Lab 4

Kyle Redfield, Sai Ruvuru, Lucy Xie

December 12, 2017

```
library(car)
```

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

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

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

```
library(sandwich)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
library(datasets)

#Load dataset
#setwd('C:\\Users\\Lucy\\Documents\\Berkeley MIDS\\W203 Statistics for Data Science\\Lab 4')
data = read.csv("crime.csv")
```

Exploratory Data Analysis (EDA)

An initial exploratory analysis of the data is conducted to identify anomalies and potential transformations. The size of the dataset is large but relatively small at 90. The variables to begin the EDA is as following:

```
str(data)
```

```
## 'data.frame':    90 obs. of  26 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : num  0.528 1.481 0.268 0.525 0.477 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

A summary and snapshot of the data is taken to examine for anomalies:

```
# Examine size and shape of data
summary(data)
```

```
##        X              county           year        crmrte
##  Min.   : 1.00   Min.   :  1.0   Min.   :87   Min.   :0.005533
##  1st Qu.:23.25   1st Qu.: 51.5   1st Qu.:87   1st Qu.:0.020604
##  Median :45.50   Median :103.0   Median :87   Median :0.030002
##  Mean   :45.50   Mean   :100.6   Mean   :87   Mean   :0.033510
##  3rd Qu.:67.75   3rd Qu.:150.5   3rd Qu.:87   3rd Qu.:0.040249
##  Max.   :90.00   Max.   :197.0   Max.   :87   Max.   :0.098966
##      prbarr           prbconv           prbpris           avgsen
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.1500   Min.   : 5.380
##  1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.:0.3642   1st Qu.: 7.375
##  Median :0.27146   Median :0.45170   Median :0.4222   Median : 9.110
##  Mean   :0.29524   Mean   :0.55086   Mean   :0.4106   Mean   : 9.689
##  3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:0.4576   3rd Qu.:11.465
##  Max.   :1.09091   Max.   :2.12121   Max.   :0.6000   Max.   :20.700
##      polpc              density           taxpc             west
##  Min.   :0.0007459   Min.   :0.2034   Min.   : 25.69   Min.   :0.0000
##  1st Qu.:0.0012378   1st Qu.:0.5472   1st Qu.: 30.73   1st Qu.:0.0000
##  Median :0.0014897   Median :0.9792   Median : 34.92   Median :0.0000
##  Mean   :0.0017080   Mean   :1.4379   Mean   : 38.16   Mean   :0.2333
##  3rd Qu.:0.0018856   3rd Qu.:1.5693   3rd Qu.: 41.01   3rd Qu.:0.0000
##  Max.   :0.0090543   Max.   :8.8277   Max.   :119.76   Max.   :1.0000
##     central           urban            pctmin80           wcon
##  Min.   :0.0000   Min.   :0.00000   Min.   : 1.284   Min.   :193.6
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:10.024   1st Qu.:250.8
##  Median :0.0000   Median :0.00000   Median :24.852   Median :281.2
##  Mean   :0.3778   Mean   :0.08889   Mean   :25.713   Mean   :285.4
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0
##  Max.   :1.0000   Max.   :1.00000   Max.   :64.348   Max.   :436.8
##       wtuc            wtrd            wfir            wser
##  Min.   :187.6   Min.   :154.2   Min.   :170.9   Min.   : 133.0
##  1st Qu.:374.3   1st Qu.:190.7   1st Qu.:285.6   1st Qu.: 229.3
##  Median :404.8   Median :203.0   Median :317.1   Median : 253.1
##  Mean   :410.9   Mean   :210.9   Mean   :321.6   Mean   : 275.3
##  3rd Qu.:440.7   3rd Qu.:224.3   3rd Qu.:342.6   3rd Qu.: 277.6
##  Max.   :613.2   Max.   :354.7   Max.   :509.5   Max.   :2177.1
##       wmfg            wfed            wsta            wloc
##  Min.   :157.4   Min.   :326.1   Min.   :258.3   Min.   :239.2
##  1st Qu.:288.6   1st Qu.:398.8   1st Qu.:329.3   1st Qu.:297.2
##  Median :321.1   Median :448.9   Median :358.4   Median :307.6
##  Mean   :336.0   Mean   :442.6   Mean   :357.7   Mean   :312.3
##  3rd Qu.:359.9   3rd Qu.:478.3   3rd Qu.:383.2   3rd Qu.:328.8
##  Max.   :646.9   Max.   :598.0   Max.   :499.6   Max.   :388.1
##       mix             pctymle
##  Min.   :0.01961   Min.   :0.06216
##  1st Qu.:0.08060   1st Qu.:0.07437
##  Median :0.10095   Median :0.07770
##  Mean   :0.12905   Mean   :0.08403
##  3rd Qu.:0.15206   3rd Qu.:0.08352
##  Max.   :0.46512   Max.   :0.24871
```

```r
head(data)
```

```
##   X county year    crmrte   prbarr   prbconv  prbpris avgsen      polpc
## 1 1      1   87 0.0356036 0.298270 0.5275960 0.436170   6.71 0.00182786
## 2 2      3   87 0.0152532 0.132029 1.4814800 0.450000   6.35 0.00074588
## 3 3      5   87 0.0129603 0.444444 0.2678570 0.600000   6.76 0.00123431
## 4 4      7   87 0.0267532 0.364760 0.5254240 0.435484   7.14 0.00152994
## 5 5      9   87 0.0106232 0.518219 0.4765630 0.442623   8.22 0.00086018
## 6 6     11   87 0.0146067 0.524664 0.0683761 0.500000  13.00 0.00288203
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 1 2.4226327 30.99368    0       1     0 20.21870 281.4259 408.7245
## 2 1.0463320 26.89208    0       1     0  7.91632 255.1020 376.2542
## 3 0.4127659 34.81605    1       0     0  3.16053 226.9470 372.2084
## 4 0.4915572 42.94759    0       1     0 47.91610 375.2345 397.6901
## 5 0.5469484 28.05474    1       0     0  1.79619 292.3077 377.3126
## 6 0.6113361 35.22974    1       0     0  1.54070 250.4006 401.3378
##       wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 1 221.2701 453.1722 274.1775 334.54 477.58 292.09 311.91 0.08016878
## 2 196.0101 258.5650 192.3077 300.38 409.83 362.96 301.47 0.03022670
## 3 229.3209 305.9441 209.6972 237.65 358.98 331.53 281.37 0.46511629
## 4 191.1720 281.0651 256.7214 281.80 412.15 328.27 299.03 0.27362204
## 5 206.8215 289.3125 215.1933 290.89 377.35 367.23 342.82 0.06008584
## 6 187.8255 258.5650 237.1507 258.60 391.48 325.71 275.22 0.31952664
##      pctymle
## 1 0.07787097
## 2 0.08260694
## 3 0.07211538
## 4 0.07353726
## 5 0.07069755
## 6 0.09891920
```

While no missing values(NAs) are identified, the summary table shows that some of the inputs expressed as probabilities have values over 100%, which is impossible. The data is subset and excluded of any rows where “prbarr”, “prbconv” or “prbpris” is greater than 1. The size of the dataset decreases from 90 to 80 rows.

```r
# Filter out values of prbarr, prbconv and prbpris >1 and count
# the remaining rows
data_sub <- data[which((data$prbarr <= 1) & (data$prbconv <= 1) & (data$prbpris <= 1)), ]
nrow(data_sub)
```

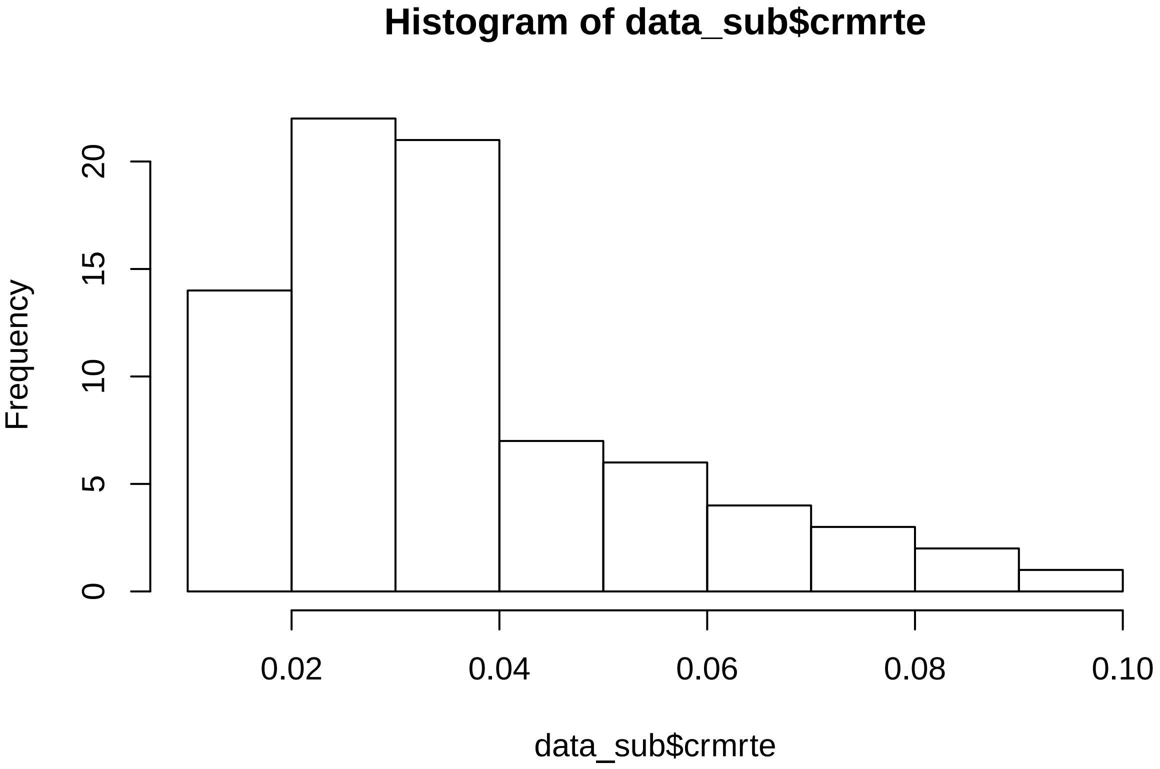
```
## [1] 80
```

```r
head(data_sub)
```

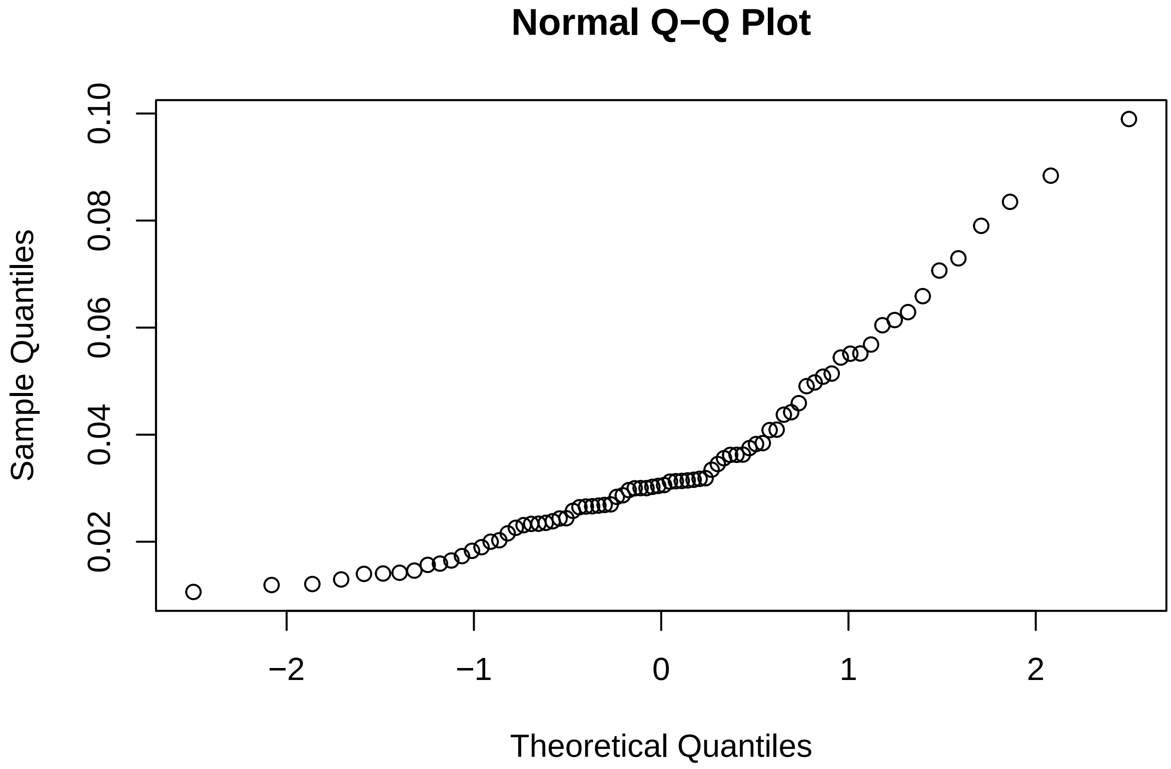
```
##   X county year    crmrte   prbarr   prbconv  prbpris avgsen      polpc
## 1 1      1   87 0.0356036 0.298270 0.5275960 0.436170   6.71 0.00182786
## 3 3      5   87 0.0129603 0.444444 0.2678570 0.600000   6.76 0.00123431
## 4 4      7   87 0.0267532 0.364760 0.5254240 0.435484   7.14 0.00152994
## 5 5      9   87 0.0106232 0.518219 0.4765630 0.442623   8.22 0.00086018
## 6 6     11   87 0.0146067 0.524664 0.0683761 0.500000  13.00 0.00288203
## 7 7     13   87 0.0296409 0.365004 0.5206070 0.420833  10.55 0.00133771
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 1 2.4226327 30.99368    0       1     0 20.21870 281.4259 408.7245
## 3 0.4127659 34.81605    1       0     0  3.16053 226.9470 372.2084
## 4 0.4915572 42.94759    0       1     0 47.91610 375.2345 397.6901
## 5 0.5469484 28.05474    1       0     0  1.79619 292.3077 377.3126
## 6 0.6113361 35.22974    1       0     0  1.54070 250.4006 401.3378
## 7 0.5169492 30.69649    0       0     0 32.17940 238.3064 366.3004
##       wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 1 221.2701 453.1722 274.1775 334.54 477.58 292.09 311.91 0.08016878
## 3 229.3209 305.9441 209.6972 237.65 358.98 331.53 281.37 0.46511629
## 4 191.1720 281.0651 256.7214 281.80 412.15 328.27 299.03 0.27362204
## 5 206.8215 289.3125 215.1933 290.89 377.35 367.23 342.82 0.06008584
## 6 187.8255 258.5650 237.1507 258.60 391.48 325.71 275.22 0.31952664
## 7 205.5358 310.1737 259.3391 303.42 449.84 350.72 283.76 0.15237226
##      pctymle
## 1 0.07787097
## 3 0.07211538
## 4 0.07353726
## 5 0.07069755
## 6 0.09891920
## 7 0.07073344
```

The dependent variable, crime rate(`crmrte`) is further explored.

```r
# Examine the dependent variable crmrte.
hist(data_sub$crmrte)
```



```
qqnorm(data_sub$crmrte)
```



The variable is positively skewed as further supported by the qqnorm plot. Since the expectation of the population model from background research supports the skewed distribution of crimes committed per person at a low mean(0.0355124) in relationship with the other normalized variables on the order of per capita, probability and per square mile, a log-log transformation will not be utilized in models.

Focusing on the potential independent variables, a preliminary correlation matrix is created to check for potential multicollinearity between inputs, as well as to identify inputs with the highest correlation to crmrte. A univariate model is also created for every input variable and compared the BIC scores.

```r
# Create correlation matrix
round(cor(data[c(-1:-3)]),2)
```

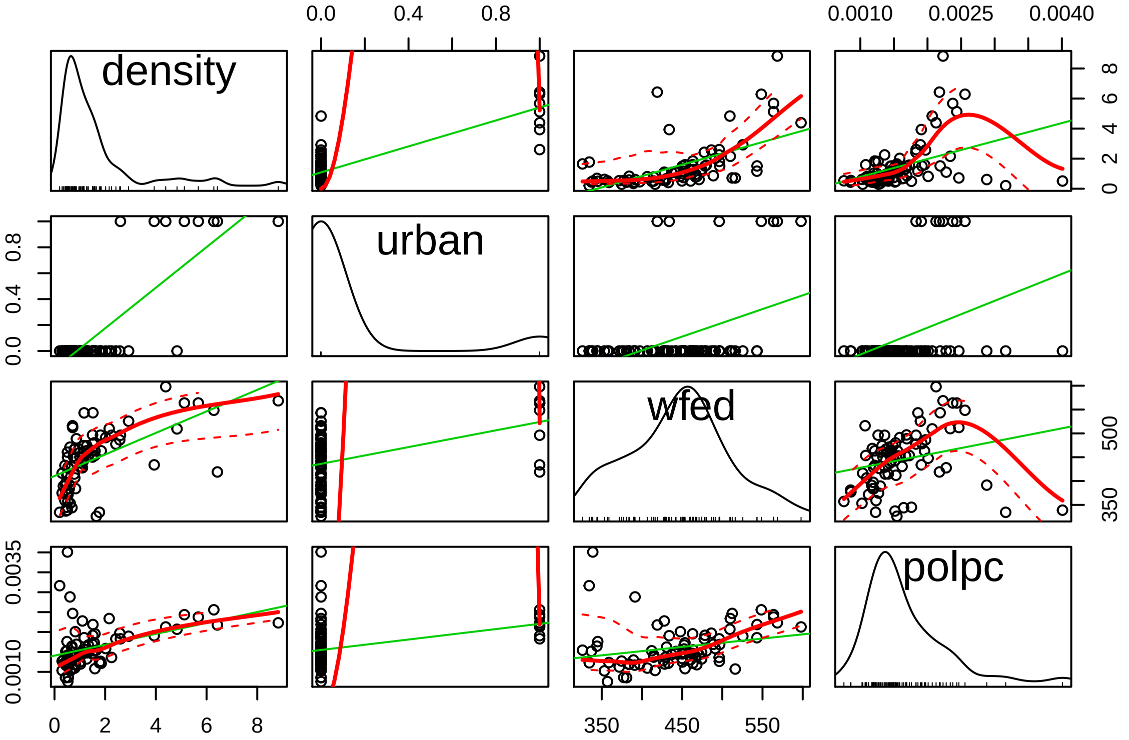
```
##          crmrte prbarr prbconv prbpris avgsen polpc density taxpc  west
## crmrte     1.00  -0.40   -0.39    0.05   0.02  0.17    0.73  0.45 -0.38
## prbarr    -0.40   1.00   -0.06    0.05   0.18  0.43   -0.30 -0.14  0.19
## prbconv   -0.39  -0.06    1.00    0.01   0.16  0.17   -0.23 -0.13  0.07
## prbpris    0.05   0.05    0.01    1.00  -0.09  0.05    0.07 -0.09 -0.04
## avgsen     0.02   0.18    0.16   -0.09   1.00  0.49    0.07  0.09  0.10
## polpc      0.17   0.43    0.17    0.05   0.49  1.00    0.16  0.28  0.14
## density    0.73  -0.30   -0.23    0.07   0.07  0.16    1.00  0.32 -0.19
## taxpc      0.45  -0.14   -0.13   -0.09   0.09  0.28    0.32  1.00 -0.17
## west      -0.38   0.19    0.07   -0.04   0.10  0.14   -0.19 -0.17  1.00
## central    0.17  -0.17   -0.05    0.16  -0.16 -0.05    0.36  0.03 -0.43
## urban      0.62  -0.21   -0.20    0.05   0.14  0.16    0.82  0.35 -0.08
## pctmin80   0.18   0.05    0.06    0.11  -0.17 -0.17   -0.07 -0.03 -0.62
## wcon       0.39  -0.25   -0.12   -0.06  -0.03 -0.02    0.45  0.26 -0.19
## wtuc       0.24  -0.07   -0.01    0.12   0.23  0.17    0.33  0.17  0.02
## wtrd       0.43  -0.10   -0.13    0.14   0.11  0.12    0.59  0.18 -0.19
## wfir       0.34  -0.17    0.03    0.03   0.18  0.20    0.55  0.13 -0.05
## wser      -0.05  -0.13    0.46    0.04  -0.15 -0.02    0.04  0.08 -0.06
## wmfg       0.35  -0.15    0.02    0.01   0.11  0.27    0.44  0.26 -0.01
## wfed       0.49  -0.21   -0.06    0.08   0.15  0.16    0.59  0.06 -0.21
## wsta       0.20  -0.16   -0.13   -0.03   0.13  0.05    0.22 -0.03 -0.08
## wloc       0.36  -0.02    0.05    0.08   0.15  0.39    0.46  0.22 -0.14
## mix       -0.13   0.41   -0.30    0.12  -0.14  0.02   -0.13 -0.04  0.00
## pctymle    0.29  -0.18   -0.16   -0.08   0.07  0.05    0.11 -0.09 -0.04
##          central urban pctmin80  wcon  wtuc  wtrd  wfir  wser  wmfg  wfed
## crmrte      0.17  0.62     0.18  0.39  0.24  0.43  0.34 -0.05  0.35  0.49
## prbarr     -0.17 -0.21     0.05 -0.25 -0.07 -0.10 -0.17 -0.13 -0.15 -0.21
## prbconv    -0.05 -0.20     0.06 -0.12 -0.01 -0.13  0.03  0.46  0.02 -0.06
## prbpris     0.16  0.05     0.11 -0.06  0.12  0.14  0.03  0.04  0.01  0.08
## avgsen     -0.16  0.14    -0.17 -0.03  0.23  0.11  0.18 -0.15  0.11  0.15
## polpc      -0.05  0.16    -0.17 -0.02  0.17  0.12  0.20 -0.02  0.27  0.16
## density     0.36  0.82    -0.07  0.45  0.33  0.59  0.55  0.04  0.44  0.59
## taxpc       0.03  0.35    -0.03  0.26  0.17  0.18  0.13  0.08  0.26  0.06
## west       -0.43 -0.08    -0.62 -0.19  0.02 -0.19 -0.05 -0.06 -0.01 -0.21
## central     1.00  0.16    -0.05  0.40  0.19  0.39  0.29  0.19  0.17  0.35
## urban       0.16  1.00     0.02  0.32  0.23  0.43  0.40  0.06  0.40  0.43
## pctmin80   -0.05  0.02     1.00 -0.11 -0.19 -0.06 -0.08  0.20 -0.12  0.03
## wcon        0.40  0.32    -0.11  1.00  0.41  0.56  0.49 -0.01  0.35  0.51
## wtuc        0.19  0.23    -0.19  0.41  1.00  0.35  0.33 -0.02  0.47  0.40
## wtrd        0.39  0.43    -0.06  0.56  0.35  1.00  0.67 -0.02  0.37  0.64
## wfir        0.29  0.40    -0.08  0.49  0.33  0.67  1.00  0.01  0.50  0.62
## wser        0.19  0.06     0.20 -0.01 -0.02 -0.02  0.01  1.00  0.01  0.02
## wmfg        0.17  0.40    -0.12  0.35  0.47  0.37  0.50  0.01  1.00  0.52
## wfed        0.35  0.43     0.03  0.51  0.40  0.64  0.62  0.02  0.52  1.00
## wsta        0.09  0.30     0.09 -0.02 -0.15  0.01  0.24  0.04  0.05  0.19
## wloc        0.33  0.34    -0.11  0.52  0.33  0.58  0.55  0.08  0.45  0.52
## mix        -0.09 -0.06     0.20 -0.20 -0.25 -0.13 -0.21 -0.17 -0.34 -0.31
## pctymle    -0.10  0.09    -0.02 -0.02 -0.10 -0.11  0.01 -0.04  0.02 -0.06
##           wsta  wloc   mix pctymle
## crmrte    0.20  0.36 -0.13    0.29
## prbarr   -0.16 -0.02  0.41   -0.18
## prbconv  -0.13  0.05 -0.30   -0.16
## prbpris  -0.03  0.08  0.12   -0.08
## avgsen    0.13  0.15 -0.14    0.07
## polpc     0.05  0.39  0.02    0.05
## density   0.22  0.46 -0.13    0.11
## taxpc    -0.03  0.22 -0.04   -0.09
## west     -0.08 -0.14  0.00   -0.04
## central   0.09  0.33 -0.09   -0.10
## urban     0.30  0.34 -0.06    0.09
## pctmin80  0.09 -0.11  0.20   -0.02
## wcon     -0.02  0.52 -0.20   -0.02
## wtuc     -0.15  0.33 -0.25   -0.10
## wtrd      0.01  0.58 -0.13   -0.11
## wfir      0.24  0.55 -0.21    0.01
## wser      0.04  0.08 -0.17   -0.04
## wmfg      0.05  0.45 -0.34    0.02
## wfed      0.19  0.52 -0.31   -0.06
## wsta      1.00  0.16 -0.08    0.22
## wloc      0.16  1.00 -0.25    0.00
## mix      -0.08 -0.25  1.00   -0.09
## pctymle   0.22  0.00 -0.09    1.00
```

```
# Print BIC score for linear model between each input and crime rate
n = 1
for (i in data) {
  #if (is.numeric(i[1])) {
    (model1 = lm(crmrte ~ i, data=data))
     print(colnames(data)[n])
     print(BIC(model1))
     last_BIC <- BIC(model1)
  #}
  n = n + 1
}
```

```
## [1] "X"
## [1] -446.5922
## [1] "county"
## [1] -446.6069
## [1] "year"
## [1] -451.0678
## [1] "crmrte"
## [1] -6978.087
## [1] "prbarr"
## [1] -461.8588
## [1] "prbconv"
## [1] -461.0857
## [1] "prbpris"
## [1] -446.7756
## [1] "avgsen"
## [1] -446.6033
## [1] "polpc"
## [1] -449.1224
## [1] "density"
## [1] -514.4552
## [1] "taxpc"
## [1] -466.8024
## [1] "west"
## [1] -460.6308
## [1] "central"
## [1] -449.0792
## [1] "urban"
## [1] -489.3452
## [1] "pctmin80"
## [1] -449.5878
## [1] "wcon"
## [1] -461.6638
## [1] "wtuc"
## [1] -451.7255
## [1] "wtrd"
## [1] -464.7055
## [1] "wfir"
## [1] -457.3512
## [1] "wser"
## [1] -446.8124
## [1] "wmfg"
## [1] -458.5138
## [1] "wfed"
## [1] -471.2695
## [1] "wsta"
## [1] -450.2363
## [1] "wloc"
## [1] -459.0475
## [1] "mix"
## [1] -448.15
## [1] "pctymle"
## [1] -454.4937
```

Both methods show that population density, urban indicator and federal wage have the highest individual influence on crime rate. In addition, examined police per capita is explored as a potential covariate to the model.

```r
# Examine population density, urban indicator, federal wage and police per capita
scatterplotMatrix(data_sub[ , c("density","urban","wfed","polpc")])
```



The positively skewed distributions of population density and police per capita indicates that lognormal transformation of the variables is needed.

Model 1

As identified in the EDA, the top three variables with the highest influence on crime rate were population density, urban indicator and federal wage. The urban indicator from the first model will be eliminated due to its collinearity with population density; by definition, regions classified as a SMSA have high population densities. The federal wage is also excluded due to lack of practical signifiance in relation to crime rate. Therefore, the first model examines the relationship between crime rate and population density, which is first transformed.

```
# Create model of crime rate based on log(density)
(model1 = lm(crmrte ~ log(density), data=data_sub))
```

```
##
## Call:
## lm(formula = crmrte ~ log(density), data = data_sub)
##
## Coefficients:
##  (Intercept)  log(density)
##      0.03440       0.01648
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = crmrte ~ log(density), data = data_sub)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.019728 -0.010660 -0.002675  0.007347  0.055666
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.034399   0.001546  22.243  < 2e-16 ***
## log(density) 0.016481   0.001957   8.421 1.44e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01378 on 78 degrees of freedom
## Multiple R-squared:  0.4762, Adjusted R-squared:  0.4695
## F-statistic: 70.91 on 1 and 78 DF,  p-value: 1.437e-12
```

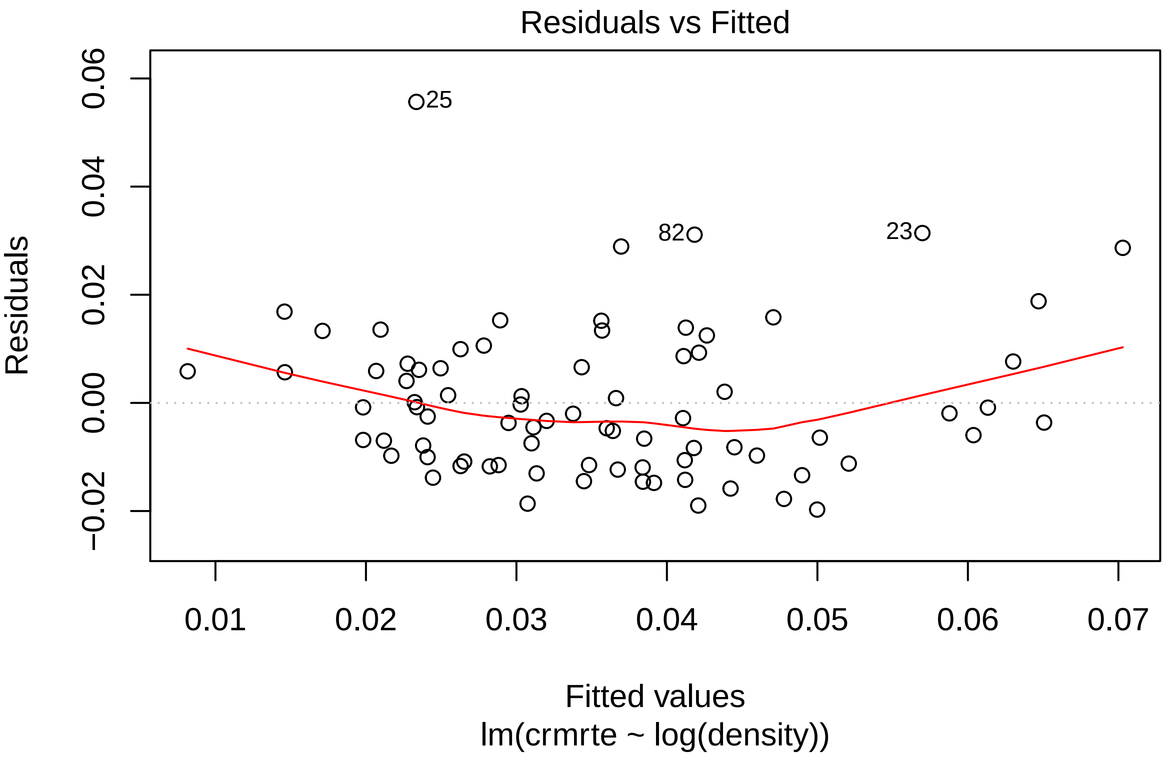
However, before considering this effect to be statistically appropriate, it must be first evaluated whether this model meets all OLS assumptions.

First, the model meets the assumption of a linear relationship since the model is a linear combination of variables.

Second, the data in the model is from a dataset about little is known. From references in the codebook, the data is exclusively from North Carolina, suggesting that any extrapolation of the population model from this data cannot be accurately performed. However, there is no evidence that the data was not collected randomly from within North Carolina. In fact, if the data is exclusively from North Carolinian counties, it represents 90% of the counties in North Carolina as of 1987. As long as the remaining 10 counties were excluded in a non-systemic fashion, we can assume the data is randomly collected but only from within North Carolina.

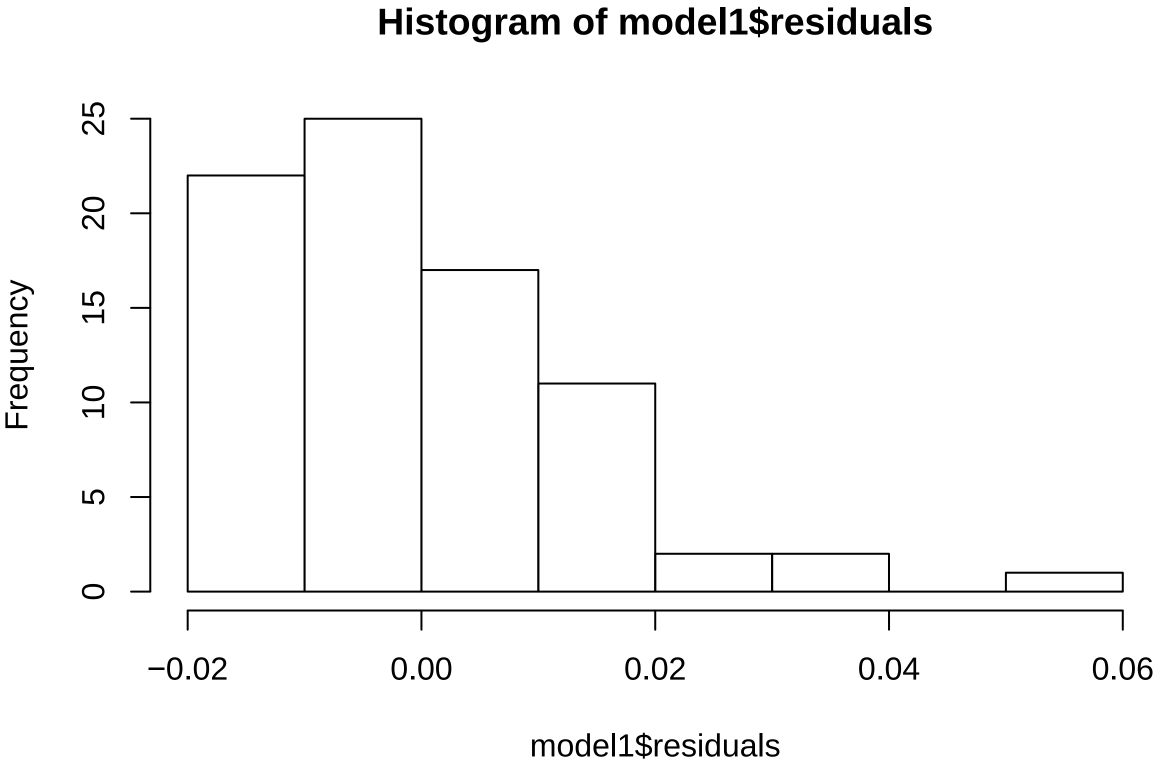
Third, because there is only one input variable to this model, there is no risk of multicollinearity between inputs. This is mitigated by eliminating the urban indicator as a model input.

```r
# Check for zero-conditional mean
# Check residuals vs. fitted plot
plot(model1, which=1)
```



From the residiuals vs. fitted values plot above, is can be seen that the values for residuals decrease, then increase along the fitted values axis. This is evidence that the zero-conditional mean assumption is violated. Furthermore, since the band of residuals are not evenly disributed, heteroskedasticity is indicated. However, this is further explored due to the relatively small sample size.

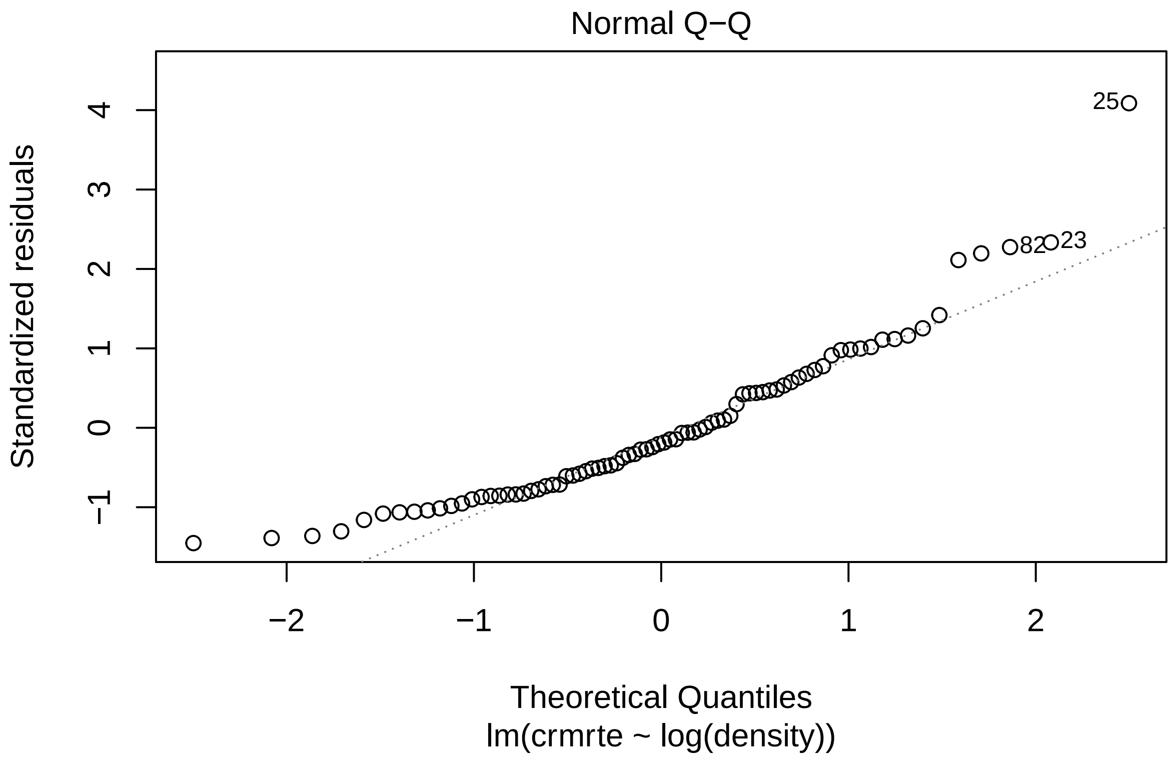
```
# Check for normality of errors
# Visualize distribution of residuals
hist(model1$residuals)
```



```
# Conduct Shapiro Test
shapiro.test(model1$residuals)
```

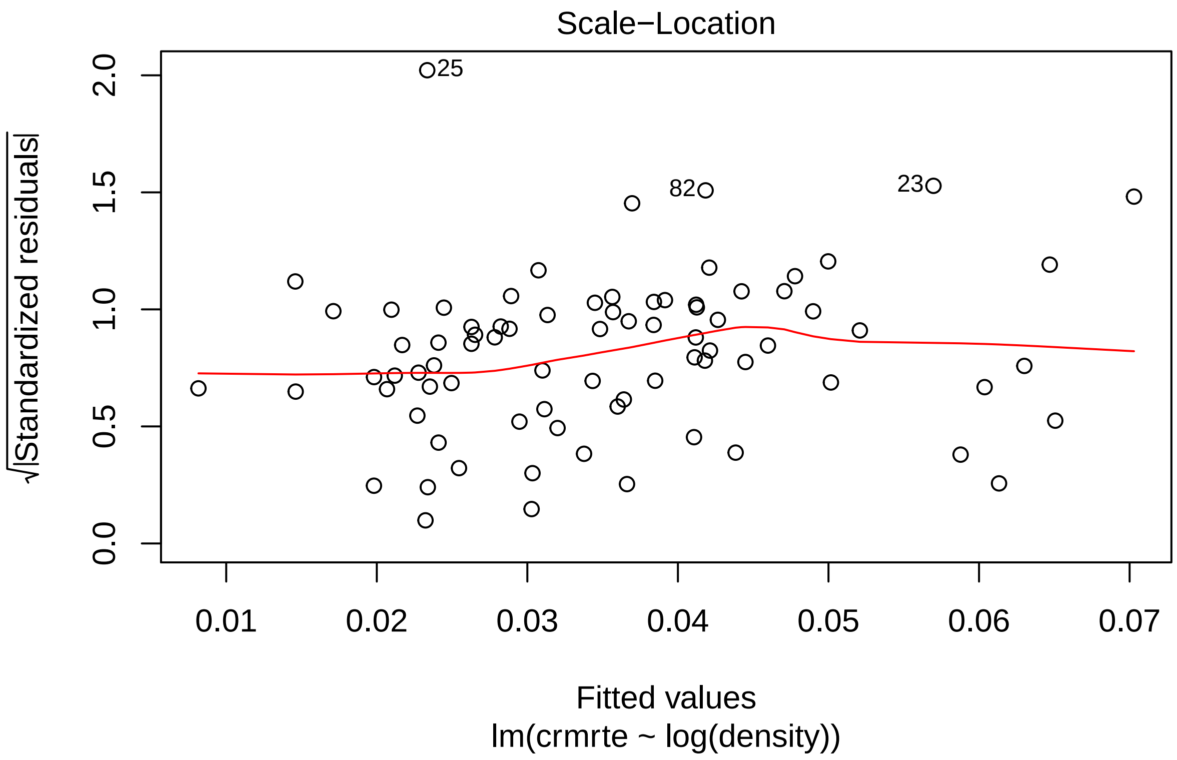
```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.91426, p-value = 5.149e-05
```

```
# Check normal Q-Q plot
plot(model1, which=2)
```



Both the histogram of residuals and Shapiro test suggest that the residuals are not normally distributed. The null hypothesis of the Shapiro states that the array is distributed normally. The p value is less than .01, suggesting that the null hypothesis is to be rejected. Finally, the residuals deviate from the diagonal line at lower and higher theoretical quantiles, which further indicates non-normality.

```r
# Check for homoskedasticity
# Check standardized residual plot
plot(model1, which=3)
```



```
# Conduct Breusch-Pagan Test
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 0.47749, df = 1, p-value = 0.4896
```

The above plot of residuals vs. fitted values shows a slight increase in the variance of residuals as fitted values increase. The standardized residuals plot confirms the increasing trend, but note that it decreases slightly at the high end of the fitted values axis, where there are much fewer data points. Therefore, a more robust method such as the Breusch-Pagan test must be used. The null hypothesis for the BP test is homoskedasticity. With this p-value, we fail to reject the null; therefore, homoskedasticity can be assumed.

Model 2

```
(model2 = lm(crmrte ~ log(density) + prbarr + pctmin80 + log(polpc), data = data_sub))
```

```
##
## Call:
## lm(formula = crmrte ~ log(density) + prbarr + pctmin80 + log(polpc),
##     data = data_sub)
##
## Coefficients:
##  (Intercept)  log(density)        prbarr      pctmin80    log(polpc)
##    0.2018744     0.0109015    -0.0517965     0.0004289     0.0251557
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = crmrte ~ log(density) + prbarr + pctmin80 + log(polpc),
##     data = data_sub)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.014951 -0.005980 -0.001534  0.004748  0.032135
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.019e-01  2.544e-02   7.935 1.58e-11 ***
## log(density)  1.090e-02  1.637e-03   6.661 3.99e-09 ***
## prbarr       -5.180e-02  1.099e-02  -4.713 1.10e-05 ***
## pctmin80      4.289e-04  6.449e-05   6.651 4.16e-09 ***
## log(polpc)    2.516e-02  3.883e-03   6.478 8.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009323 on 75 degrees of freedom
## Multiple R-squared:  0.7695, Adjusted R-squared:  0.7572
## F-statistic: 62.59 on 4 and 75 DF,  p-value: < 2.2e-16
```

Because population density is likely not the only factor that influences the crime rate of an area, additional variables are included in the specification of Model 2, listed above as the effect of population density, police per capita, the probability of arrest, and the percent of minorities in the county in 1980 on the crime rate.

These variables are chosen since, - Police per capita should be closely related to crime since it is the primary deterrant of crime - The probability of arrest is a proxy for how averse people are to committing crime in that community. The crime rate should decrease as people's aversion to arrest increases. - Previous studies have drawn a link between the presence of minority populations and the crime rate. The inclusion of this factor is consistent with those studies.

It can be seen that all of these variables have a statistically significant effect on the crime rate. Further the adjusted Rsquared has increased to .6552 from the value of .5243 in model 1. As a result, it can be determined that these variables increase the predictive power of the model over the decreases in parsmiony.

In the tests below, we see no major deviations from the diagnostics we saw in Model 1:

```r
vif(model2)
```

```
## log(density)       prbarr     pctmin80   log(polpc)
##     1.528103     1.308020     1.062402     1.224635
```

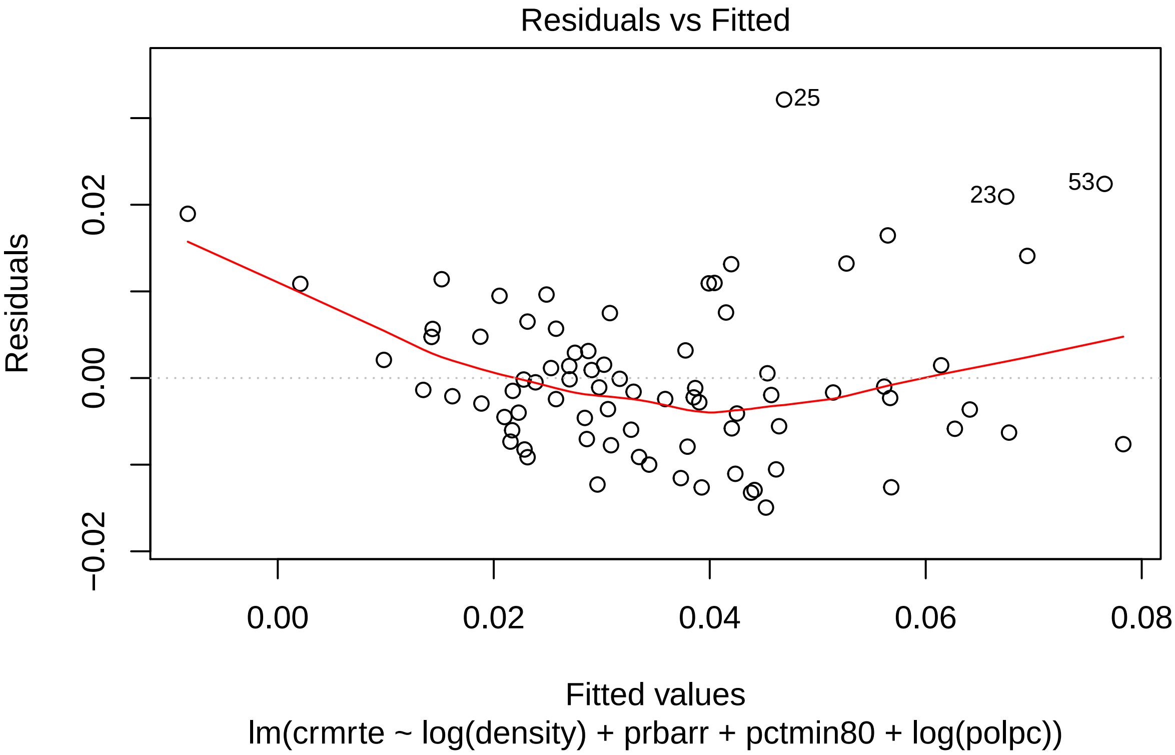
```r
shapiro.test(model2$residuals)
```

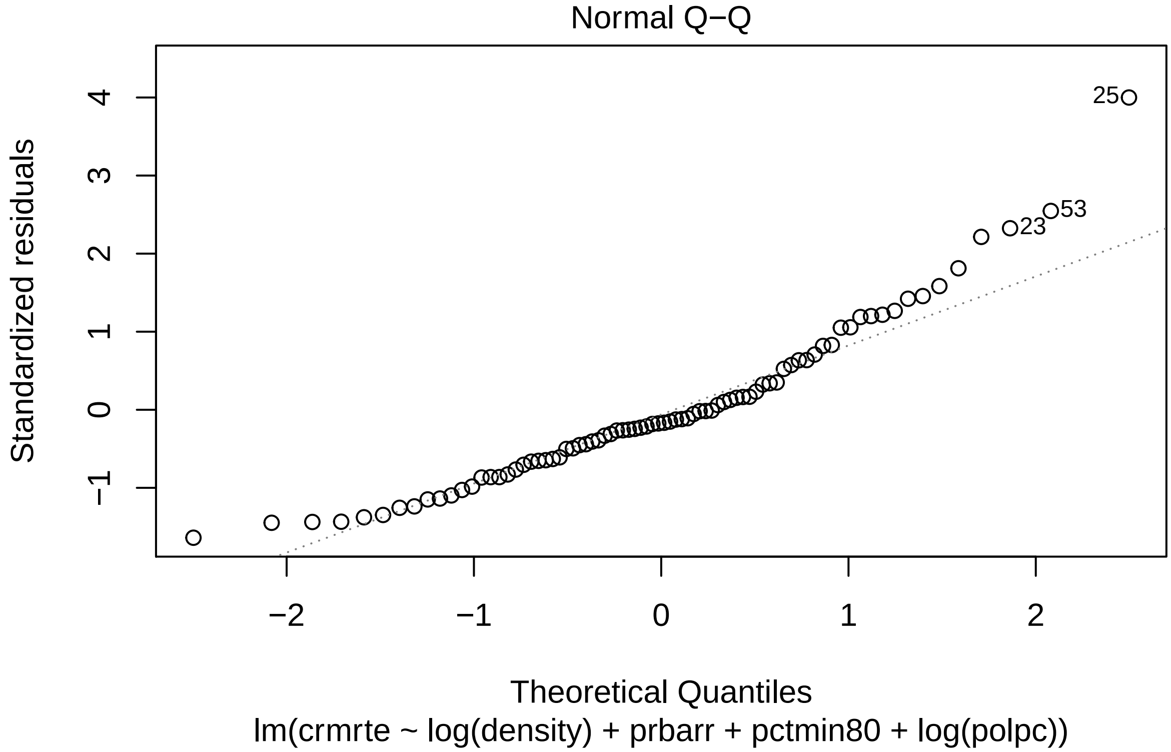
```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.9435, p-value = 0.00149
```

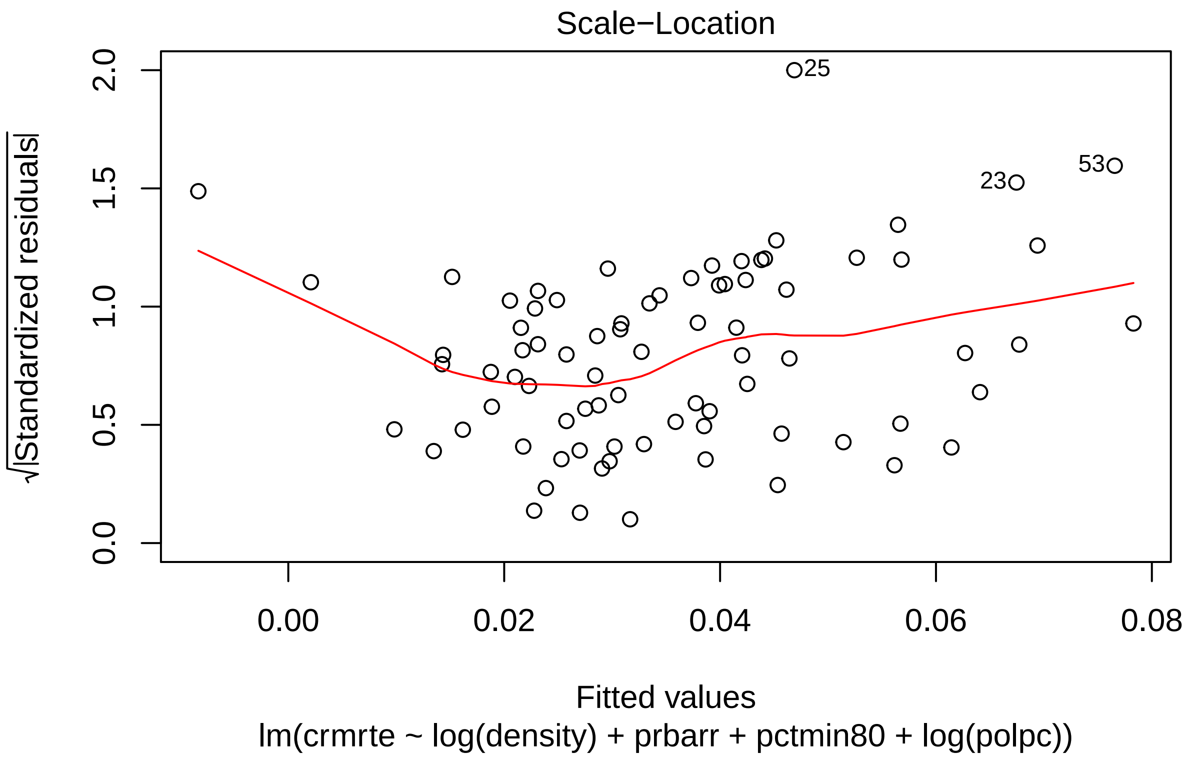
```r
bptest(model2)
```

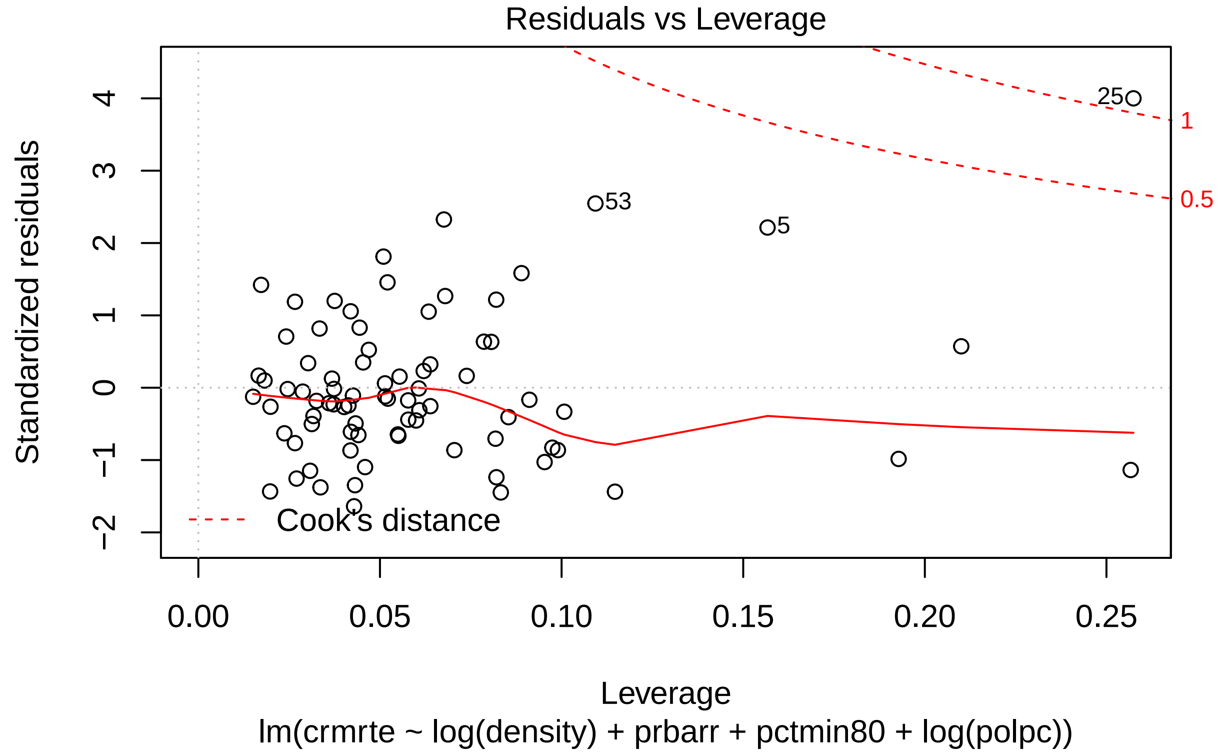
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 14.404, df = 4, p-value = 0.006111
```

```r
plot(model2)
```









As a result, all of the confirmed and unconfirmed assumptions from the previous model hold into this model as well.

Finally, a model is specified with all variables in the dataset included.

```
(model3 = lm(crmrte ~ county + density + prbarr + prbconv + prbpris + avgsen + polpc + taxpc + west + c
```

```
##
```

```
## Call:
## lm(formula = crmrte ~ county + density + prbarr + prbconv + prbpris +
##     avgsen + polpc + taxpc + west + central + urban + pctmin80 +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     mix + pctymle, data = data)
##
## Coefficients:
## (Intercept)       county      density       prbarr      prbconv
##   1.350e-02    9.834e-06    5.090e-03   -5.107e-02   -1.866e-02
##     prbpris       avgsen        polpc        taxpc         west
##   4.878e-03   -3.936e-04    6.818e+00    1.713e-04   -2.663e-03
##     central        urban     pctmin80         wcon         wtuc
##  -4.175e-03    1.068e-03    3.197e-04    2.225e-05    4.585e-06
##        wtrd         wfir         wser         wmfg         wfed
##   2.729e-05   -3.400e-05   -2.101e-06   -8.347e-06    3.117e-05
##        wsta         wloc          mix      pctymle
##  -2.553e-05    1.332e-05   -1.913e-02    1.017e-01
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = crmrte ~ county + density + prbarr + prbconv + prbpris +
##     avgsen + polpc + taxpc + west + central + urban + pctmin80 +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     mix + pctymle, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0169837 -0.0038785 -0.0005335  0.0045365  0.0220250
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.350e-02  1.981e-02   0.682 0.497879
## county       9.834e-06  1.616e-05   0.608 0.545052
## density      5.090e-03  1.411e-03   3.607 0.000596 ***
## prbarr      -5.107e-02  9.977e-03  -5.119 2.86e-06 ***
## prbconv     -1.866e-02  3.793e-03  -4.920 6.05e-06 ***
## prbpris      4.878e-03  1.221e-02   0.400 0.690716
## avgsen      -3.936e-04  4.261e-04  -0.924 0.359073
## polpc        6.818e+00  1.561e+00   4.367 4.55e-05 ***
## taxpc        1.713e-04  9.594e-05   1.786 0.078714 .
## west        -2.663e-03  4.230e-03  -0.630 0.531158
## central     -4.175e-03  2.883e-03  -1.448 0.152338
## urban        1.068e-03  6.461e-03   0.165 0.869236
## pctmin80     3.197e-04  1.002e-04   3.192 0.002168 **
## wcon         2.225e-05  2.823e-05   0.788 0.433294
## wtuc         4.585e-06  1.522e-05   0.301 0.764210
## wtrd         2.729e-05  4.671e-05   0.584 0.561049
## wfir        -3.400e-05  2.765e-05  -1.230 0.223194
## wser        -2.101e-06  5.716e-06  -0.368 0.714373
## wmfg        -8.347e-06  1.443e-05  -0.578 0.565039
## wfed         3.117e-05  2.583e-05   1.206 0.231967
## wsta        -2.553e-05  2.636e-05  -0.968 0.336367
## wloc         1.332e-05  4.920e-05   0.271 0.787429
## mix         -1.913e-02  1.479e-02  -1.294 0.200335
## pctymle      1.017e-01  4.553e-02   2.233 0.028943 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008357 on 66 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8042
## F-statistic:  16.9 on 23 and 66 DF,  p-value: < 2.2e-16
```

```
vif(model3)
```

```
##   county  density   prbarr  prbconv  prbpris   avgsen    polpc    taxpc
## 1.132419 5.859470 2.404323 2.300461 1.235592 1.859038 3.051109 2.016876
##     west  central    urban pctmin80     wcon     wtuc     wtrd     wfir
## 4.124729 2.518082 4.357536 3.689596 2.315631 1.766915 3.189553 2.841079
##     wser     wmfg     wfed     wsta     wloc      mix  pctymle
## 1.791002 2.067023 3.057280 1.660168 2.441284 1.864209 1.452435
```

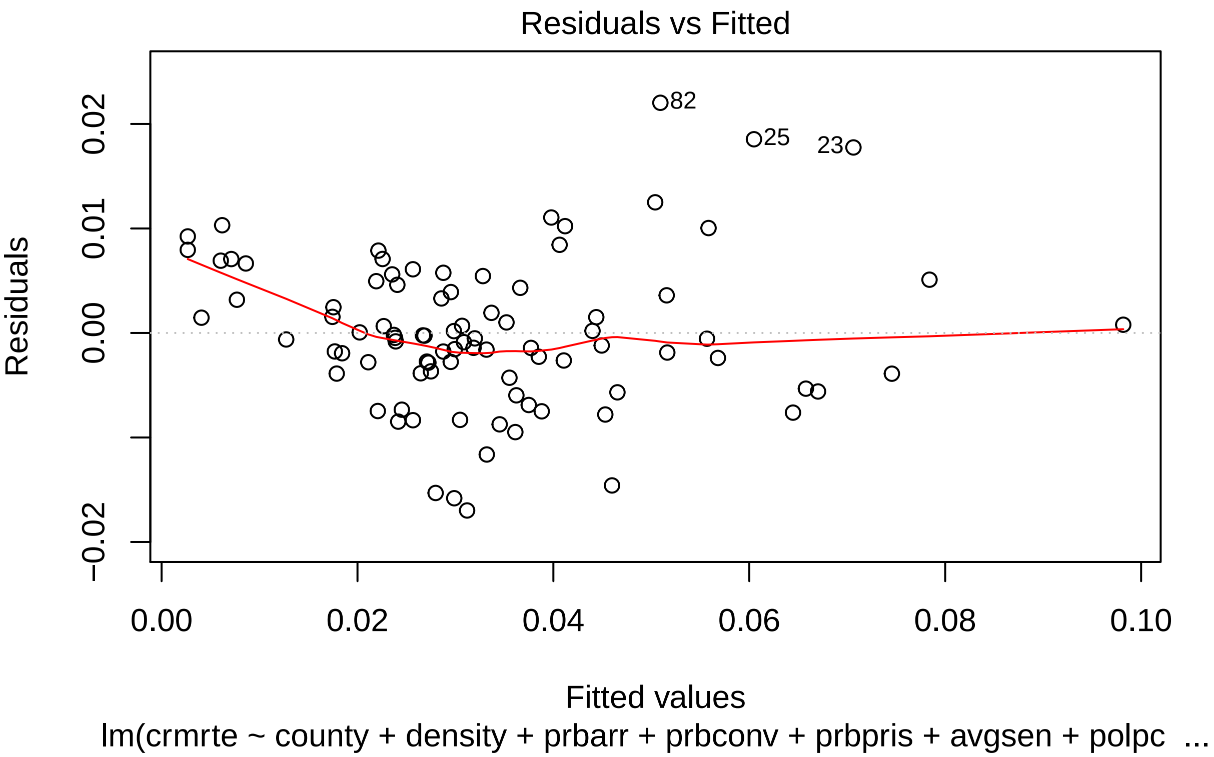
```
shapiro.test(model3$residuals)
```

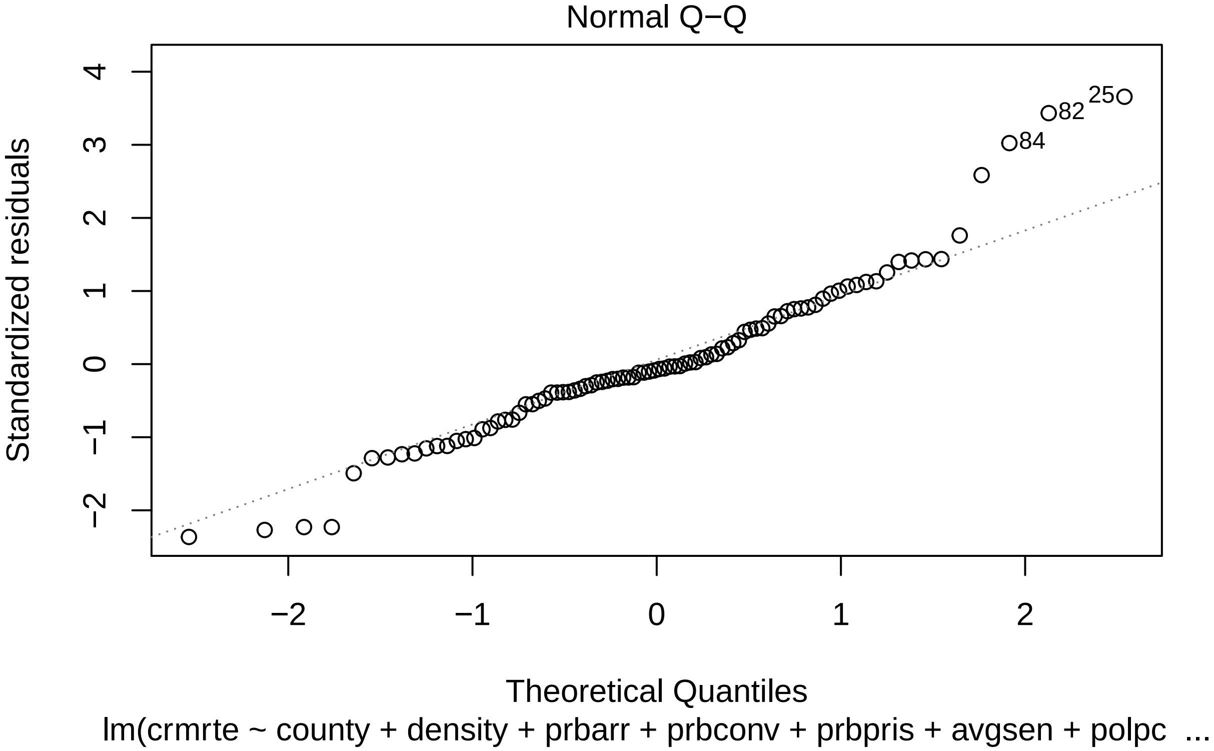
```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.97975, p-value = 0.1735
```

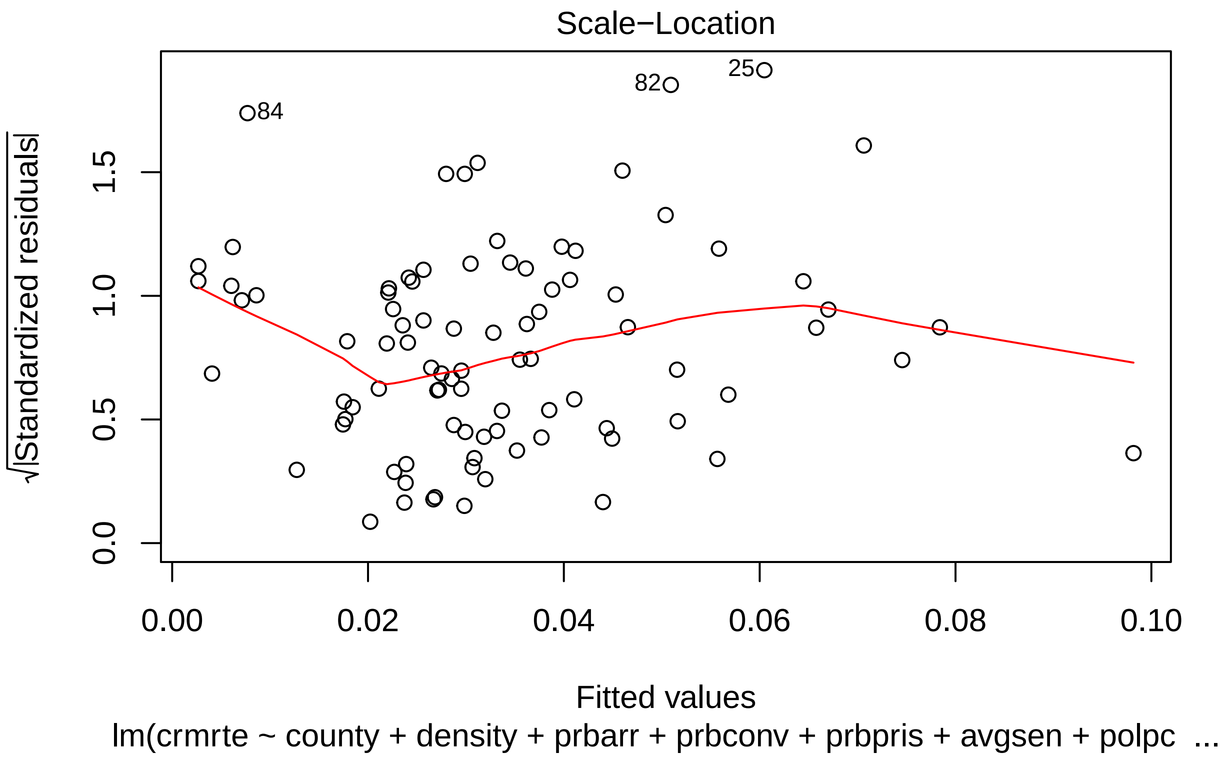
```
bptest(model3)
```

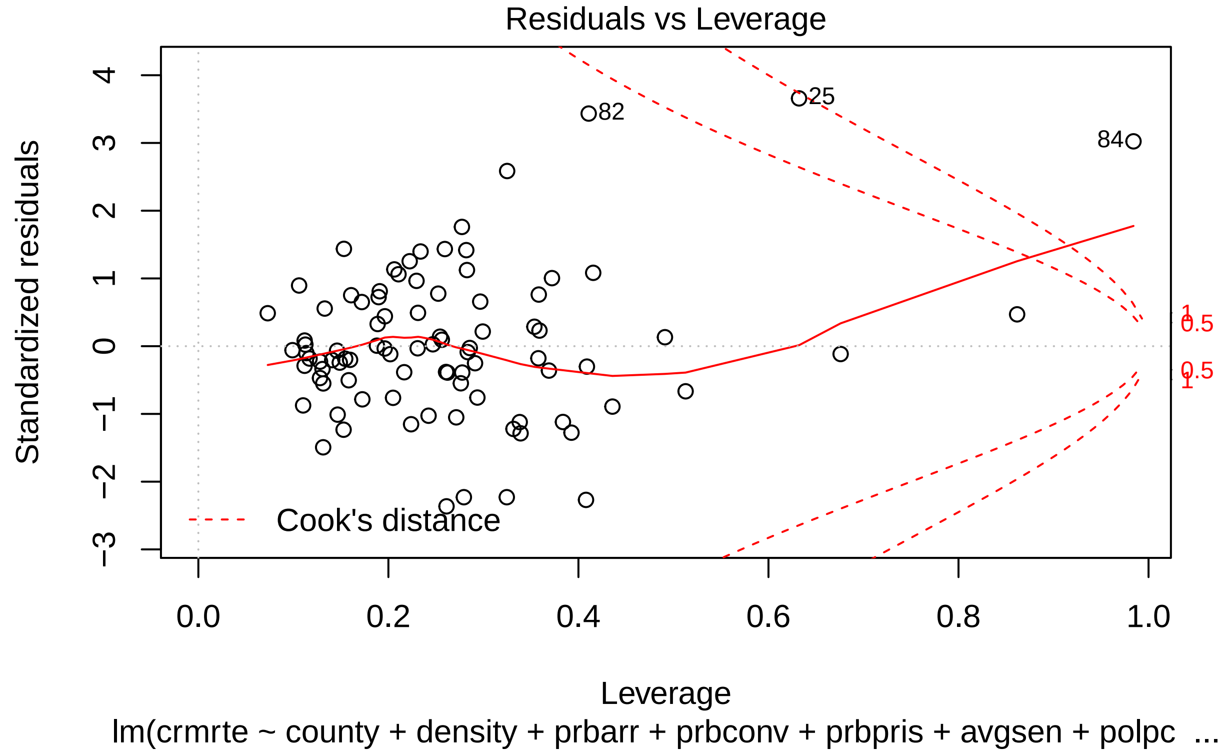
```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 35.706, df = 23, p-value = 0.0442
```

```
plot(model3)
```


Residuals vs Fitted
Residuals
0.02
0.01
0.00
−0.02
82
25
23
0.00
0.02
0.04
0.06
0.08
0.10
Fitted values
lm(crmrte ~ county + density + prbarr + prbconv + prbpris + avgsen + polpc ...


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
25
82
84
−2
−1
0
1
2
Theoretical Quantiles
lm(crmrte ~ county + density + prbarr + prbconv + prbpris + avgsen + polpc ...





However, in this model, many of the assumptions are violated. The residuals are not near 0, the error is not normally distributed, there are several points with high residuals and leverage, some variables have high degrees of collinearity, and heteroskedacity has been introduced. Therefore, this model is considered to be inaccurate and a more robust Model 2 is used with the compiled regression table as following:

```
se.model2 = sqrt(diag(vcovHC(model2)))
stargazer(model2, type = "text", omit.stat = "f", se = list(se.model2), star.cutoffs = c(0.05, 0.01, 0.0
```

```
##
## =================================================
##                          Dependent variable:
##                     -----------------------------
##                                crmrte
## -------------------------------------------------
## log(density)                  0.011***
##                                (0.003)
##
## prbarr                        -0.052***
##                                (0.013)
##
## pctmin80                      0.0004***
##                               (0.0001)
##
## log(polpc)                     0.025**
##                                (0.009)
##
## Constant                      0.202***
##                                (0.061)
##
## -------------------------------------------------
## Observations                     80
## R2                              0.769
## Adjusted R2                     0.757
## Residual Std. Error        0.009 (df = 75)
## =================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

Causality in the Model

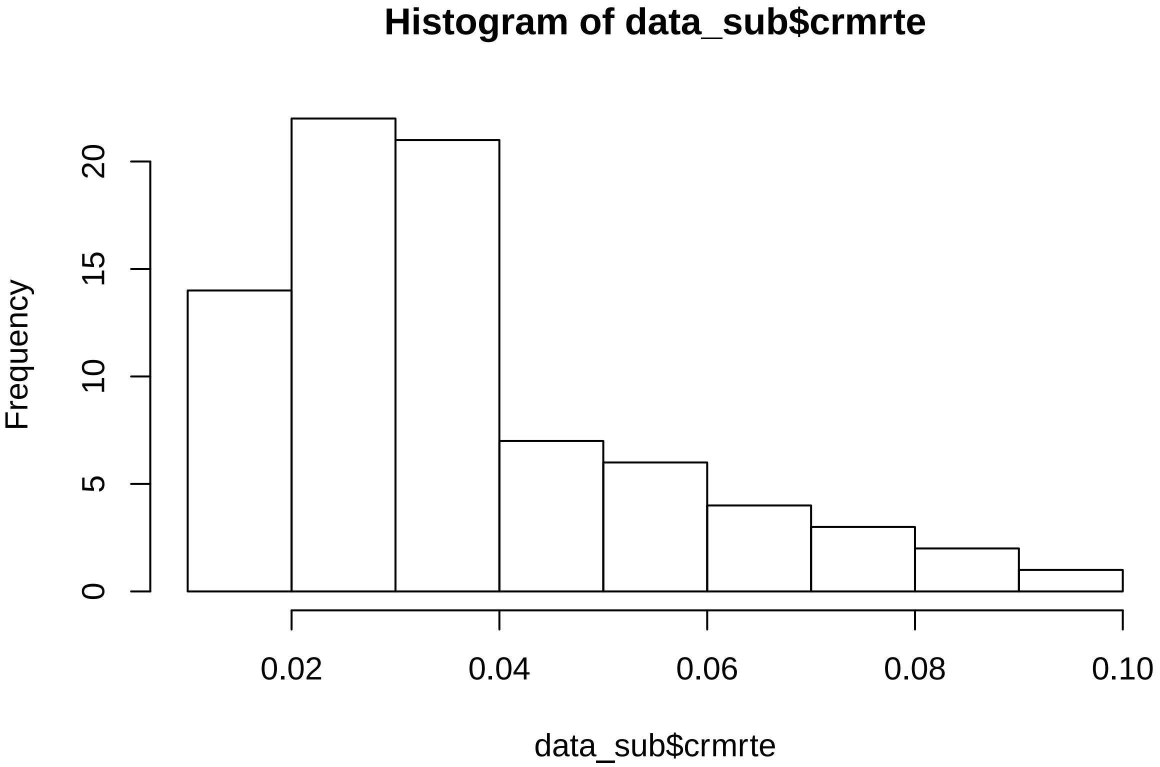
The three models we specify above range in their ability to be interpreted causally. The first model has only one variable. Though it meets the assumptions required to be considered unbiased, it almost certainly has omitted variable bias. Model 2 inclues more variables and still meets the assumptions. There is likely less ommitted variable bias and more ability to discuss causality in Model 2. However, Model 3 does not meet the assumptions required for unbiasedness or consistency. Therefore, while we can perhaps observe the direction of the coefficients, we are unable to draw any conclusions about causality from it.

Appendix

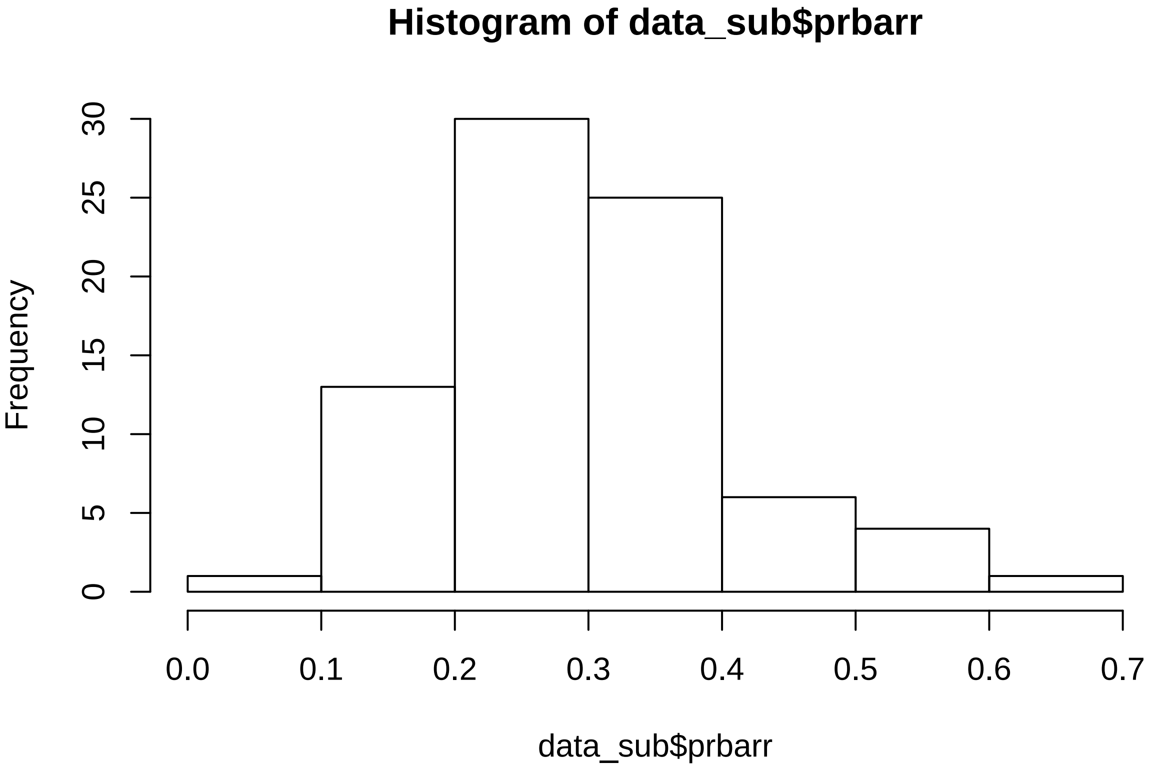
(code that has been removed - we could add back in depending on space)

Exploring the crime rates and their associated probabilities even further:

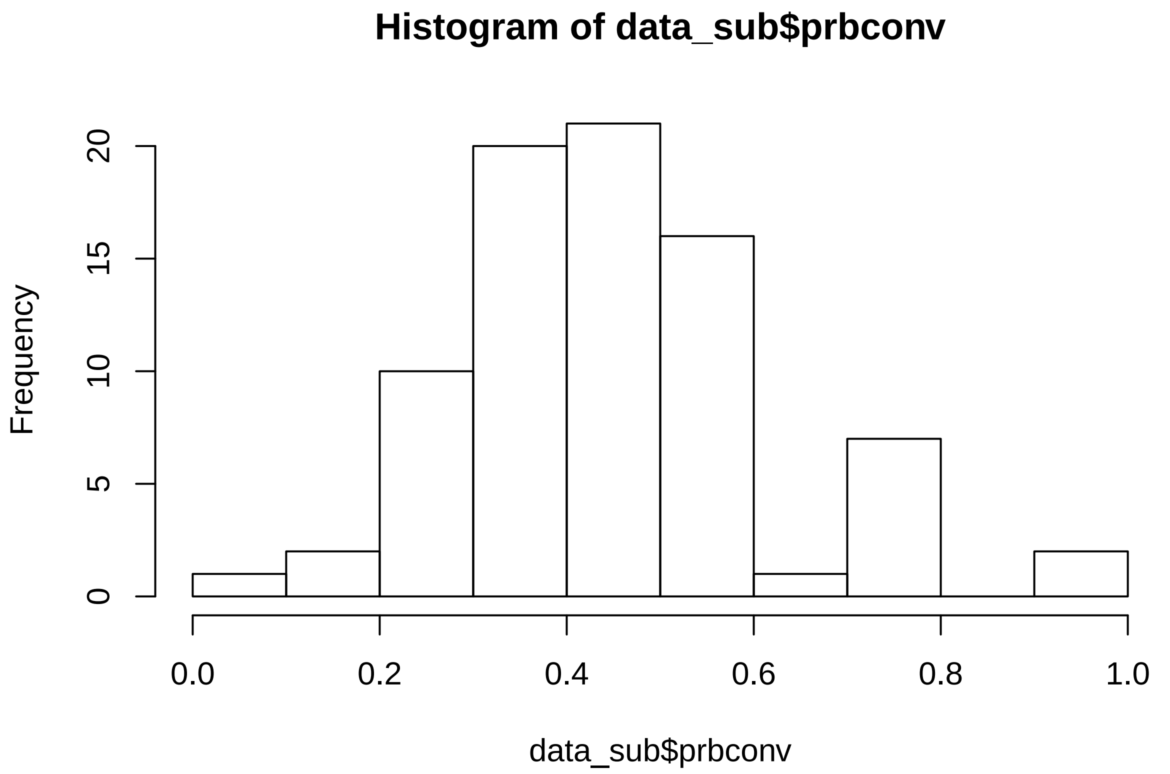
```
hist(data_sub$crmrte)
```



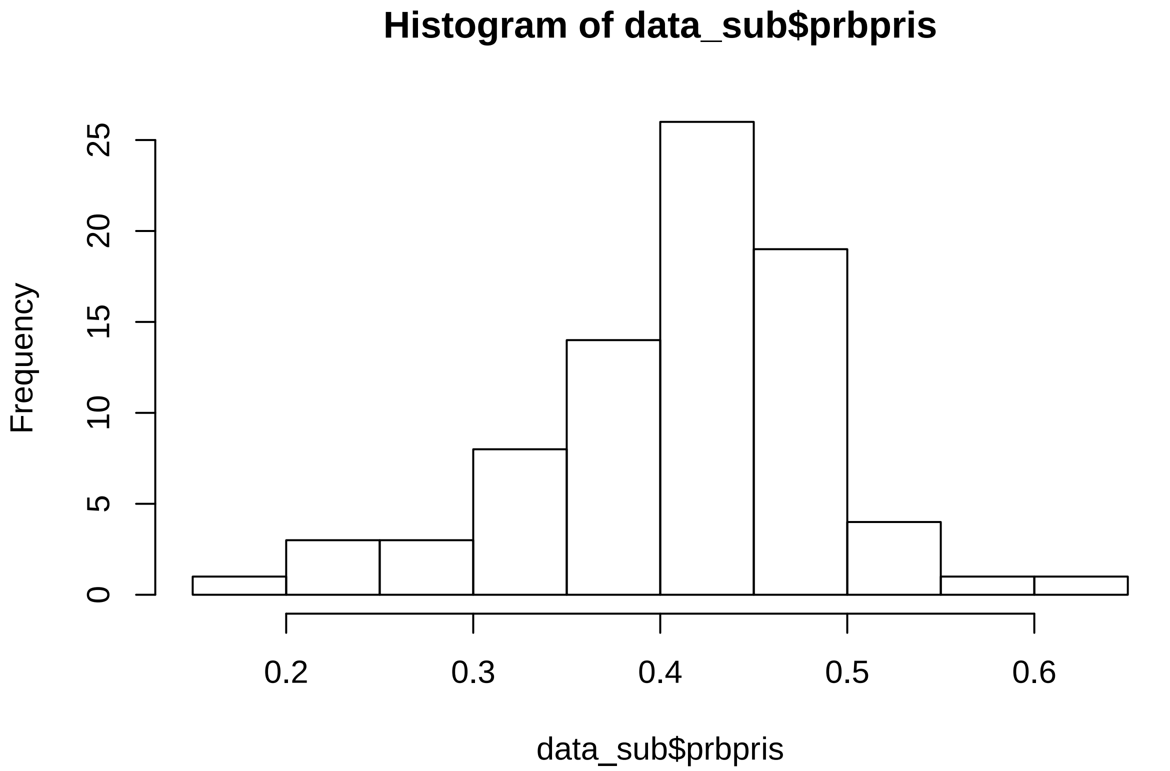
```
hist(data_sub$prbarr)
```



```
hist(data_sub$prbconv)
```



```
hist(data_sub$prbpris)
```



The crimes committed per person, the probability of arrest and the probability of conviction have a very positively skewed distribution. On the other hand, the probability of prison sentence has a relatively normal distribution.