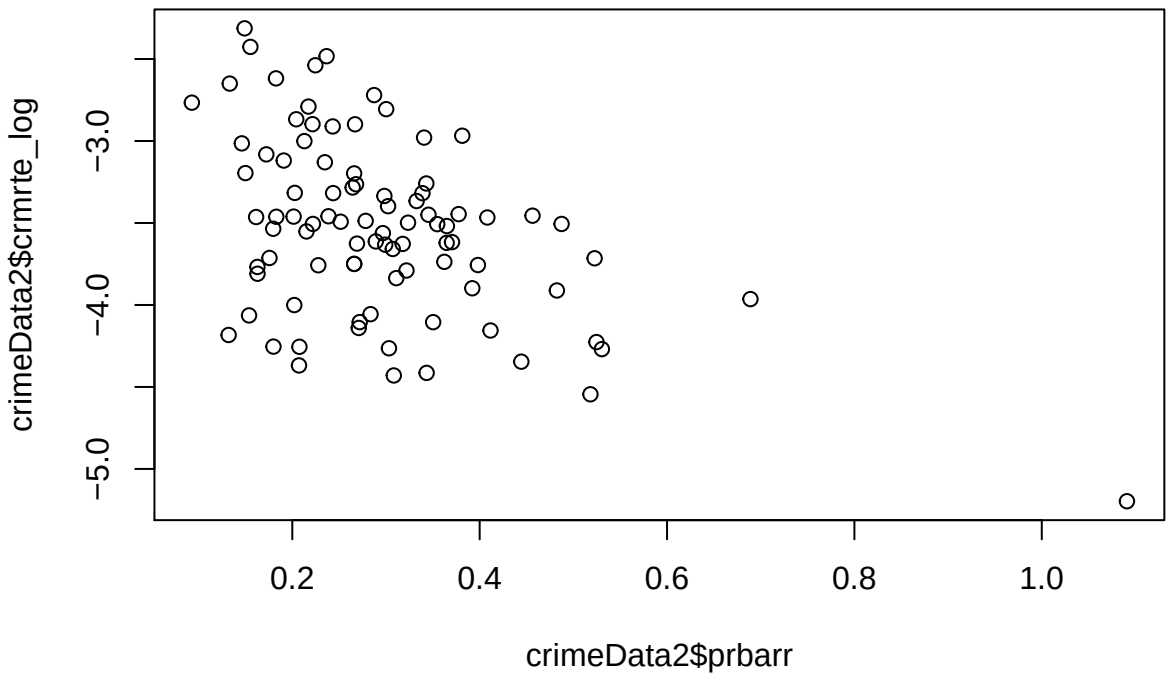
```
cor(crimeData2$crmrte, crimeData2$prbpris)
```

```
## [1] 0.05284061
```

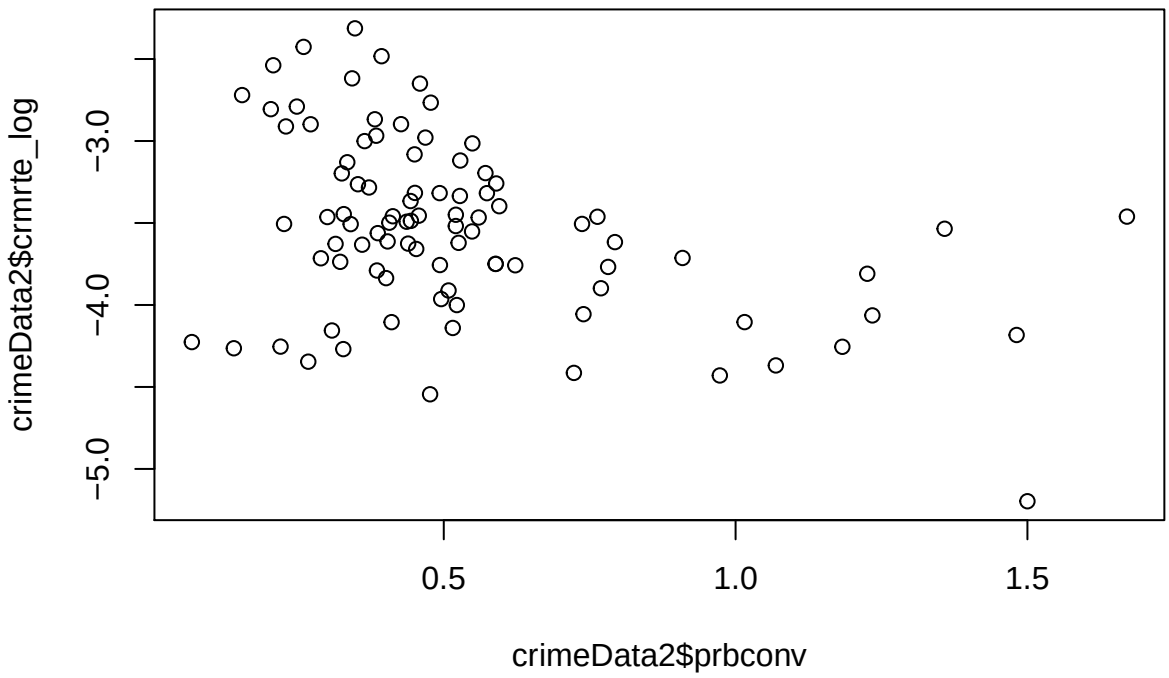
```
cor(crimeData2$crmrte, crimeData2$avgsen)
```

```
## [1] 0.007397583
```

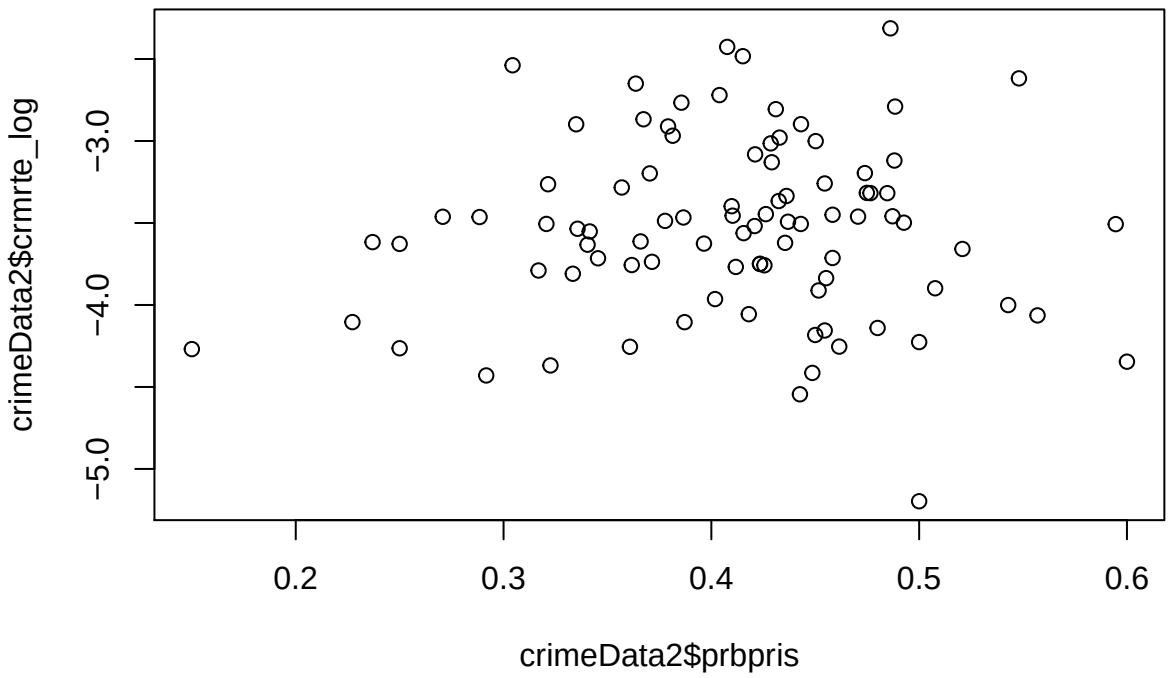
```
crimeData2$crmrte_log = log(crimeData2$crmrte)
plot(crimeData2$prbarr, crimeData2$crmrte_log)
```



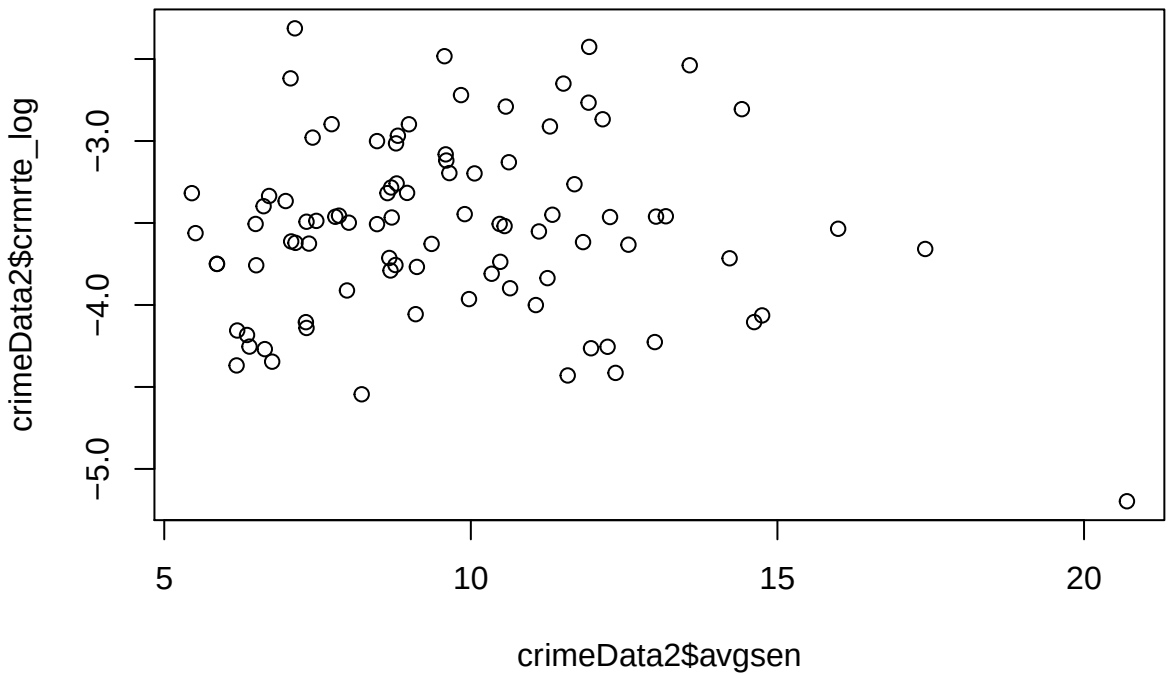

```
plot(crimeData2$prbconv, crimeData2$crmrte_log)
```




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




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




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

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

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

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

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

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

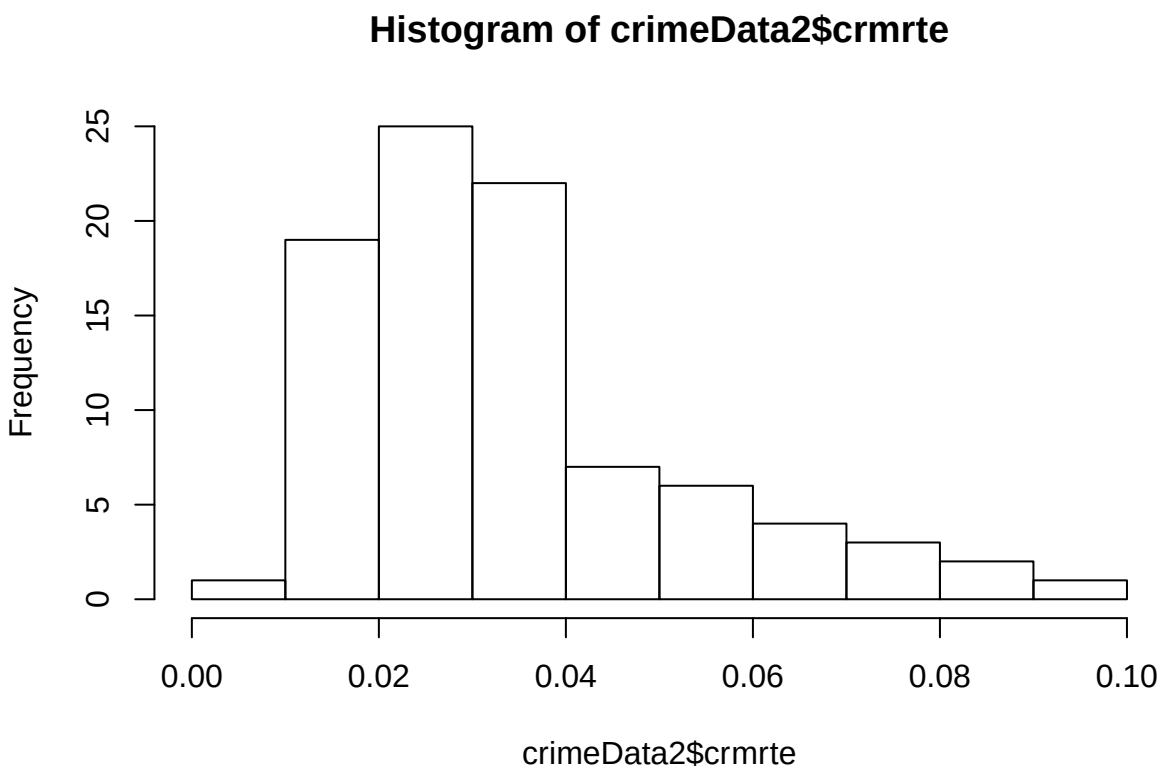
```
cor(crimeData2$crmrte_log,crimeData2$avgsen)
```

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

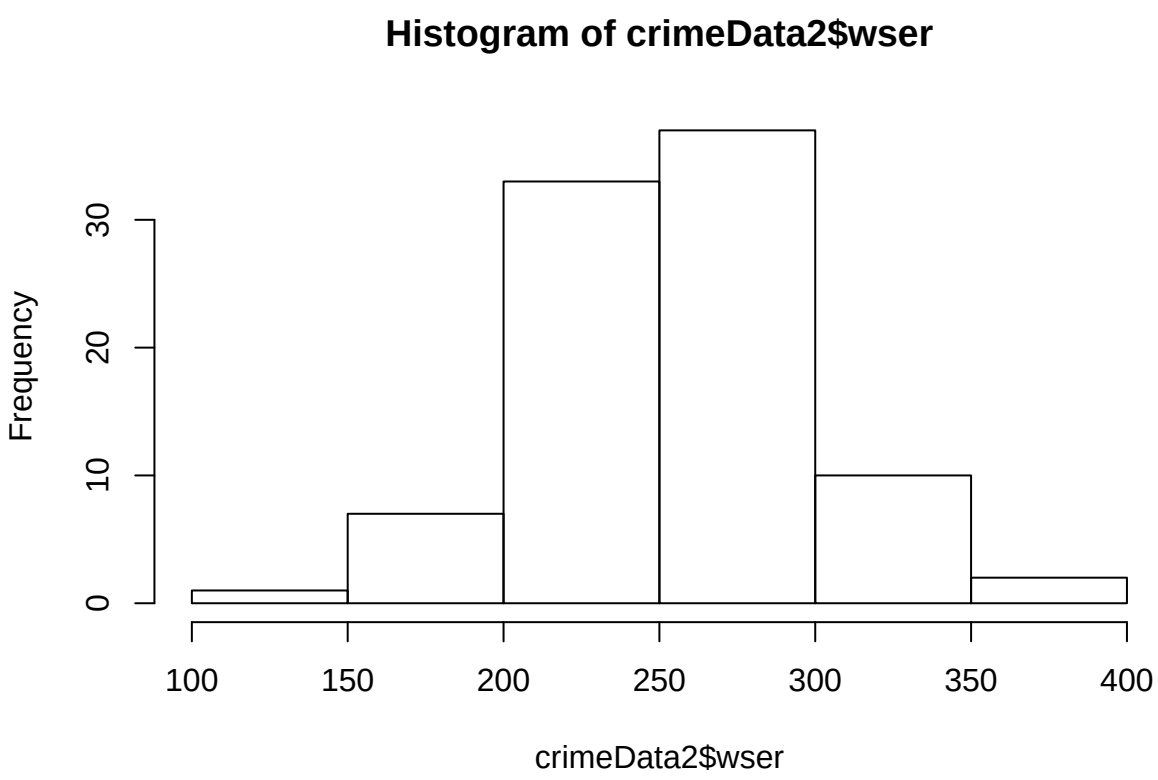
```
#crimeData2_drop <- c("year") # extra column
#crimeData3 = crimeData2[ , !(names(crimeData2) %in% crimeData2_drop)]
#correlations <- cor(na.omit(crimeData3[,-1,]))

# correlations
#row_indic <- apply(correlations, 1, function(x) sum(x > 0.3 | x < -0.3) > 1)
#correlations<- correlations[row_indic ,row_indic ]
#install.packages("corrplot", dependencies = TRUE)
#library(corrplot)
#corrplot(correlations, method="square")
```

```
hist(crimeData2$crmrte)
```




```
hist(crimeData2$wser)
```

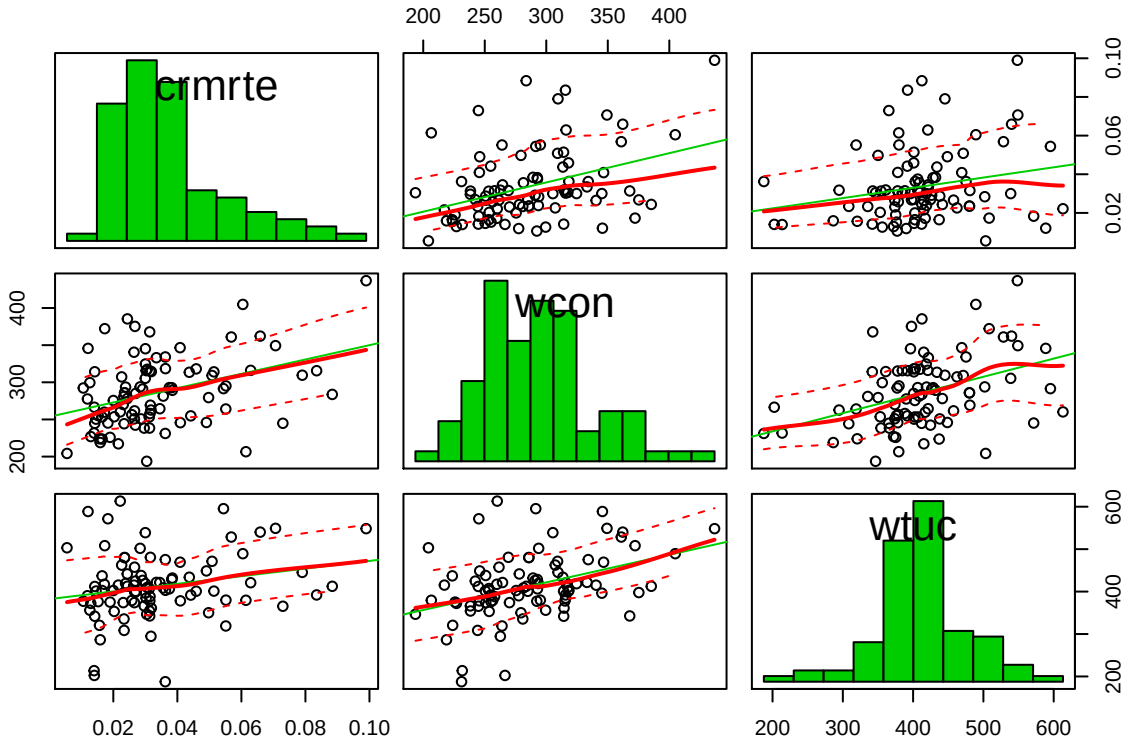


histogram + correlation graph => bin? transformation? Bhuvnesh: prbarr - pctmin80 Angela: wcon - pctmle meet 4pm CST tomorrow due Monday 8pm CST

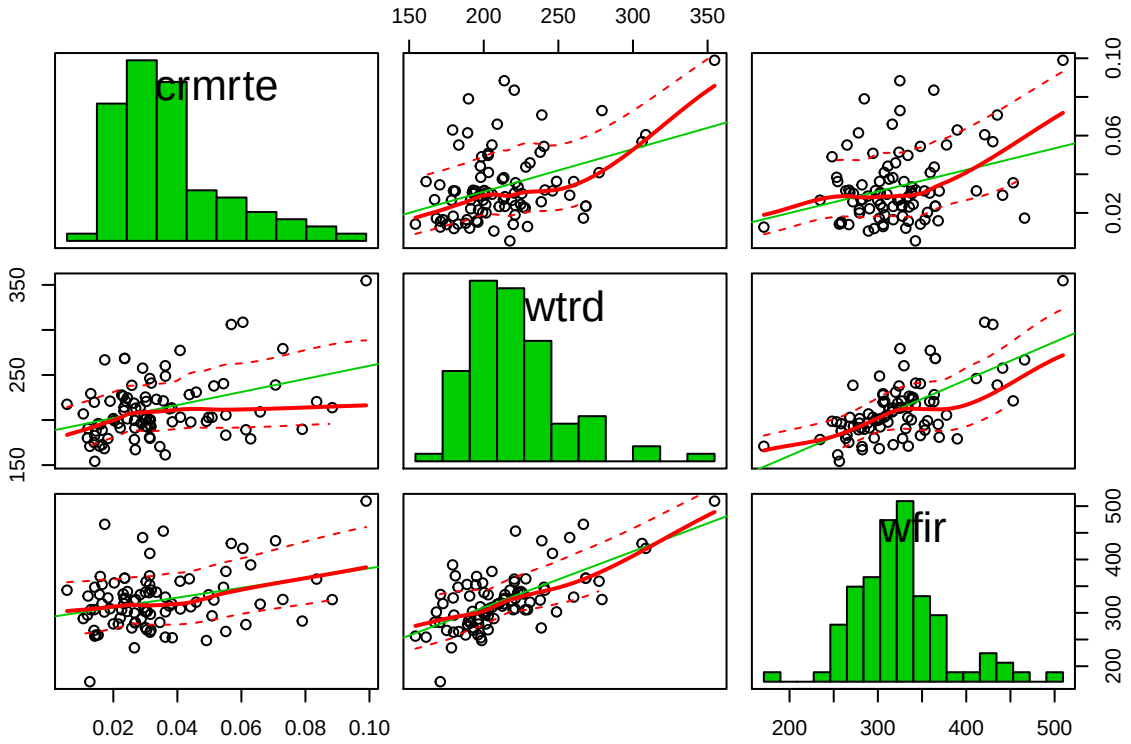
```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

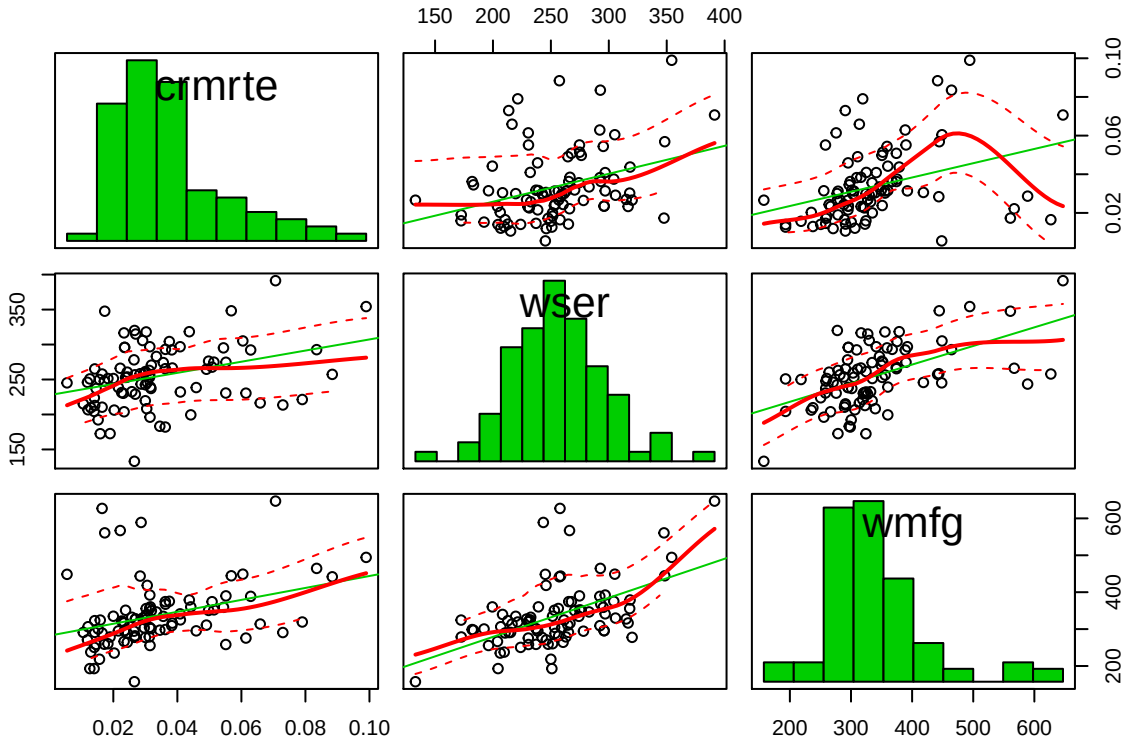
```
scatterplotMatrix(~ crmrte+wcon+wtuc, data=crimeData2, diagonal="histogram")
```



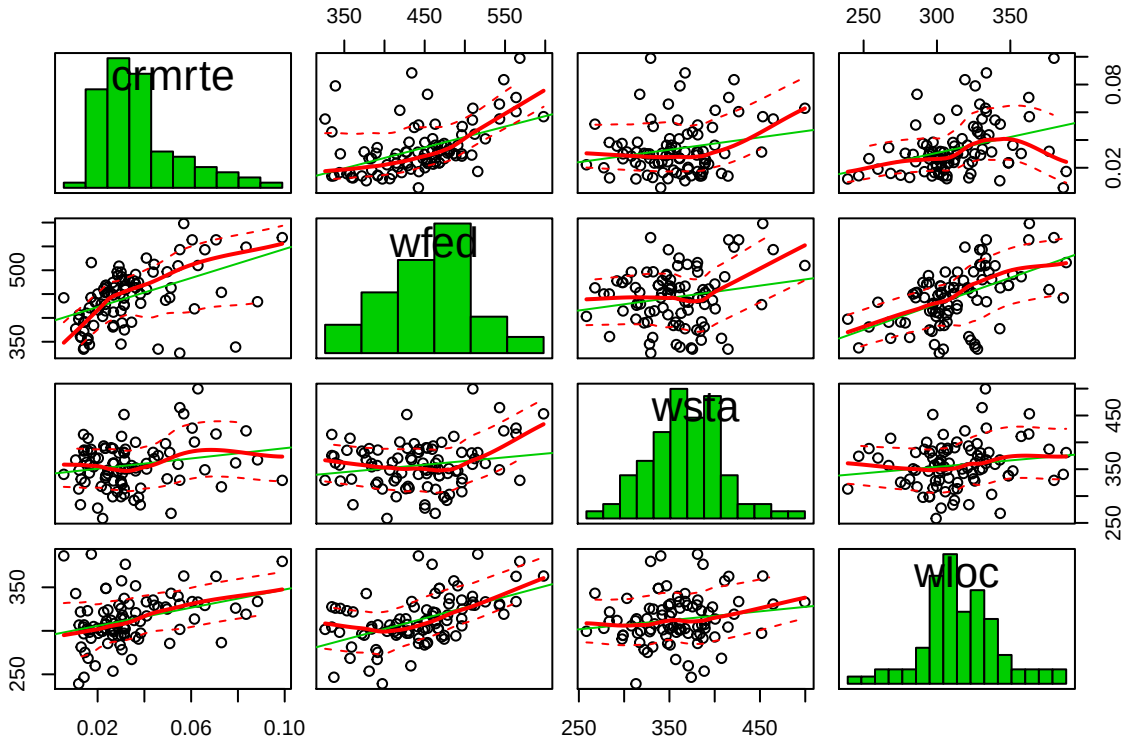

```
scatterplotMatrix(~ crmrte+wtrd+wfir, data=crimeData2, diagonal="histogram")
```



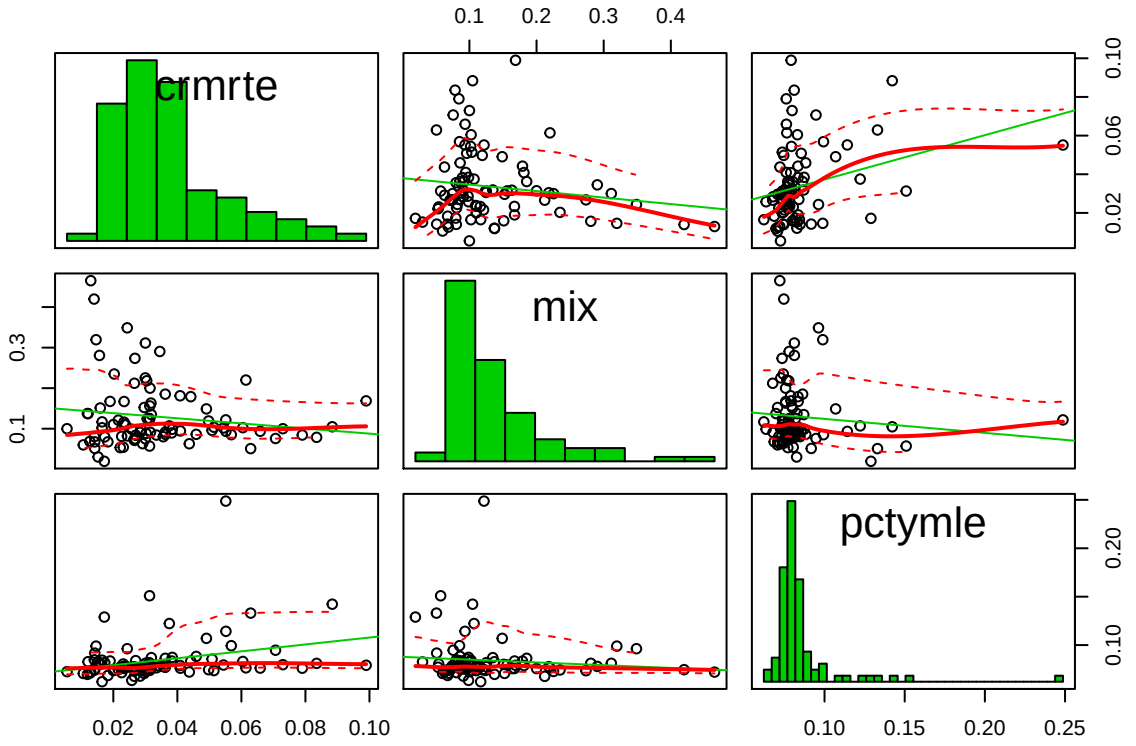

```
scatterplotMatrix(~ crmrte+wser+wmfg, data=crimeData2, diagonal="histogram")
```




```
scatterplotMatrix(~ crmrte+wfed+wsta+wloc, data=crimeData2, diagonal="histogram")
```




```
scatterplotMatrix(~ crmrte+mix+pctymle, data=crimeData2, diagonal="histogram")
```



I think I would delete wser = 2177.068 because it significantly distorted the trend. wmfg has a weird shape. wsta has a weird flat then increase trend. not sure how to interpret the variable 'mix' pctymle has one extreme value as well. This data point is highly leveraged and could potentially be influential. Most other variables look like they have linear relationships with crmrte.

```
#crimeData3[which(crimeData2$pctymle>0.1),]
#crimeData3$w_sum <- crimeData3$wcon+crimeData3$wtuc+crimeData3$wtrd+crimeData3$wfir+crimeData3$wser+cr
#cor(crimeData3$crmrte,crimeData3$w_sum)
```

Model Building 1

```
step(lm(crmrte ~ prbarr+prbconv+prbpris+avgsen+polpc+density+taxpc+west+central+urban+pctmin80_2+wcon+w
```

```
## Start:  AIC=-856.8
## crmrte ~ prbarr + prbconv + prbpris + avgsen + polpc + density +
##     taxpc + west + central + urban + pctmin80_2 + wcon + wtuc +
##     wtrd + wfir + wser + wmfg + wfed + wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - prbpris     1 0.00000048 0.0039611 -858.79
## - urban       1 0.00000133 0.0039620 -858.77
## - wmfg        1 0.00000622 0.0039669 -858.66
## - wtrd        1 0.00002922 0.0039899 -858.14
## - west        1 0.00003014 0.0039908 -858.12
## - wfir        1 0.00003033 0.0039910 -858.12
```

```
## - wtuc        1 0.00003308 0.0039937 -858.06
## - wloc        1 0.00006341 0.0040241 -857.37
## <none>                     0.0039607 -856.80
## - wcon        1 0.00009106 0.0040517 -856.76
## - wsta        1 0.00010424 0.0040649 -856.47
## - avgsen      1 0.00012107 0.0040817 -856.09
## - mix         1 0.00014070 0.0041013 -855.66
## - central     1 0.00021965 0.0041803 -853.95
## - wfed        1 0.00024072 0.0042014 -853.49
## - taxpc       1 0.00025196 0.0042126 -853.25
## - pctymle     1 0.00048977 0.0044504 -848.31
## - pctmin80_2  1 0.00061741 0.0045781 -845.77
## - wser        1 0.00065950 0.0046202 -844.94
## - density     1 0.00106644 0.0050271 -837.35
## - polpc       1 0.00141130 0.0053720 -831.37
## - prbconv     1 0.00191438 0.0058750 -823.32
## - prbarr      1 0.00199338 0.0059540 -822.11
##
## Step:  AIC=-858.79
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     west + central + urban + pctmin80_2 + wcon + wtuc + wtrd +
##     wfir + wser + wmfg + wfed + wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - urban       1 0.00000141 0.0039625 -860.76
## - wmfg        1 0.00000638 0.0039675 -860.65
## - west        1 0.00002974 0.0039909 -860.12
## - wtrd        1 0.00003071 0.0039918 -860.10
## - wfir        1 0.00003079 0.0039919 -860.10
## - wtuc        1 0.00003572 0.0039968 -859.99
## - wloc        1 0.00006433 0.0040255 -859.34
## <none>                     0.0039611 -858.79
## - wcon        1 0.00009171 0.0040528 -858.73
## - wsta        1 0.00010466 0.0040658 -858.45
## - avgsen      1 0.00012511 0.0040862 -857.99
## - mix         1 0.00014041 0.0041015 -857.66
## + prbpris     1 0.00000048 0.0039607 -856.80
## - central     1 0.00022173 0.0041829 -855.89
## - wfed        1 0.00024127 0.0042024 -855.47
## - taxpc       1 0.00025174 0.0042129 -855.25
## - pctymle     1 0.00048964 0.0044508 -850.30
## - pctmin80_2  1 0.00063006 0.0045912 -847.51
## - wser        1 0.00066479 0.0046259 -846.83
## - density     1 0.00106596 0.0050271 -839.35
## - polpc       1 0.00142823 0.0053894 -833.08
## - prbconv     1 0.00191423 0.0058754 -825.31
## - prbarr      1 0.00201116 0.0059723 -823.84
##
## Step:  AIC=-860.76
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     west + central + pctmin80_2 + wcon + wtuc + wtrd + wfir +
##     wser + wmfg + wfed + wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
```

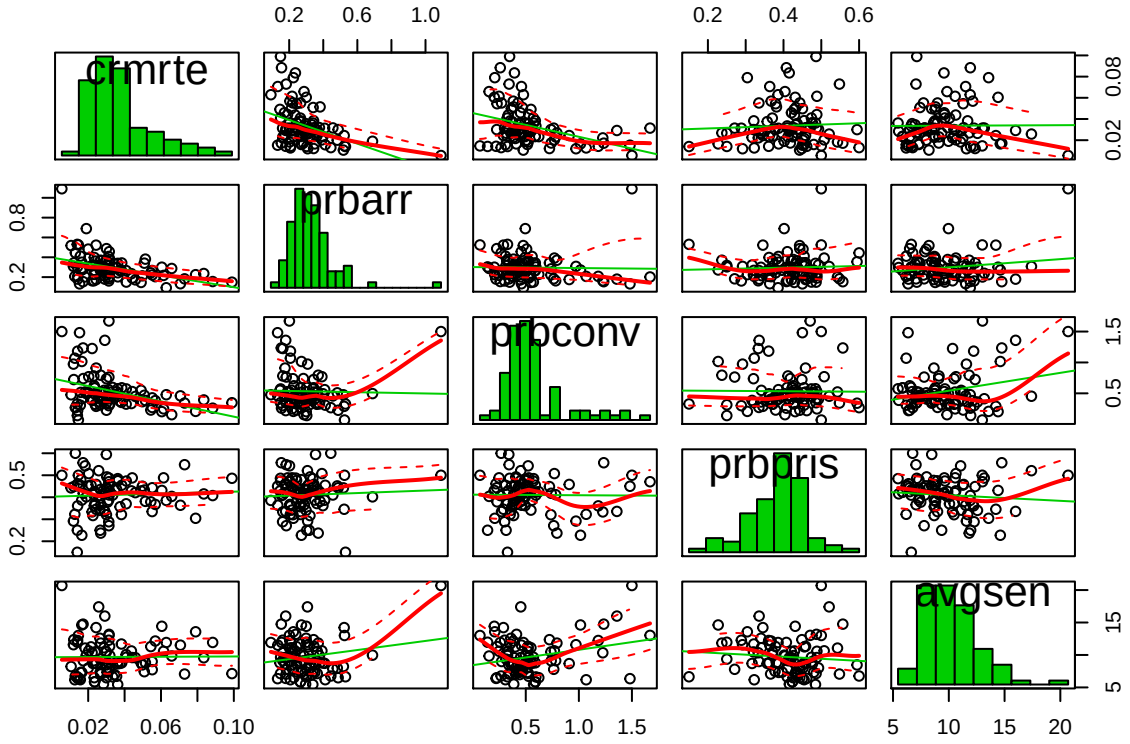
```
## - wmfg        1 0.00000543 0.0039680 -862.64
## - west        1 0.00002880 0.0039913 -862.11
## - wtrd        1 0.00003167 0.0039942 -862.04
## - wfir        1 0.00003352 0.0039961 -862.00
## - wtuc        1 0.00003486 0.0039974 -861.97
## - wloc        1 0.00006382 0.0040264 -861.32
## <none>                     0.0039625 -860.76
## - wcon        1 0.00009328 0.0040558 -860.67
## - wsta        1 0.00010533 0.0040679 -860.40
## - avgsen      1 0.00012422 0.0040868 -859.98
## - mix         1 0.00013912 0.0041017 -859.66
## + urban       1 0.00000141 0.0039611 -858.79
## + prbpris     1 0.00000056 0.0039620 -858.77
## - central     1 0.00022980 0.0041923 -857.69
## - wfed        1 0.00024004 0.0042026 -857.47
## - taxpc       1 0.00025970 0.0042222 -857.05
## - pctymle     1 0.00048823 0.0044508 -852.30
## - pctmin80_2  1 0.00064761 0.0046101 -849.14
## - wser        1 0.00066342 0.0046260 -848.83
## - polpc       1 0.00143320 0.0053957 -834.98
## - prbconv     1 0.00191282 0.0058754 -827.31
## - prbarr      1 0.00201270 0.0059752 -825.79
## - density     1 0.00244101 0.0064035 -819.56
##
## Step:  AIC=-862.64
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     west + central + pctmin80_2 + wcon + wtuc + wtrd + wfir +
##     wser + wfed + wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - west        1 0.00002951 0.0039975 -863.97
## - wtuc        1 0.00003039 0.0039984 -863.95
## - wtrd        1 0.00003621 0.0040042 -863.82
## - wfir        1 0.00004016 0.0040081 -863.73
## - wloc        1 0.00006236 0.0040303 -863.23
## <none>                     0.0039680 -862.64
## - wcon        1 0.00009684 0.0040648 -862.47
## - wsta        1 0.00010319 0.0040712 -862.33
## - avgsen      1 0.00011967 0.0040876 -861.96
## - mix         1 0.00013373 0.0041017 -861.65
## + wmfg        1 0.00000543 0.0039625 -860.76
## + prbpris     1 0.00000067 0.0039673 -860.65
## + urban       1 0.00000046 0.0039675 -860.65
## - central     1 0.00023125 0.0041992 -859.54
## - wfed        1 0.00023466 0.0042026 -859.47
## - taxpc       1 0.00025428 0.0042222 -859.05
## - pctymle     1 0.00048550 0.0044535 -854.25
## - pctmin80_2  1 0.00064346 0.0046114 -851.11
## - wser        1 0.00068473 0.0046527 -850.31
## - polpc       1 0.00143149 0.0053995 -836.91
## - prbconv     1 0.00191741 0.0058854 -829.16
## - prbarr      1 0.00201545 0.0059834 -827.67
## - density     1 0.00243902 0.0064070 -821.52
##
```

```
## Step:  AIC=-863.97
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     central + pctmin80_2 + wcon + wtuc + wtrd + wfir + wser +
##     wfed + wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - wtuc        1 0.00002118 0.0040186 -865.50
## - wtrd        1 0.00004158 0.0040391 -865.04
## - wfir        1 0.00005414 0.0040516 -864.76
## - wloc        1 0.00006977 0.0040672 -864.41
## <none>                     0.0039975 -863.97
## - wcon        1 0.00010225 0.0040997 -863.70
## - avgsen      1 0.00010410 0.0041016 -863.66
## - wsta        1 0.00011233 0.0041098 -863.48
## - mix         1 0.00014846 0.0041459 -862.69
## + west        1 0.00002951 0.0039680 -862.64
## + wmfg        1 0.00000613 0.0039913 -862.11
## + prbpris     1 0.00000015 0.0039973 -861.97
## + urban       1 0.00000002 0.0039974 -861.97
## - central     1 0.00022421 0.0042217 -861.06
## - wfed        1 0.00023950 0.0042370 -860.73
## - taxpc       1 0.00034686 0.0043443 -858.48
## - pctymle     1 0.00054578 0.0045433 -854.45
## - wser        1 0.00070387 0.0047013 -851.37
## - polpc       1 0.00140686 0.0054043 -838.83
## - prbconv     1 0.00189738 0.0058948 -831.01
## - pctmin80_2  1 0.00198607 0.0059835 -829.67
## - prbarr      1 0.00209576 0.0060932 -828.03
## - density     1 0.00241411 0.0064116 -823.45
##
## Step:  AIC=-865.5
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     central + pctmin80_2 + wcon + wtrd + wfir + wser + wfed +
##     wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - wtrd        1 0.00003839 0.0040570 -866.64
## - wfir        1 0.00005179 0.0040704 -866.34
## - wloc        1 0.00007230 0.0040909 -865.89
## - avgsen      1 0.00008975 0.0041084 -865.51
## <none>                     0.0040186 -865.50
## - wcon        1 0.00012008 0.0041387 -864.85
## - wsta        1 0.00014692 0.0041656 -864.26
## + wtuc        1 0.00002118 0.0039975 -863.97
## + west        1 0.00002029 0.0039984 -863.95
## - mix         1 0.00016504 0.0041837 -863.87
## + wmfg        1 0.00000167 0.0040170 -863.53
## + prbpris     1 0.00000161 0.0040170 -863.53
## + urban       1 0.00000001 0.0040186 -863.50
## - central     1 0.00022225 0.0042409 -862.65
## - wfed        1 0.00025189 0.0042705 -862.02
## - taxpc       1 0.00034215 0.0043608 -860.14
## - pctymle     1 0.00052759 0.0045462 -856.39
## - wser        1 0.00068491 0.0047036 -853.33
```

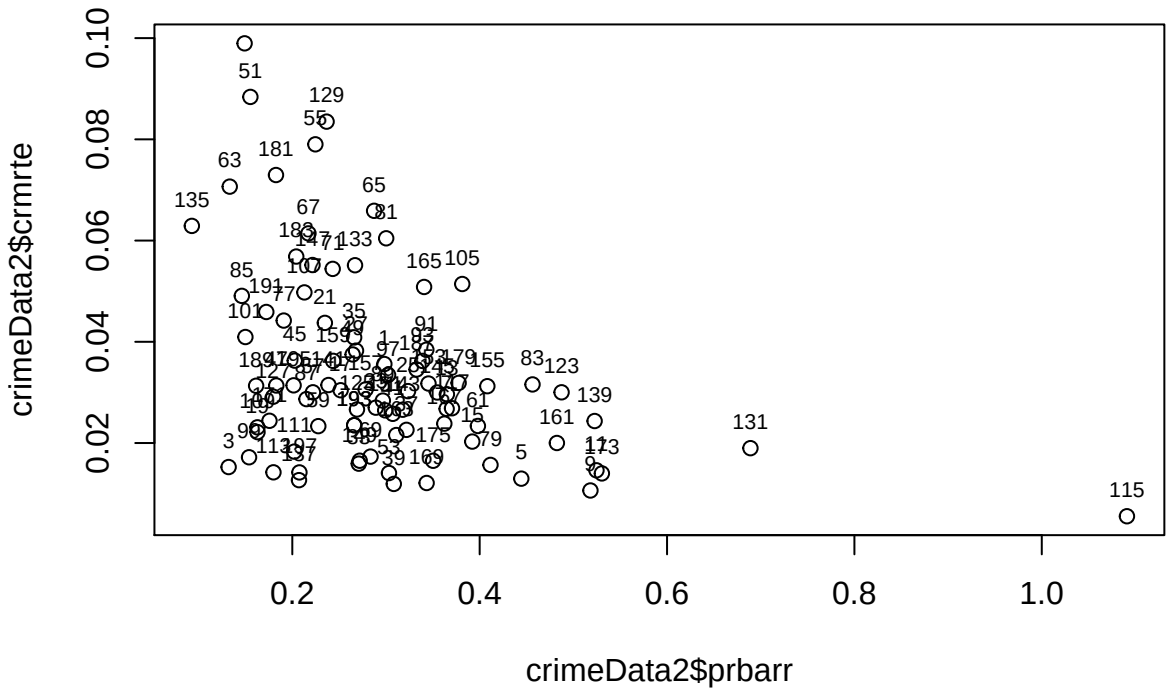
```
## - polpc       1 0.00141429 0.0054329 -840.36
## - prbconv     1 0.00193761 0.0059563 -832.08
## - pctmin80_2  1 0.00196914 0.0059878 -831.61
## - prbarr      1 0.00209117 0.0061098 -829.79
## - density     1 0.00247997 0.0064986 -824.24
##
## Step:  AIC=-866.64
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     central + pctmin80_2 + wcon + wfir + wser + wfed + wsta +
##     wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - wfir        1 0.00002783 0.0040849 -868.02
## - avgsen      1 0.00007995 0.0041370 -866.88
## <none>                     0.0040570 -866.64
## - wloc        1 0.00011554 0.0041726 -866.11
## - wcon        1 0.00013079 0.0041878 -865.78
## + wtrd        1 0.00003839 0.0040186 -865.50
## - mix         1 0.00015449 0.0042115 -865.28
## + west        1 0.00002515 0.0040319 -865.20
## + wtuc        1 0.00001798 0.0040391 -865.04
## + wmfg        1 0.00000495 0.0040521 -864.75
## + prbpris     1 0.00000401 0.0040530 -864.73
## + urban       1 0.00000008 0.0040570 -864.64
## - wsta        1 0.00020642 0.0042635 -864.17
## - central     1 0.00021107 0.0042681 -864.08
## - wfed        1 0.00030884 0.0043659 -862.04
## - taxpc       1 0.00034389 0.0044009 -861.32
## - pctymle     1 0.00050482 0.0045619 -858.09
## - wser        1 0.00069346 0.0047505 -854.44
## - polpc       1 0.00137631 0.0054333 -842.35
## - pctmin80_2  1 0.00195421 0.0060112 -833.25
## - prbconv     1 0.00202618 0.0060832 -832.18
## - prbarr      1 0.00206068 0.0061177 -831.67
## - density     1 0.00272130 0.0067783 -822.45
##
## Step:  AIC=-868.02
## crmrte ~ prbarr + prbconv + avgsen + polpc + density + taxpc +
##     central + pctmin80_2 + wcon + wser + wfed + wsta + wloc +
##     mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## - avgsen      1 0.00008720 0.0041721 -868.12
## <none>                     0.0040849 -868.02
## - wloc        1 0.00010126 0.0041861 -867.82
## - wcon        1 0.00011897 0.0042038 -867.44
## + west        1 0.00003315 0.0040517 -866.76
## + wfir        1 0.00002783 0.0040570 -866.64
## + wtuc        1 0.00001727 0.0040676 -866.41
## - mix         1 0.00017050 0.0042554 -866.34
## + wtrd        1 0.00001443 0.0040704 -866.34
## + wmfg        1 0.00000897 0.0040759 -866.22
## + prbpris     1 0.00000402 0.0040808 -866.11
## + urban       1 0.00000058 0.0040843 -866.04
```

```
## - central     1 0.00021259 0.0042975 -865.46
## - wsta        1 0.00024799 0.0043329 -864.72
## - wfed        1 0.00028360 0.0043685 -863.98
## - taxpc       1 0.00034784 0.0044327 -862.67
## - pctymle     1 0.00050849 0.0045934 -859.47
## - wser        1 0.00079940 0.0048843 -853.94
## - polpc       1 0.00140357 0.0054884 -843.44
## - pctmin80_2  1 0.00196639 0.0060513 -834.66
## - prbarr      1 0.00207748 0.0061624 -833.02
## - prbconv     1 0.00223143 0.0063163 -830.80
## - density     1 0.00271355 0.0067984 -824.18
##
## Step:  AIC=-868.12
## crmrte ~ prbarr + prbconv + polpc + density + taxpc + central +
##     pctmin80_2 + wcon + wser + wfed + wsta + wloc + mix + pctymle
##
##              Df  Sum of Sq       RSS     AIC
## <none>                     0.0041721 -868.12
## + avgsen      1 0.00008720 0.0040849 -868.02
## - wcon        1 0.00011252 0.0042846 -867.73
## - wloc        1 0.00012500 0.0042971 -867.47
## - mix         1 0.00014185 0.0043139 -867.11
## + wfir        1 0.00003508 0.0041370 -866.88
## + west        1 0.00002110 0.0041510 -866.58
## + prbpris     1 0.00000772 0.0041643 -866.29
## + wtrd        1 0.00000747 0.0041646 -866.29
## - central     1 0.00018414 0.0043562 -866.24
## + wtuc        1 0.00000498 0.0041671 -866.23
## + wmfg        1 0.00000485 0.0041672 -866.23
## + urban       1 0.00000018 0.0041719 -866.13
## - wfed        1 0.00024775 0.0044198 -864.93
## - wsta        1 0.00030568 0.0044777 -863.76
## - taxpc       1 0.00032746 0.0044995 -863.32
## - pctymle     1 0.00048270 0.0046548 -860.27
## - wser        1 0.00074689 0.0049190 -855.30
## - polpc       1 0.00131838 0.0054904 -845.41
## - pctmin80_2  1 0.00206603 0.0062381 -833.92
## - prbarr      1 0.00218706 0.0063591 -832.19
## - prbconv     1 0.00246792 0.0066400 -828.30
## - density     1 0.00266991 0.0068420 -825.60
##
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + polpc + density + taxpc +
##     central + pctmin80_2 + wcon + wser + wfed + wsta + wloc +
##     mix + pctymle, data = crimeData2)
##
## Coefficients:
## (Intercept)       prbarr      prbconv        polpc      density
##   1.451e-02   -5.446e-02   -2.157e-02    6.254e+00    5.411e-03
##       taxpc      central   pctmin80_2         wcon         wser
##   1.943e-04   -3.510e-03    3.312e-02    3.393e-05   -1.019e-04
##        wfed         wsta         wloc          mix      pctymle
##   4.631e-05   -4.779e-05    6.200e-05   -1.998e-02    1.173e-01
```

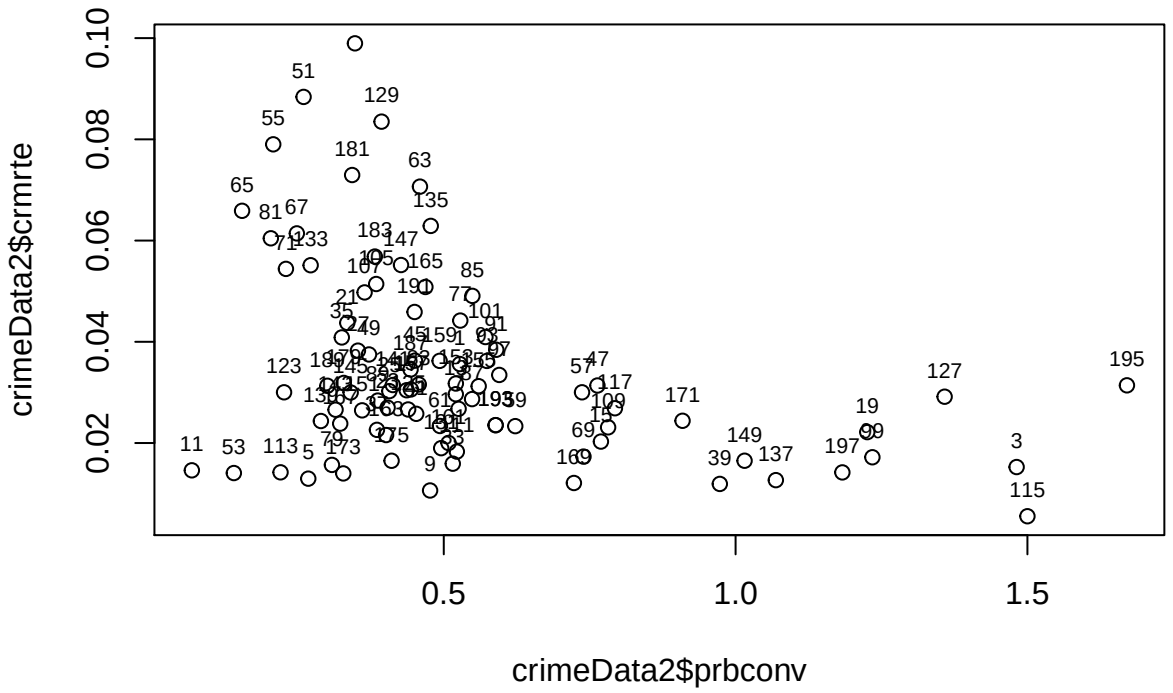
```
scatterplotMatrix(~ crmrte+prbarr+prbconv+prbpris+avgsen, data=crimeData2, diagonal="histogram")
```



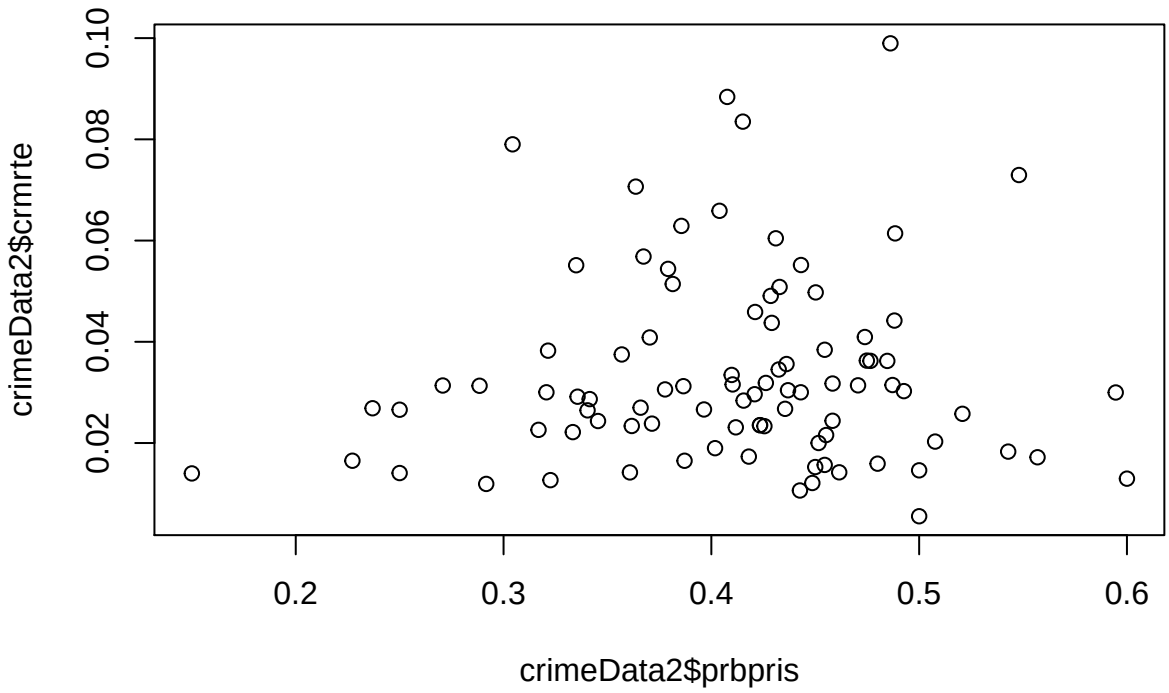

```
plot(crimeData2$prbarr, crimeData2$crmrte)
text(crimeData2$prbarr, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3)
```



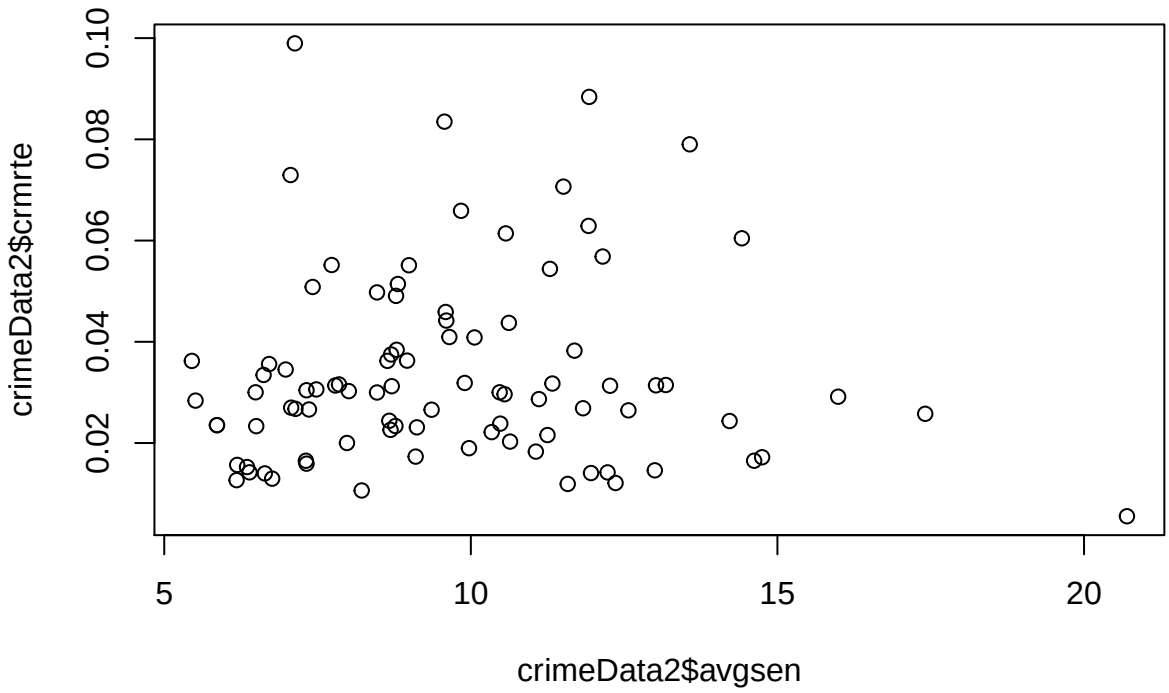

```
plot(crimeData2$prbconv, crimeData2$crmrte)
text(crimeData2$prbconv, crimeData2$crmrte, labels = crimeData2$county, cex=0.7, pos=3)
```




```
plot(crimeData2$prbpris, crimeData2$crmrte)
```




```
plot(crimeData2$avgsen, crimeData2$crmrte)
```




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

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

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

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

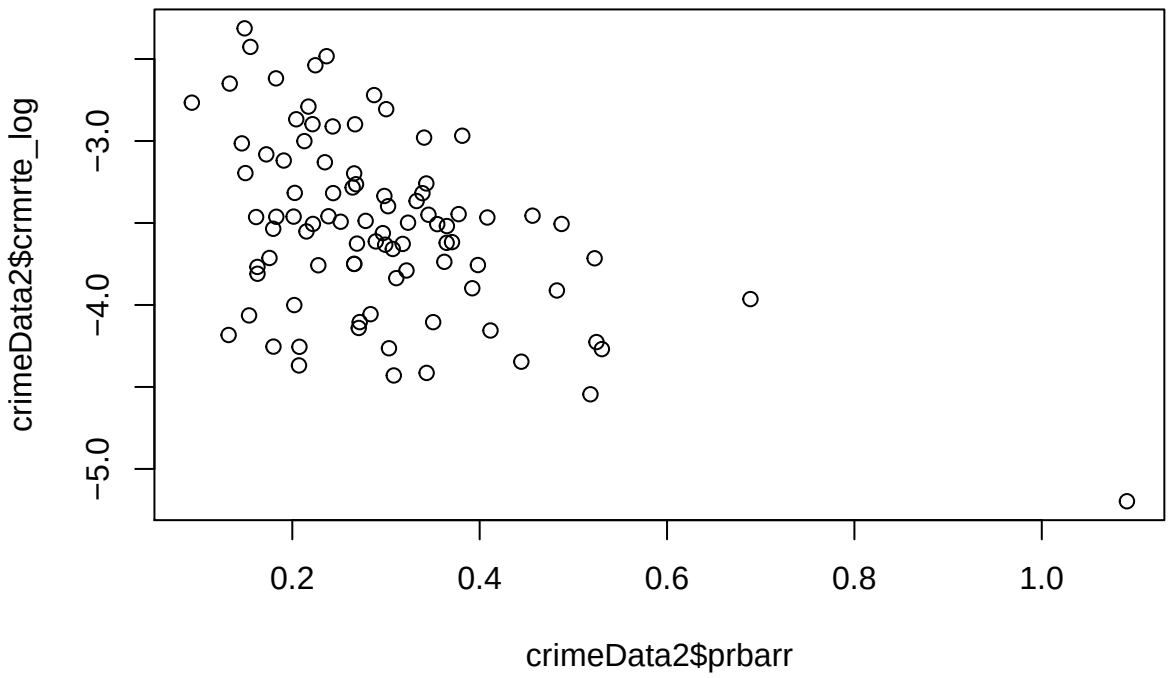
```
cor(crimeData2$crmrte, crimeData2$prbpris)
```

```
## [1] 0.05284061
```

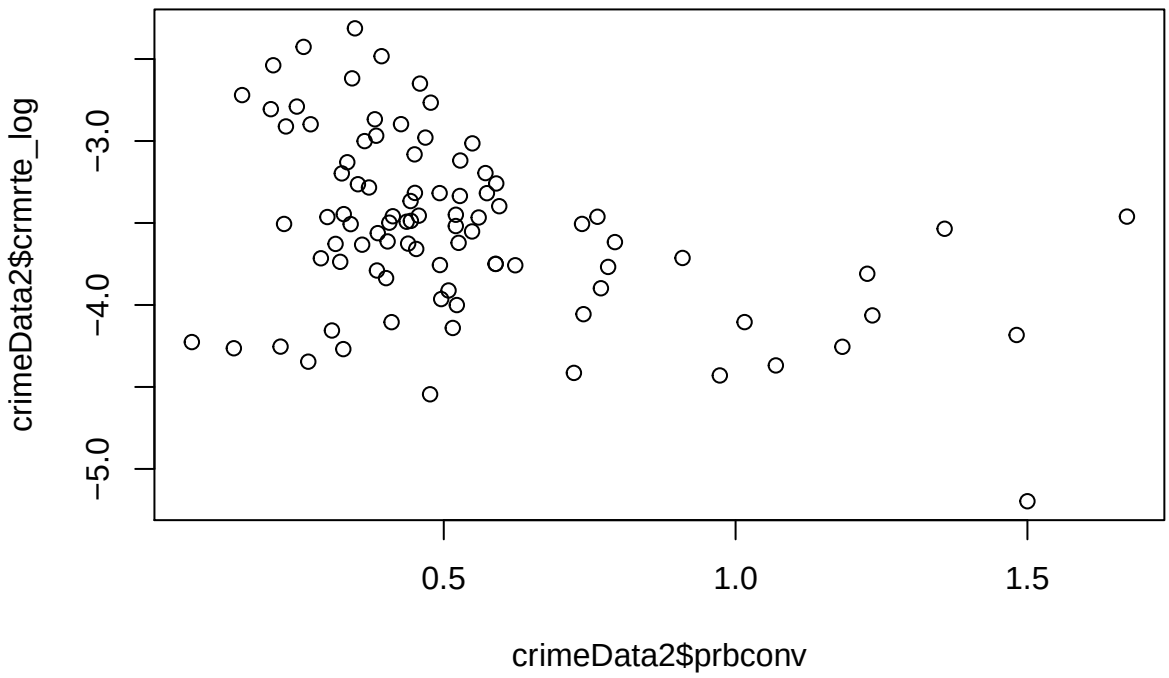
```
cor(crimeData2$crmrte, crimeData2$avgsen)
```

```
## [1] 0.007397583
```

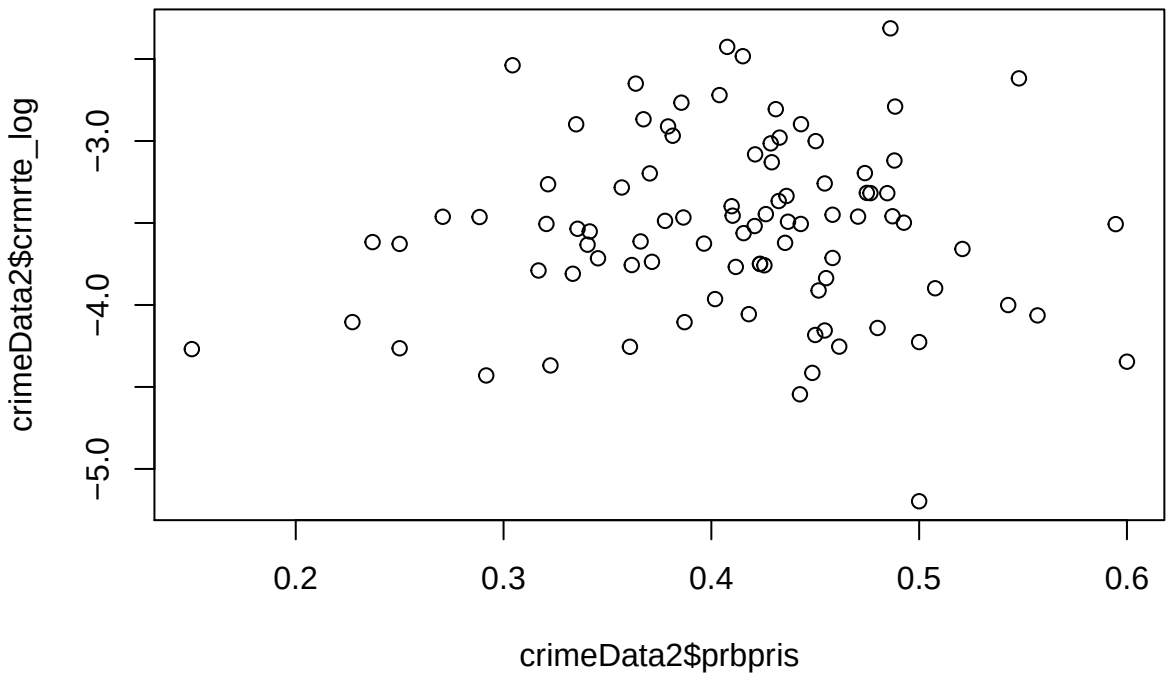
```
crimeData2$crmrte_log = log(crimeData2$crmrte)
plot(crimeData2$prbarr, crimeData2$crmrte_log)
```



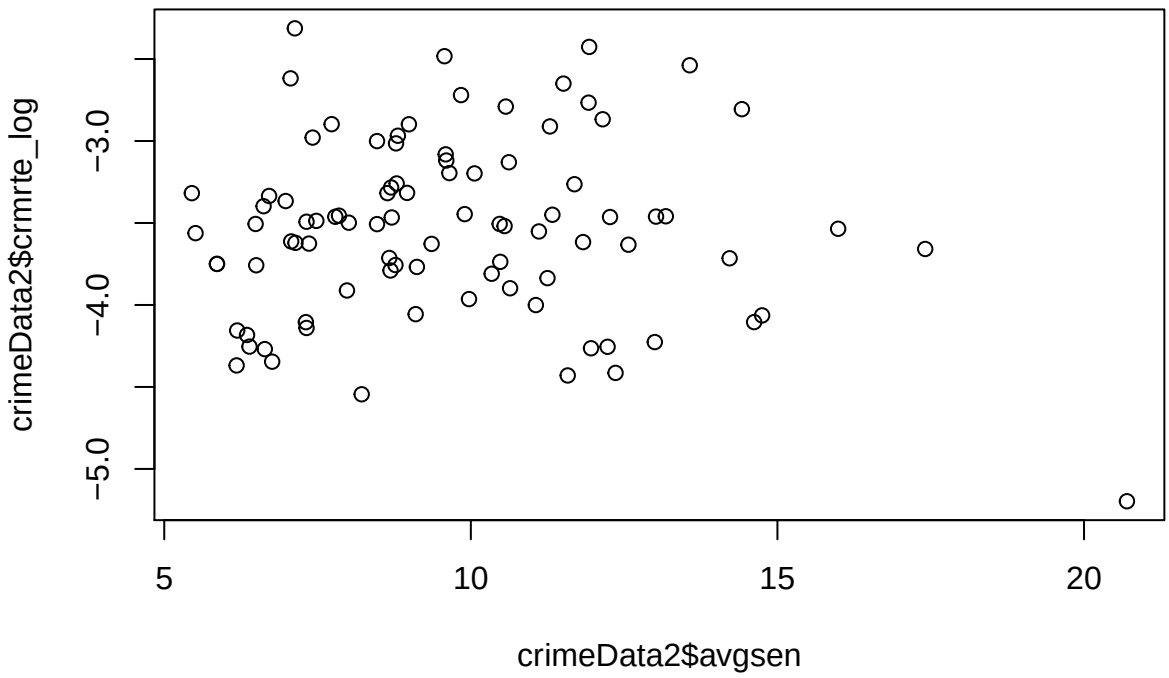

```
plot(crimeData2$prbconv, crimeData2$crmrte_log)
```




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




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




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

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

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

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

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

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

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

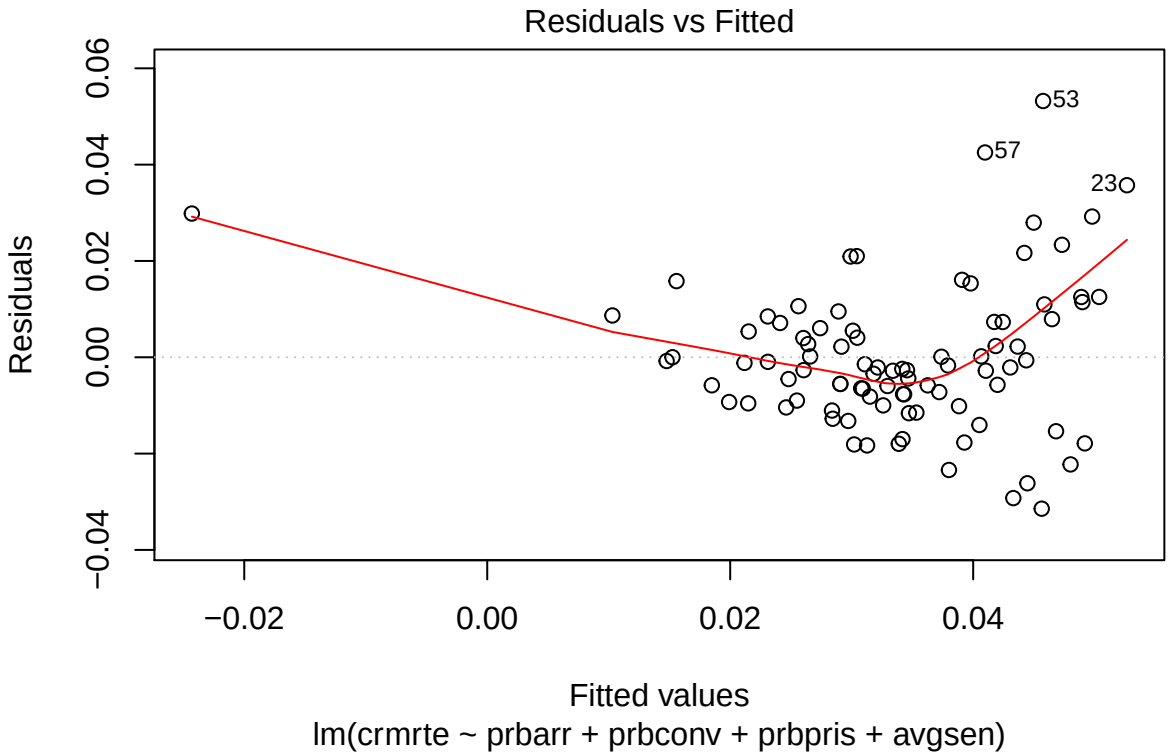
```
## [1] -0.07567514
```

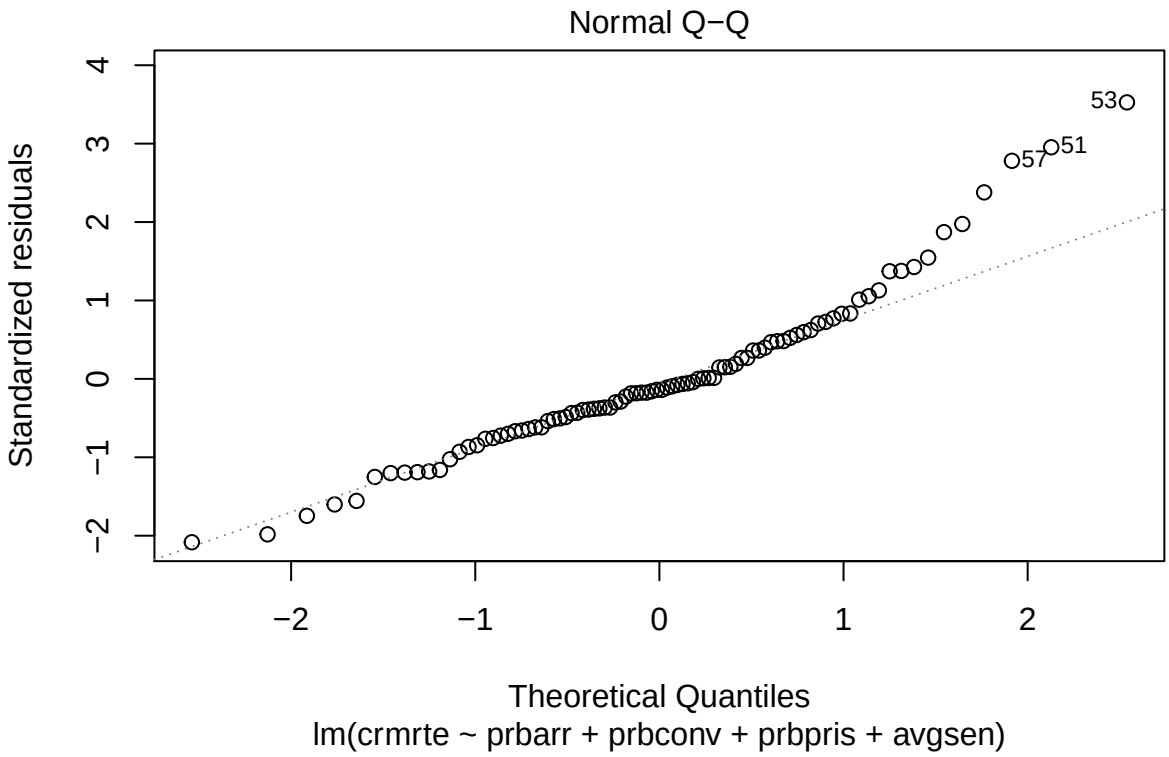
```
model1 <- lm(crmrte ~ prbarr+prbconv+prbpris+avgsen, data=crimeData2)
summary(model1)
```

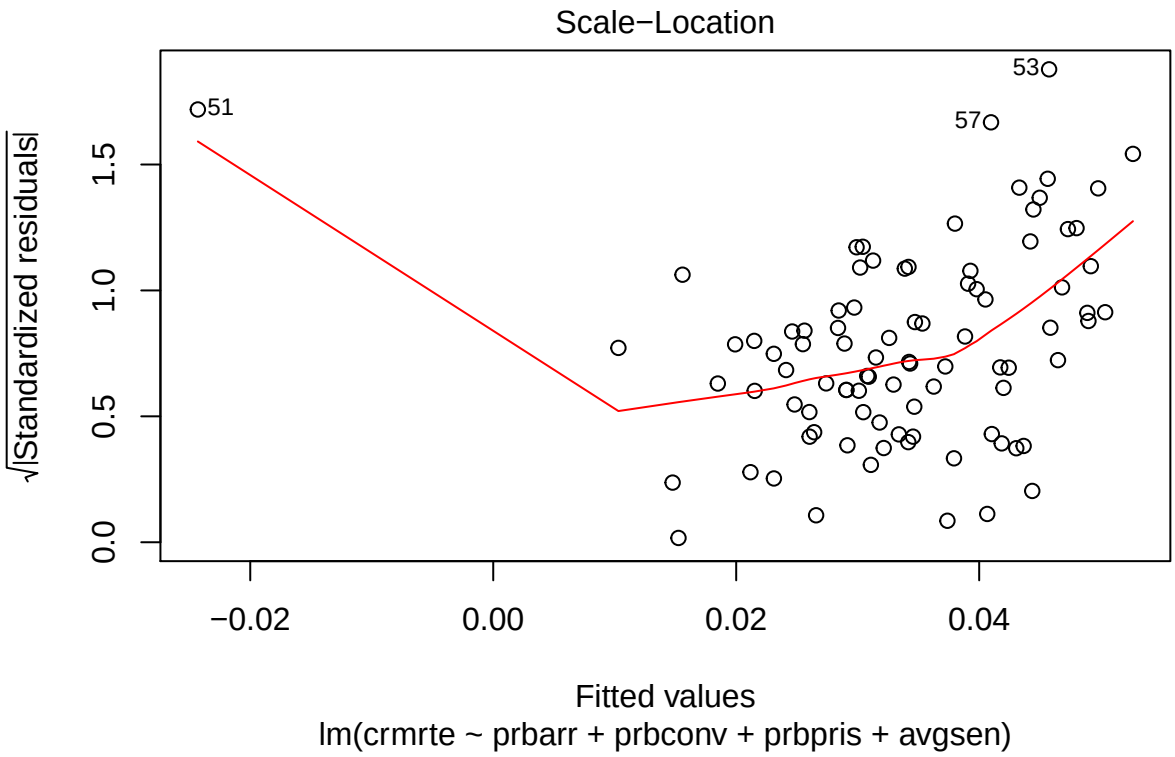
```
##
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + prbpris + avgsen, data = crimeData2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.031436 -0.009214 -0.002118  0.007333  0.053210
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

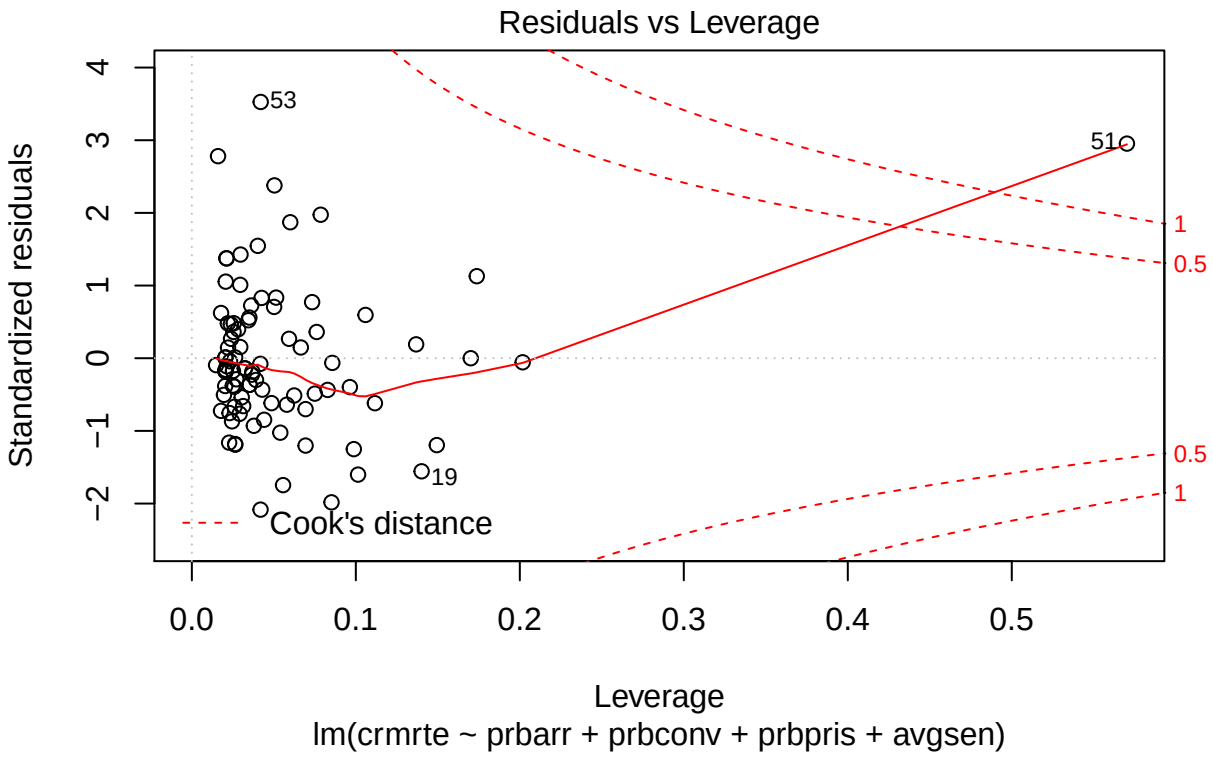
```
## (Intercept)  0.0442230  0.0108935   4.060 0.000109 ***
## prbarr      -0.0624726  0.0121348  -5.148 1.67e-06 ***
## prbconv     -0.0262359  0.0054372  -4.825 6.09e-06 ***
## prbpris      0.0208630  0.0204007   1.023 0.309370
## avgsen       0.0013786  0.0006124   2.251 0.026959 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01541 on 85 degrees of freedom
## Multiple R-squared:  0.3554, Adjusted R-squared:  0.3251
## F-statistic: 11.72 on 4 and 85 DF,  p-value: 1.265e-07
```

```
plot(model1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(crmrte ~ prbarr + prbconv + prbpris + avgsen)
53
51
57


Scale−Location
51
53
57
√|Standardized residuals|
0.0
0.5
1.0
1.5
−0.02
0.00
0.02
0.04
Fitted values
lm(crmrte ~ prbarr + prbconv + prbpris + avgsen)

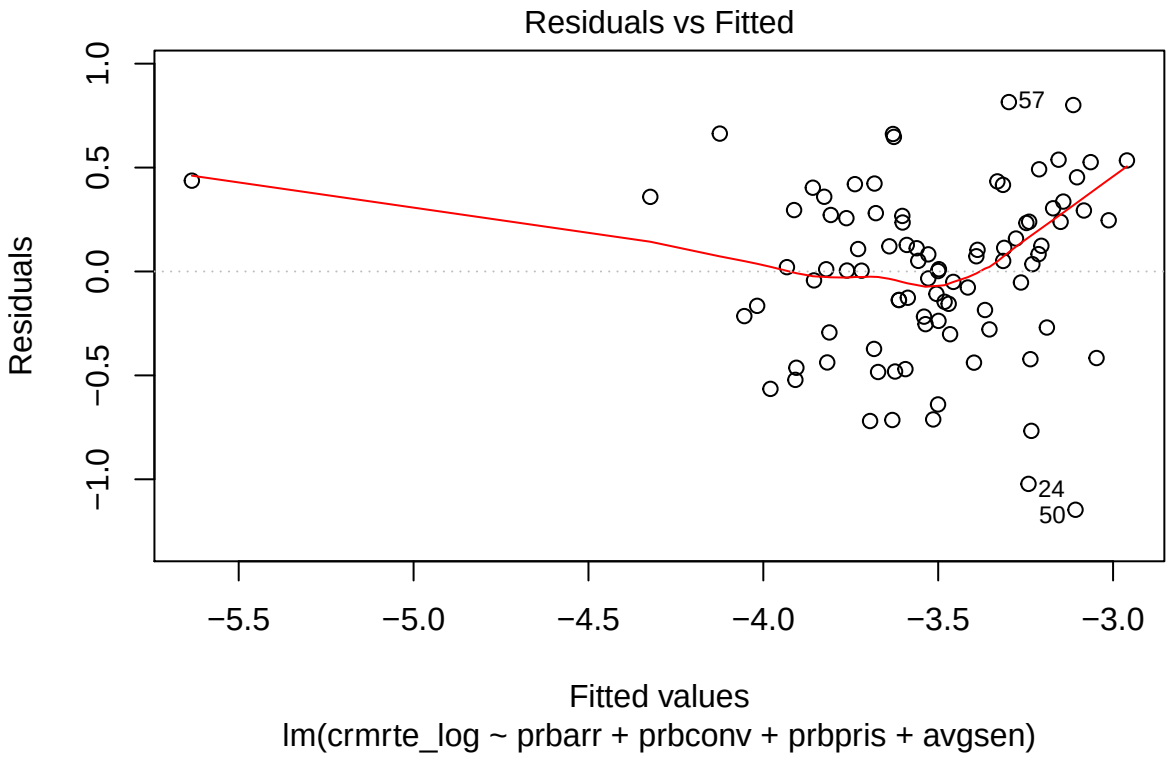


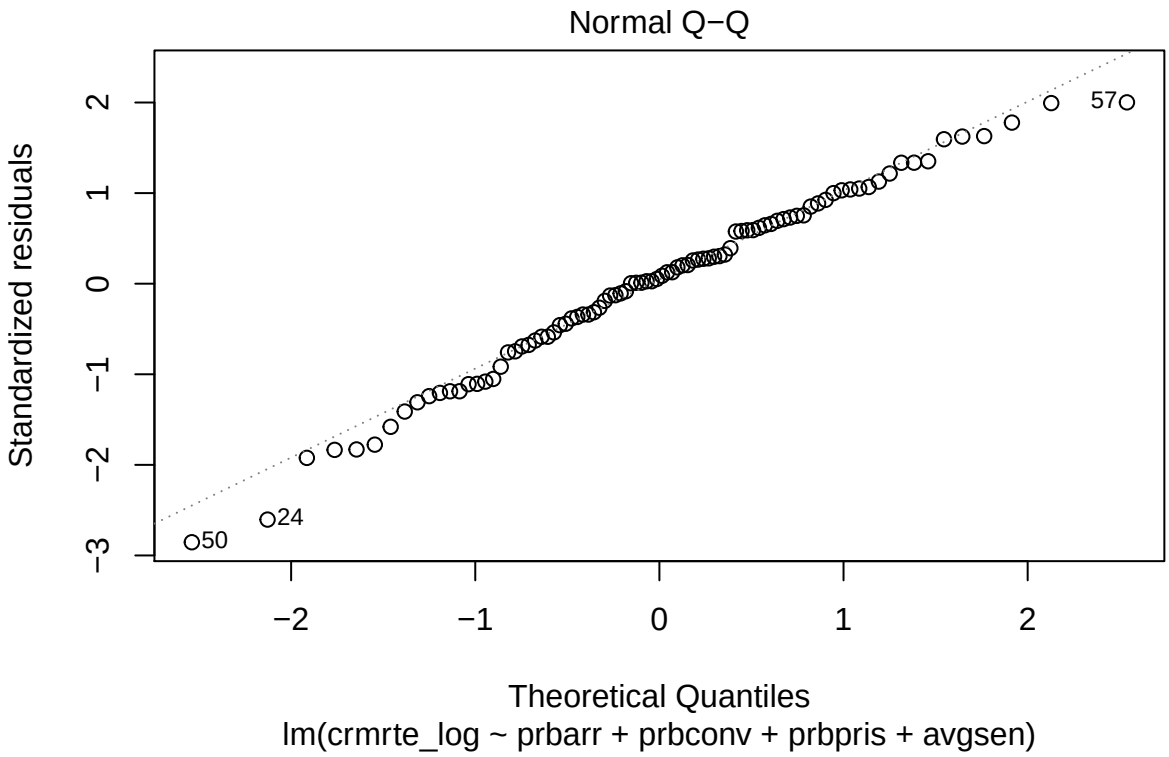

```
## Detect heteroskedasticity, var(y) seems correlated with E(y)
## Take the log of y
crimeData2$crmrte_log = log(crimeData2$crmrte)
model2 <- lm(crmrte_log ~ prbarr+prbconv+prbpris+avgsen, data=crimeData2)
summary(model2)
```

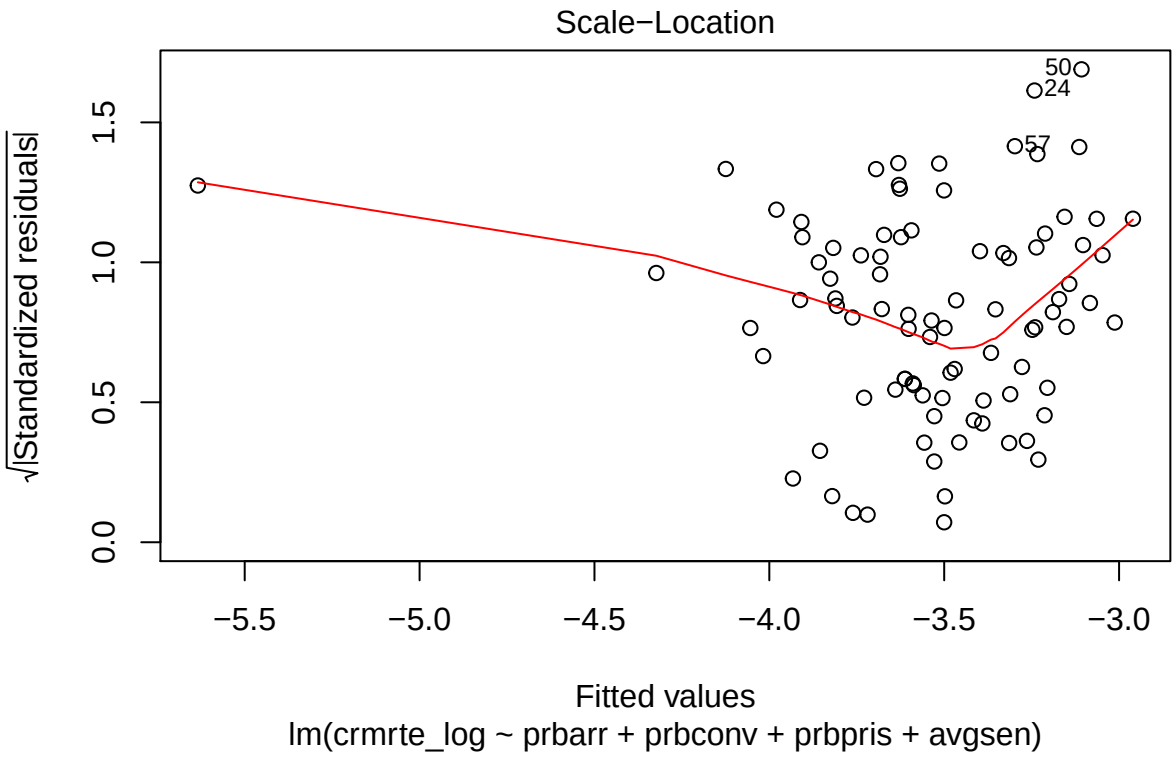
```
##
## Call:
## lm(formula = crmrte_log ~ prbarr + prbconv + prbpris + avgsen,
##     data = crimeData2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1466 -0.2495  0.0280  0.2785  0.8151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.92426    0.29003 -10.082 3.51e-16 ***
## prbarr      -2.09740    0.32308  -6.492 5.42e-09 ***
## prbconv     -0.79655    0.14476  -5.502 3.89e-07 ***
## prbpris      0.42628    0.54316   0.785    0.435
## avgsen       0.02705    0.01630   1.659    0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4104 on 85 degrees of freedom
```

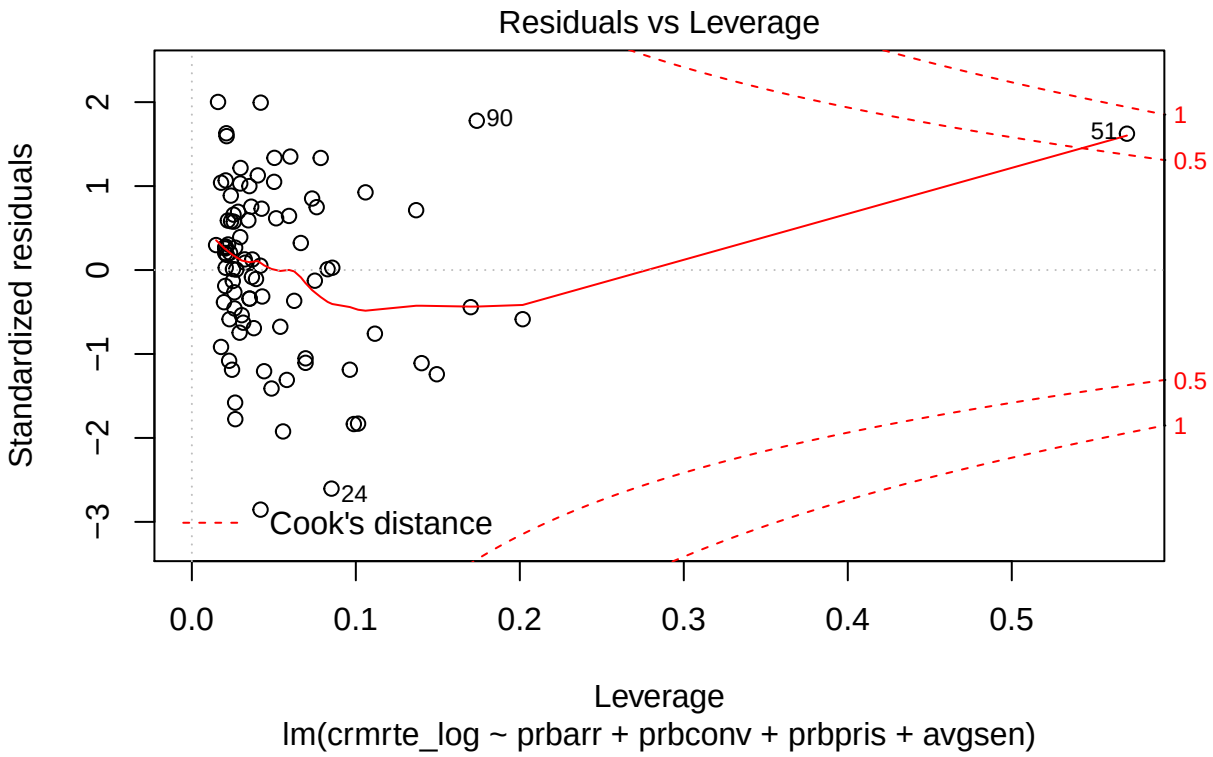
```
## Multiple R-squared:  0.4467, Adjusted R-squared:  0.4207
## F-statistic: 17.16 on 4 and 85 DF,  p-value: 2.382e-10
```

```
plot(model2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(crmrte_log ~ prbarr + prbconv + prbpris + avgsen)
57
24
50
−3
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
50
24
57
−5.5
−5.0
−4.5
−4.0
−3.5
−3.0
Fitted values
lm(crmrte_log ~ prbarr + prbconv + prbpris + avgsen)



Model Building 2

Model Building 3

Model Display

Omitted Variables