Lab4_CaseyMicheline_MamrothAndrew_ArunimaKayath_Draf

Andrew Mamroth

August 13, 2017

Introduction

The purpose of this analysis is to review a set crime data in support of understanding determinants of crime and generating policy recommendations for local governments for this political campaign. This analysis is based on a set of county-level data from North Carolina from 1987. However, this data set is not inclusive of all North Carolina counties.

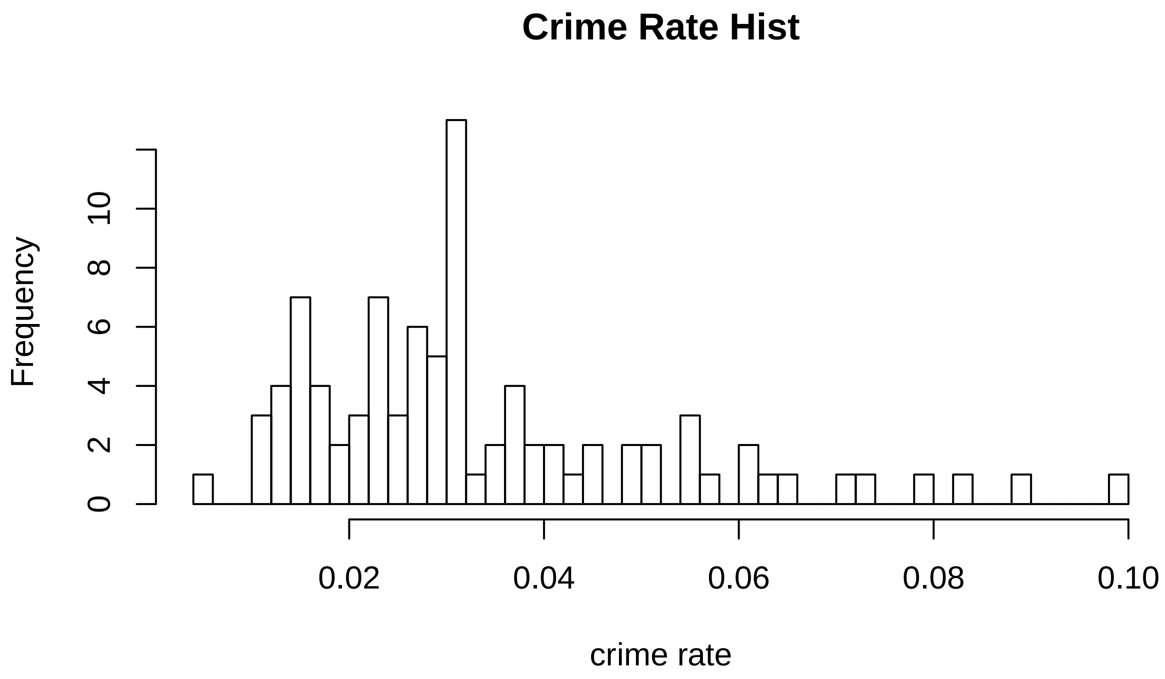
In our first stage, an exploratory data analysis was undertaken in order to better understand the data set, the variables, anomolies, and any variable transformations that may need to be done to provide a robust set of models. We also examined the relationships between variables from a variety of perspectives to support model building. Our second stage consisted of building and analyzing a set of linear models to determine the right mix of signficant explanatory variables, to identify best fit models, and to test all CLM assumptions. We use a backwards approach to model construction, meaning we first build a model that includes all the data we are given then remove the data with the least explanatory power.

Following that, we then explain why the information for the variables removed is already incorporated into the model and thus why it is excluded from the final model. We address bias, omitted variables, parsimony, and discussion of causality here. We conclude this paper with high-level set of recommendations and related notes.

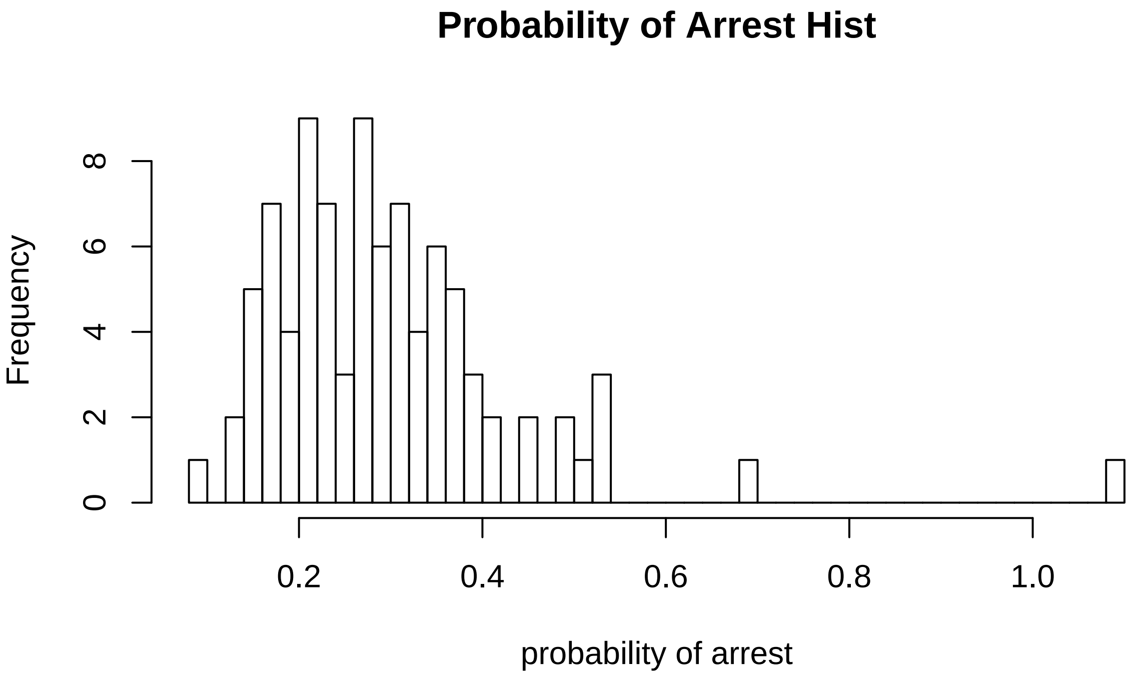
Exploratory Analysis

A thorough exploratory analysis of the crimes data set was performed to better understand the variables, anomolies, and any variable transformations that may need to be done to provide a robust set of models. Outlier data was reviewed for significance and possible patterns. We also examined the relationships between variables from a variety of perspectives to support model building. Histograms were run on all variables to understand their distributions. A select set for our key variables (crmrte, prbarr, prbconv, polpc, pctmin80) is provided here.

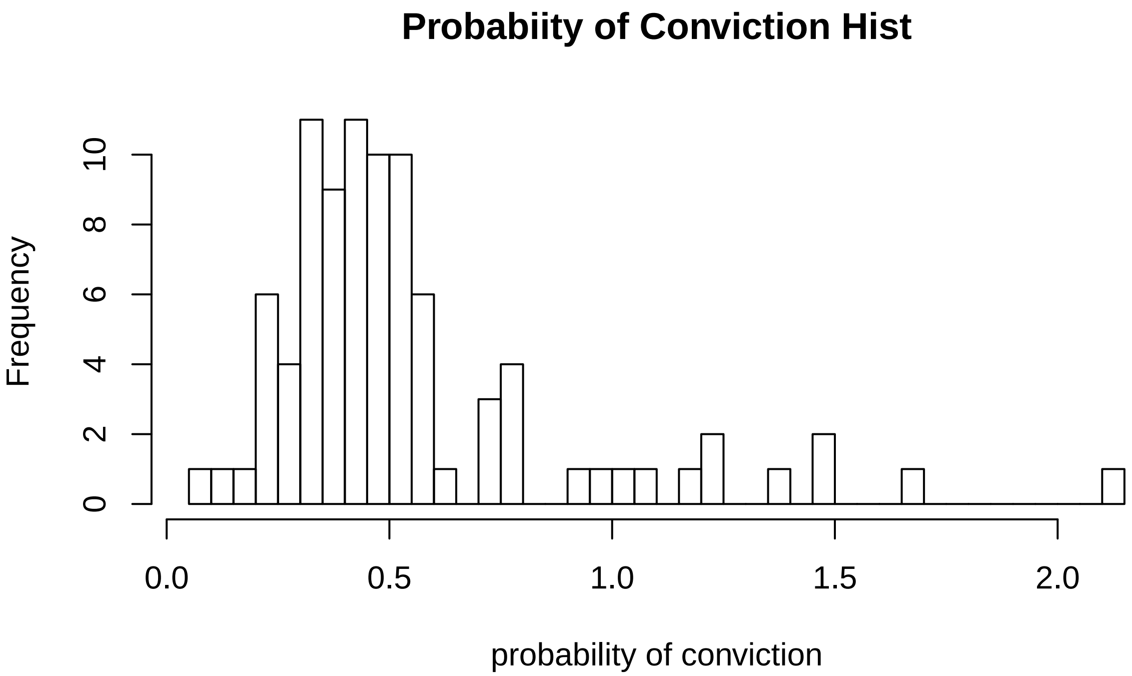
```r
hist(crime$crmrte, breaks = 50, xlab = "crime rate", ylab = "Frequency", main = "Crime Rate Hist")
```



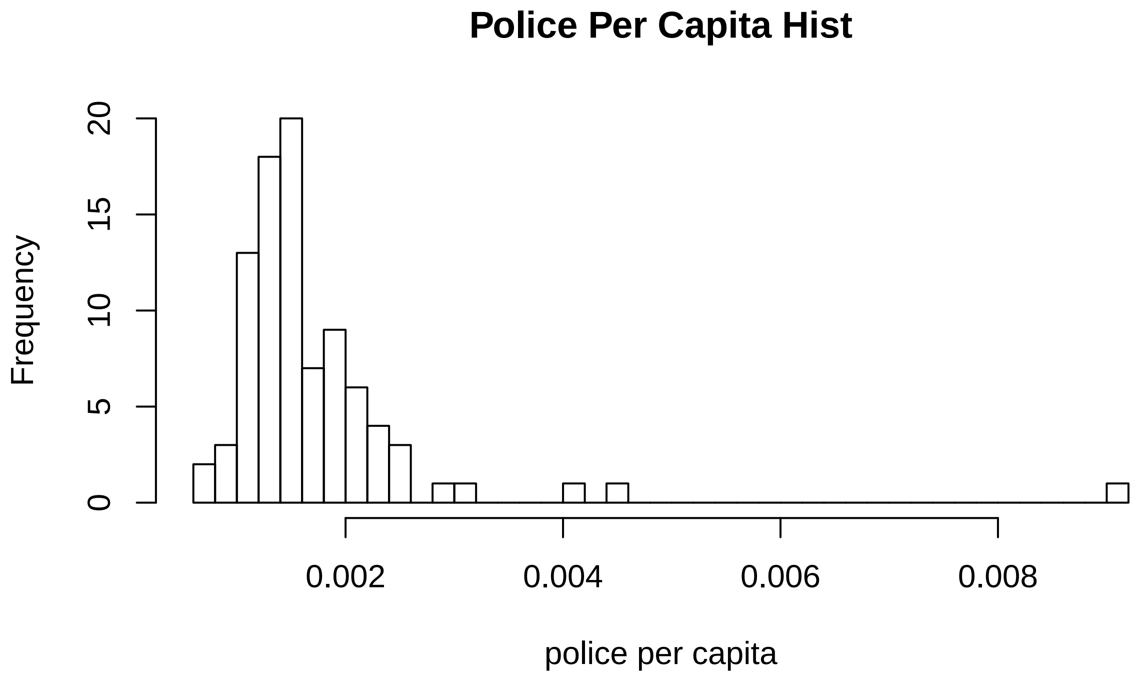
```r
hist(crime$prbarr, breaks = 50, xlab = "probability of arrest", ylab = "Frequency", main = "Probability
```



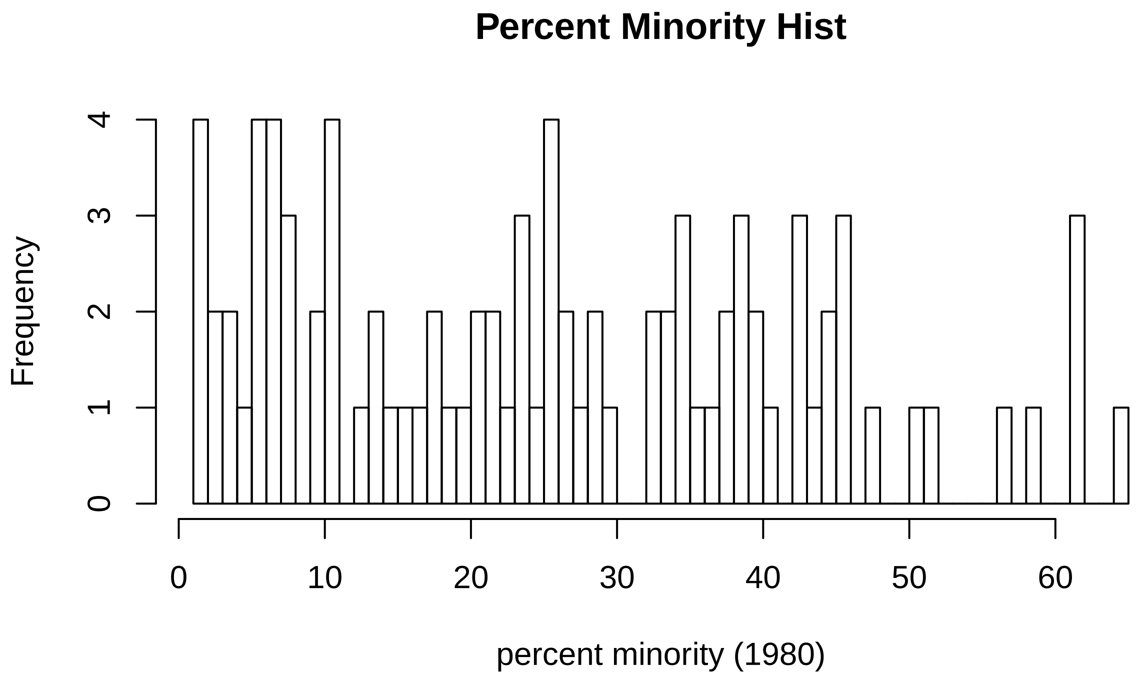
```
hist(crime$prbconv, breaks = 50, xlab = "probability of conviction", ylab = "Frequency", main = "Probabi
```



```
hist(crime$polpc, breaks = 50, xlab = "police per capita", ylab = "Frequency", main = "Police Per Capita
```



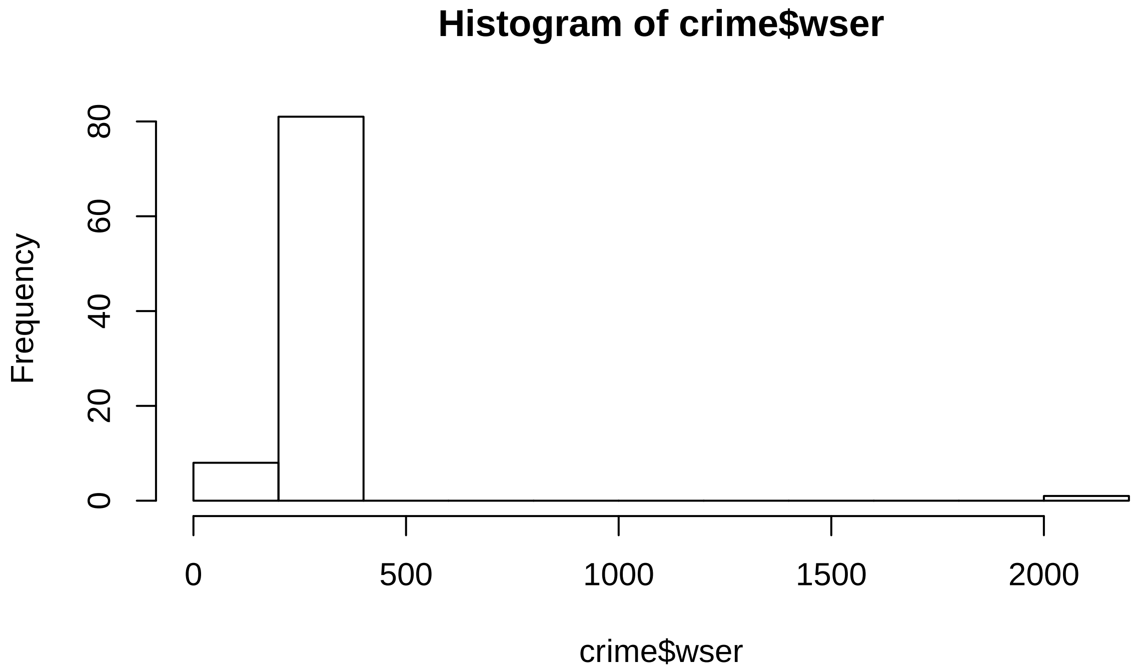
```
hist(crime$pctmin80, breaks = 50, xlab = "percent minority (1980)", ylab = "Frequency", main = "Percent
```



It should be noted that we do see “probabilities” greater than 1 here for the probability of conviction variable. A brief investigation into common crime statistics shows that the most common measures of convictions is actually conviction per thousands of people. (doesn’t seem to be consistancy between states to use 1000, 10,000, or 100,000.) We believe this is what this variable is measuring for each county.

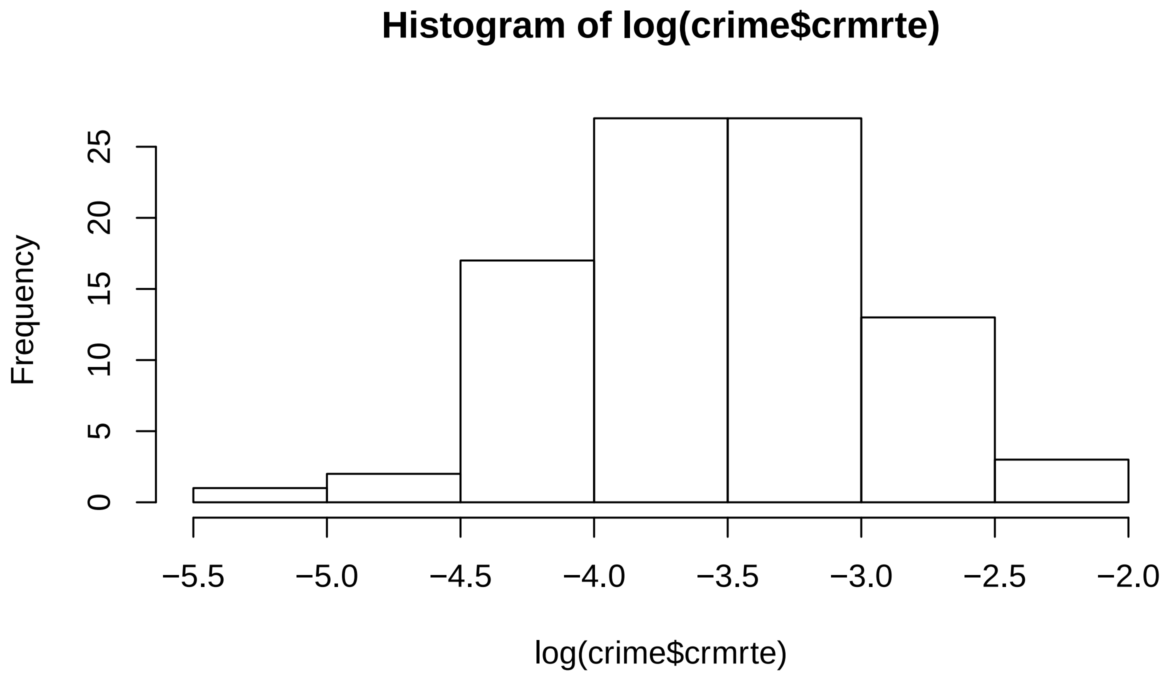
Additionally, it should be noted, there was a significant outlier in the wser variable. But this will not matter for reasons explained later.

```r
hist(crime$wser)
```

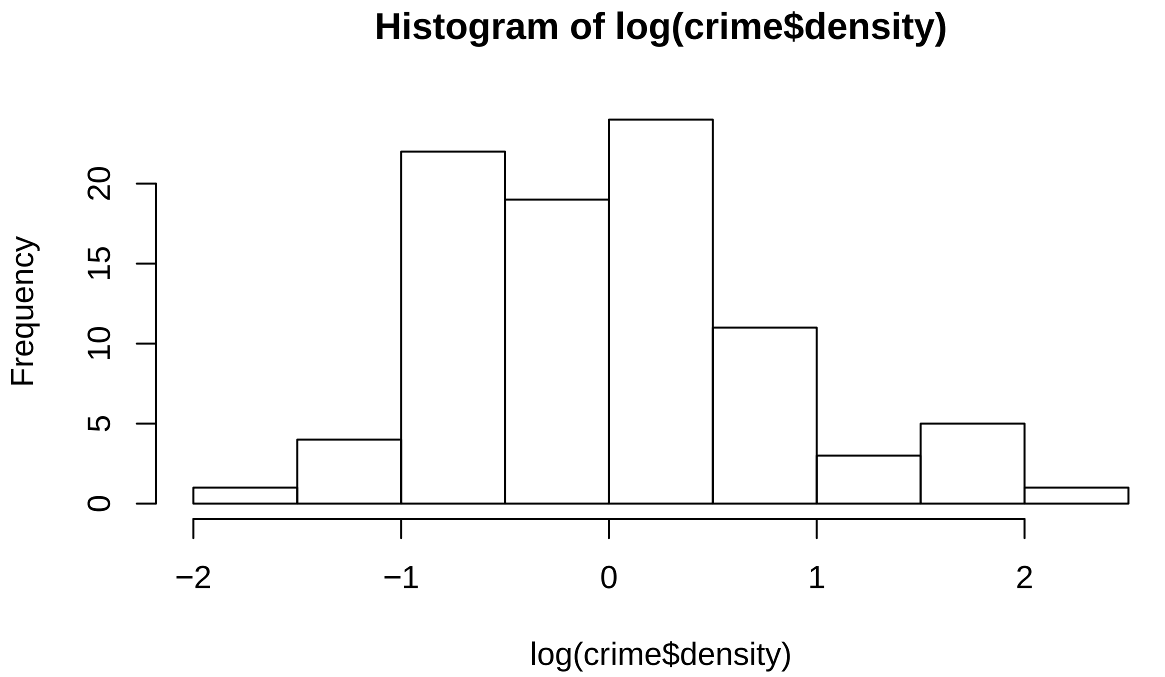


For some of the most skewed variables, polpc, density, and crmrte, we also look at the log transform.

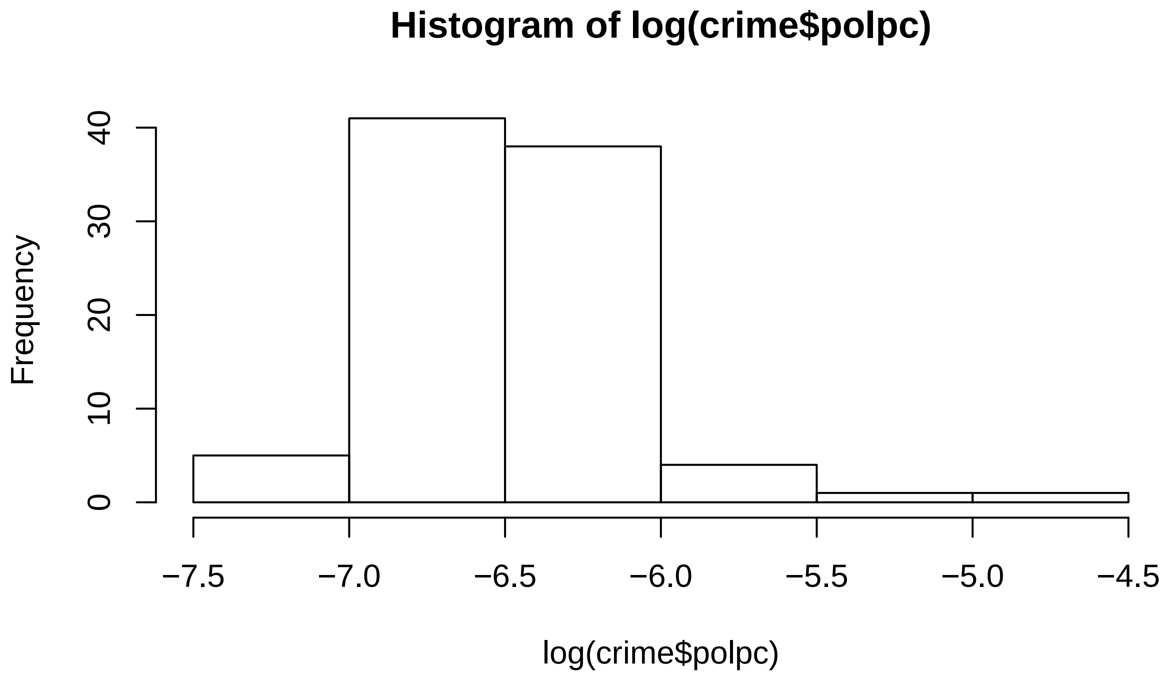
```r
hist(log(crime$crmrte))
```



```
hist(log(crime$density))
```



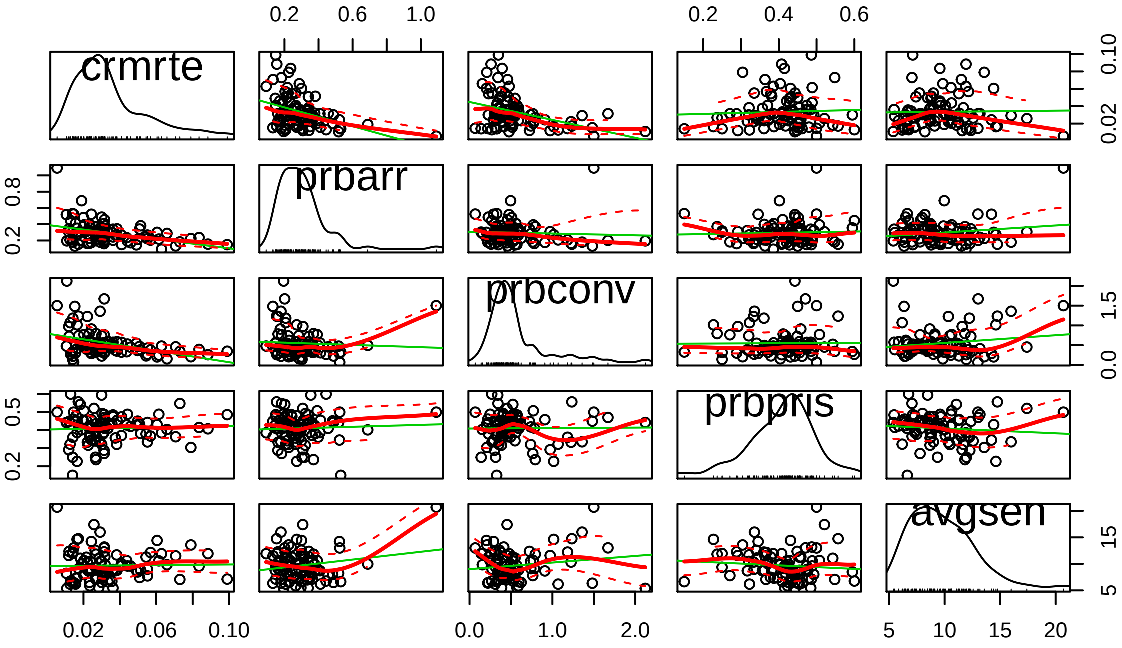
```
hist(log(crime$polpc))
```



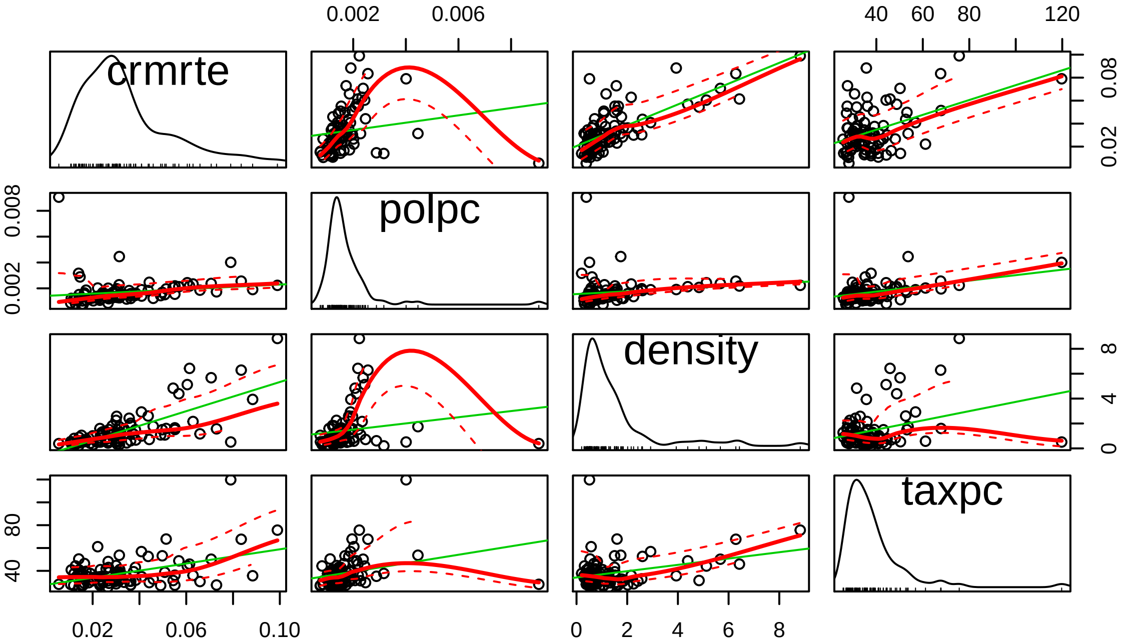
Here, we see that these variables take on a near normal distribution under a log transform.

Scatterplot matrices were developed for several variables to get an initial understanding of variable relationships.

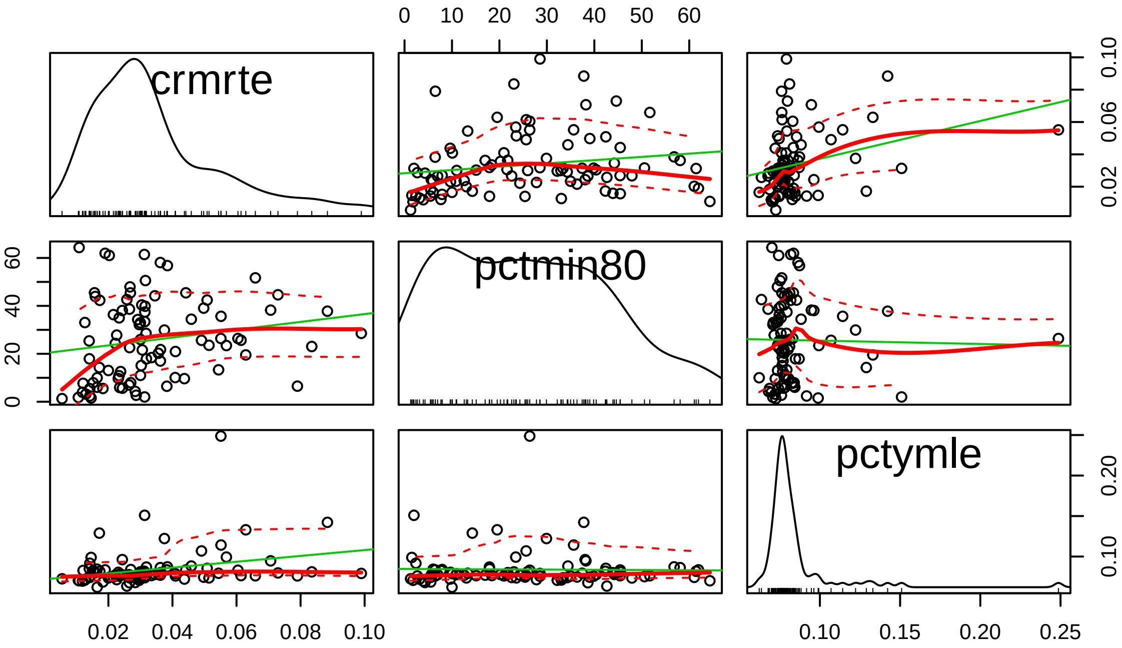
```
scatterplotMatrix(~crmrte + prbarr + prbconv + prbpris + avgsen, data=crime)
```



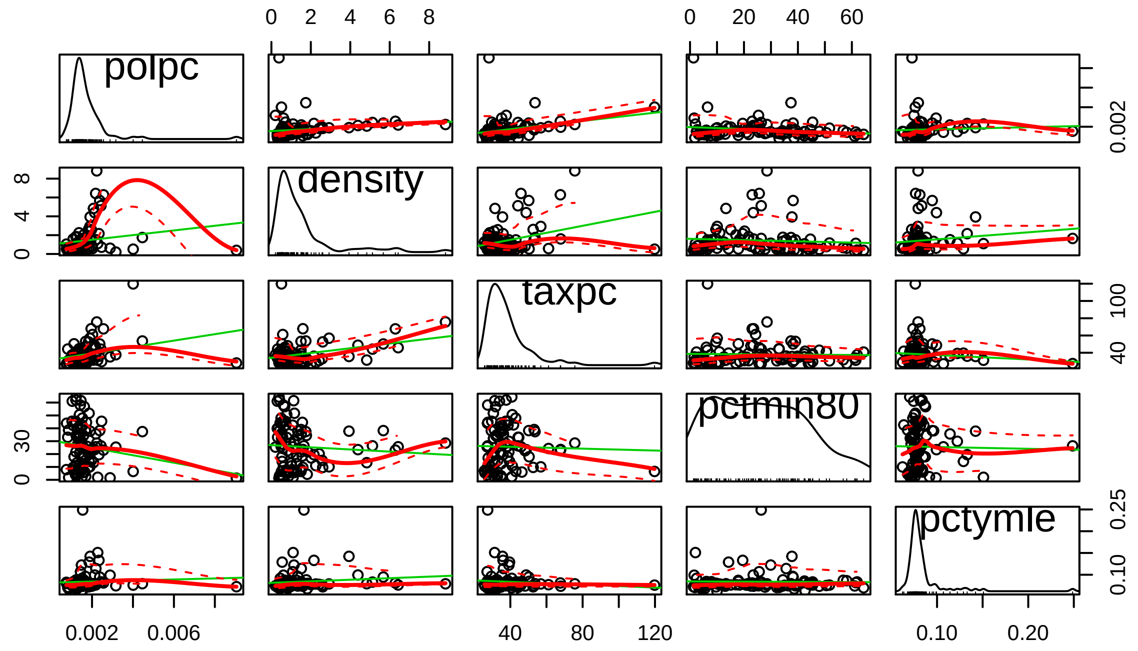
```
scatterplotMatrix(~crmrte + polpc + density + taxpc, data=crime)
```



```
scatterplotMatrix(~crmrte + pctmin80 + pctymle, data=crime)
```



```
scatterplotMatrix(~polpc + density + taxpc + pctmin80 + pctymle, data=crime)
```



Finally, we added to this by analyzing bivariate and multivariate relationships between a number of variables and combinations of variables. We looked at the bivariate correlations, R squared values, p values, for significance. Several of these are included below.

```
cor(crime[4:26])
```

```
              crmrte      prbarr     prbconv      prbpris      avgsen
crmrte    1.00000000 -0.39528302 -0.38596559  0.047995395  0.01979653
prbarr   -0.39528302  1.00000000 -0.05579621  0.045833245  0.17869425
prbconv  -0.38596559 -0.05579621  1.00000000  0.011022645  0.15585232
prbpris   0.04799540  0.04583324  0.01102265  1.000000000 -0.09468083
avgsen    0.01979653  0.17869425  0.15585232 -0.094680833  1.00000000
polpc     0.16728162  0.42596481  0.17186516  0.048207825  0.48815230
density   0.72777835 -0.30053317 -0.22791204  0.072609846  0.07159560
taxpc     0.44871512 -0.13719105 -0.12738963 -0.092360509  0.08654323
west     -0.38033874  0.18649897  0.07198455 -0.035044526  0.09855151
central   0.16588032 -0.16888612 -0.04640007  0.164520114 -0.15816897
urban     0.61506307 -0.20856276 -0.19709186  0.050354117  0.14391388
pctmin80  0.18165059  0.04907002  0.06249824  0.106136091 -0.16633664
wcon      0.39296155 -0.25183650 -0.11745577 -0.059611223 -0.03030263
wtuc      0.23599574 -0.07035781 -0.00716159  0.124730237  0.23116592
wtrd      0.42722262 -0.09948428 -0.13454762  0.139338689  0.10822274
wfir      0.33602609 -0.17253501  0.03217747  0.032777974  0.17792907
wser     -0.05206995 -0.13133303  0.45666832  0.038011073 -0.15103677
wmfg      0.35256117 -0.15316974  0.01757978  0.009408759  0.11045461
wfed      0.48991634 -0.20792619 -0.06085923  0.084965065  0.15240383
wsta      0.19984674 -0.16253921 -0.12843449 -0.031213974  0.12840868
wloc      0.35982934 -0.02447781  0.05060548  0.081193439  0.14575388
```

```
mix      -0.13200035  0.41289804 -0.30425124  0.116588825 -0.14170497
pctymle   0.29033966 -0.18096201 -0.16222602 -0.082759753  0.07099989
               polpc     density       taxpc          west     central
crmrte    0.16728162  0.72777835  0.44871512 -0.3803387441  0.16588032
prbarr    0.42596481 -0.30053317 -0.13719105  0.1864989678 -0.16888612
prbconv   0.17186516 -0.22791204 -0.12738963  0.0719845544 -0.04640007
prbpris   0.04820783  0.07260985 -0.09236051 -0.0350445258  0.16452011
avgsen    0.48815230  0.07159560  0.08654323  0.0985515064 -0.15816897
polpc     1.00000000  0.16152857  0.28055315  0.1441740336 -0.04600949
density   0.16152857  1.00000000  0.32047367 -0.1945888906  0.35682850
taxpc     0.28055315  0.32047367  1.00000000 -0.1738761174  0.03361974
west      0.14417403 -0.19458889 -0.17387612  1.0000000000 -0.42986348
central  -0.04600949  0.35682850  0.03361974 -0.4298634768  1.00000000
urban     0.15770869  0.82068254  0.34574617 -0.0800034085  0.15927023
pctmin80 -0.16911752 -0.07470698 -0.02797739 -0.6245144292 -0.05487554
wcon     -0.02379236  0.45134939  0.26395677 -0.1923563276  0.39786886
wtuc      0.17277373  0.33119447  0.17129001  0.0217295819  0.18855844
wtrd      0.12384123  0.59414742  0.18392144 -0.1913375272  0.38668510
wfir      0.19522607  0.54597415  0.13094363 -0.0528935846  0.29060049
wser     -0.01638582  0.04344734  0.07594777 -0.0633567781  0.19261249
wmfg      0.27043619  0.43766213  0.25860844 -0.0106669263  0.17368746
wfed      0.16187035  0.58693219  0.06207230 -0.2106600886  0.34923553
wsta      0.04891417  0.22310548 -0.03498830 -0.0785093734  0.08527707
wloc      0.38698768  0.46001747  0.21990116 -0.1429070937  0.33323127
mix       0.02411189 -0.13172771 -0.04355958  0.0008159465 -0.09210923
pctymle   0.05022177  0.11478144 -0.09154375 -0.0362124738 -0.10371067
               urban    pctmin80        wcon        wtuc        wtrd
crmrte    0.61506307  0.18165059  0.39296155  0.23599574  0.427222622
prbarr   -0.20856276  0.04907002 -0.25183650 -0.07035781 -0.099484278
prbconv  -0.19709186  0.06249824 -0.11745577 -0.00716159 -0.134547618
prbpris   0.05035412  0.10613609 -0.05961122  0.12473024  0.139338689
avgsen    0.14391388 -0.16633664 -0.03030263  0.23116592  0.108222741
polpc     0.15770869 -0.16911752 -0.02379236  0.17277373  0.123841229
density   0.82068254 -0.07470698  0.45134939  0.33119447  0.594147416
taxpc     0.34574617 -0.02797739  0.26395677  0.17129001  0.183921439
west     -0.08000341 -0.62451443 -0.19235633  0.02172958 -0.191337527
central   0.15927023 -0.05487554  0.39786886  0.18855844  0.386685101
urban     1.00000000  0.01619569  0.31926691  0.22632785  0.431728441
pctmin80  0.01619569  1.00000000 -0.10793251 -0.18913279 -0.064824402
wcon      0.31926691 -0.10793251  1.00000000  0.40937889  0.564058565
wtuc      0.22632785 -0.18913279  0.40937889  1.00000000  0.351683658
wtrd      0.43172844 -0.06482440  0.56405857  0.35168366  1.000000000
wfir      0.40171167 -0.07717356  0.48893774  0.32761956  0.668154525
wser      0.05589097  0.19672114 -0.01316438 -0.01924404 -0.020741268
wmfg      0.40362299 -0.11688213  0.34739282  0.46892658  0.371487416
wfed      0.42602595  0.03081152  0.50666394  0.39866915  0.640521866
wsta      0.30194045  0.09274887 -0.01885609 -0.15340397  0.007267295
wloc      0.33835635 -0.10590108  0.51704129  0.33301976  0.581463886
mix      -0.06417238  0.20123542 -0.19587213 -0.25346871 -0.125754703
pctymle   0.09396449 -0.01925657 -0.02214779 -0.10249879 -0.109277017
                wfir         wser         wmfg        wfed         wsta
crmrte    0.33602609 -0.052069955  0.352561171  0.48991634  0.199846744
prbarr   -0.17253501 -0.131333029 -0.153169735 -0.20792619 -0.162539207
prbconv   0.03217747  0.456668322  0.017579785 -0.06085923 -0.128434487
```

```
prbpris   0.03277797  0.038011073  0.009408759  0.08496507 -0.031213974
avgsen    0.17792907 -0.151036772  0.110454606  0.15240383  0.128408685
polpc     0.19522607 -0.016385818  0.270436194  0.16187035  0.048914168
density   0.54597415  0.043447343  0.437662125  0.58693219  0.223105478
taxpc     0.13094363  0.075947768  0.258608438  0.06207230 -0.034988299
west     -0.05289358 -0.063356778 -0.010666926 -0.21066009 -0.078509373
central   0.29060049  0.192612486  0.173687456  0.34923553  0.085277071
urban     0.40171167  0.055890972  0.403622994  0.42602595  0.301940450
pctmin80 -0.07717356  0.196721137 -0.116882126  0.03081152  0.092748871
wcon      0.48893774 -0.013164375  0.347392817  0.50666394 -0.018856088
wtuc      0.32761956 -0.019244042  0.468926579  0.39866915 -0.153403974
wtrd      0.66815452 -0.020741268  0.371487416  0.64052187  0.007267295
wfir      1.00000000  0.013716140  0.497583408  0.62317882  0.240700059
wser      0.01371614  1.000000000  0.008986754  0.02067471  0.037471156
wmfg      0.49758341  0.008986754  1.000000000  0.51823047  0.052336590
wfed      0.62317882  0.020674709  0.518230474  1.00000000  0.188250660
wsta      0.24070006  0.037471156  0.052336590  0.18825066  1.000000000
wloc      0.55443563  0.076971337  0.450453501  0.51941357  0.164641269
mix      -0.21232339 -0.173562869 -0.344125134 -0.31220529 -0.075726032
pctymle   0.01075555 -0.043107714  0.024179451 -0.06046726  0.218316221
                 wloc           mix      pctymle
crmrte    0.359829341 -0.1320003539  0.290339658
prbarr   -0.024477813  0.4128980444 -0.180962011
prbconv   0.050605485 -0.3042512443 -0.162226023
prbpris   0.081193439  0.1165888249 -0.082759753
avgsen    0.145753884 -0.1417049658  0.070999887
polpc     0.386987678  0.0241118925  0.050221768
density   0.460017473 -0.1317277105  0.114781444
taxpc     0.219901156 -0.0435595792 -0.091543750
west     -0.142907094  0.0008159465 -0.036212474
central   0.333231267 -0.0921092281 -0.103710667
urban     0.338356350 -0.0641723765  0.093964486
pctmin80 -0.105901082  0.2012354175 -0.019256570
wcon      0.517041291 -0.1958721285 -0.022147787
wtuc      0.333019765 -0.2534687080 -0.102498785
wtrd      0.581463886 -0.1257547028 -0.109277017
wfir      0.554435635 -0.2123233861  0.010755553
wser      0.076971337 -0.1735628695 -0.043107714
wmfg      0.450453501 -0.3441251344  0.024179451
wfed      0.519413570 -0.3122052928 -0.060467265
wsta      0.164641269 -0.0757260320  0.218316221
wloc      1.000000000 -0.2535193780 -0.001651489
mix      -0.253519378  1.0000000000 -0.092856609
pctymle  -0.001651489 -0.0928566094  1.000000000
```

```r
plot(crime$density, crime$crmrte)
modela <- lm(crmrte~density, data = crime)
modela
```

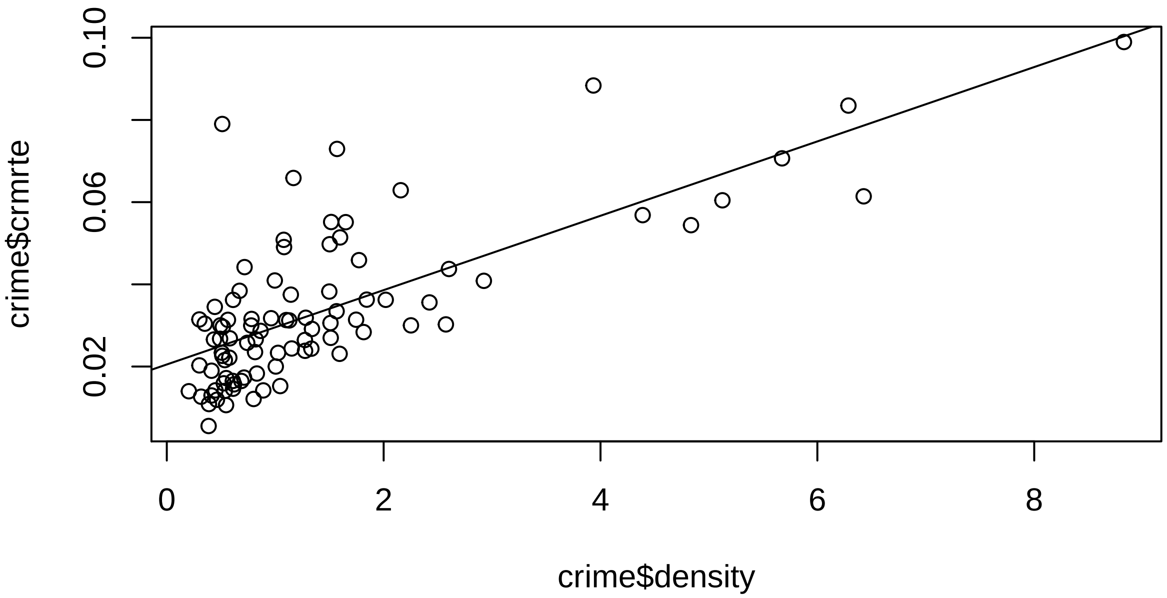
```

Call:
lm(formula = crmrte ~ density, data = crime)

Coefficients:
(Intercept)      density
```

```
   0.020503     0.009046
```

```
abline(modela)
```



```
summary(modela)
```

```

Call:
lm(formula = crmrte ~ density, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.018459 -0.009471 -0.002741  0.004902  0.053887 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0205027  0.0018954  10.817  < 2e-16 ***
density     0.0090458  0.0009087   9.955 4.45e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01303 on 88 degrees of freedom
Multiple R-squared:  0.5297,	Adjusted R-squared:  0.5243 
F-statistic:  99.1 on 1 and 88 DF,  p-value: 4.45e-16
```

```
cor(crime$crmrte, crime$density, use="pairwise.complete.obs")
```

```
[1] 0.7277783
```

```
modelb <- lm(crmrte~density+urban, data = crime)
modelb
```

```
Call:
lm(formula = crmrte ~ density + urban, data = crime)

Coefficients:
(Intercept)      density        urban  
   0.020982     0.008490     0.003596  
```

```r
summary(modelb)
```

```
Call:
lm(formula = crmrte ~ density + urban, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.018725 -0.009311 -0.003069  0.004768  0.053691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.020982   0.002215   9.473 4.84e-15 ***
density     0.008490   0.001598   5.314 8.18e-07 ***
urban       0.003596   0.008484   0.424    0.673    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01309 on 87 degrees of freedom
Multiple R-squared:  0.5306,	Adjusted R-squared:  0.5198 
F-statistic: 49.18 on 2 and 87 DF,  p-value: 5.139e-15
```

```r
modelc <- lm(crmrte~density+prbarr+prbconv+prbpris+avgsen, data = crime)
modelc
```

```
Call:
lm(formula = crmrte ~ density + prbarr + prbconv + prbpris + 
    avgsen, data = crime)

Coefficients:
(Intercept)      density       prbarr      prbconv      prbpris  
  0.0343712    0.0072427   -0.0341768   -0.0148251    0.0062388  
     avgsen  
  0.0004561  
```

```r
summary(modelc)
```

```
Call:
lm(formula = crmrte ~ density + prbarr + prbconv + prbpris + 
    avgsen, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.019794 -0.007971 -0.003249  0.005560  0.043610 

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0343712  0.0084552   4.065 0.000108 ***
density      0.0072427  0.0009144   7.921 8.73e-12 ***
prbarr      -0.0341768  0.0099446  -3.437 0.000918 ***
prbconv     -0.0148251  0.0037484  -3.955 0.000159 ***
prbpris      0.0062388  0.0157094   0.397 0.692273
avgsen       0.0004561  0.0004665   0.978 0.330978
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01179 on 84 degrees of freedom
Multiple R-squared:  0.6325,    Adjusted R-squared:  0.6106
F-statistic: 28.91 on 5 and 84 DF,  p-value: < 2.2e-16
```

```
modeld <- lm(crmrte~density+wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+wloc, data = crime)
modeld
```

```

Call:
lm(formula = crmrte ~ density + wcon + wtuc + wtrd + wfir + wser +
    wmfg + wfed + wsta + wloc, data = crime)

Coefficients:
(Intercept)      density         wcon         wtuc         wtrd
  2.701e-03    8.568e-03    4.304e-05   -1.112e-05   -5.444e-06
       wfir         wser         wmfg         wfed         wsta
 -7.462e-05   -7.864e-06    1.321e-05    4.167e-05    2.770e-05
       wloc
  1.697e-05
```

```
summary(modeld)
```

```

Call:
lm(formula = crmrte ~ density + wcon + wtuc + wtrd + wfir + wser +
    wmfg + wfed + wsta + wloc, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.017548 -0.008526 -0.002629  0.004543  0.053684

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.701e-03  2.195e-02   0.123   0.9024
density      8.568e-03  1.275e-03   6.720 2.55e-09 ***
wcon         4.304e-05  3.870e-05   1.112   0.2694
wtuc        -1.112e-05  2.211e-05  -0.503   0.6163
wtrd        -5.444e-06  6.929e-05  -0.079   0.9376
wfir        -7.462e-05  4.029e-05  -1.852   0.0678 .
wser        -7.864e-06  6.777e-06  -1.160   0.2494
wmfg         1.321e-05  2.087e-05   0.633   0.5285
wfed         4.167e-05  3.560e-05   1.170   0.2454
wsta         2.770e-05  3.700e-05   0.749   0.4564
wloc         1.697e-05  6.818e-05   0.249   0.8041
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.01314 on 79 degrees of freedom
Multiple R-squared:  0.5707,    Adjusted R-squared:  0.5163
F-statistic:  10.5 on 10 and 79 DF,  p-value: 4.627e-11
```

```
waldtest(modela,modeld,vcov = vcovHC)
```

```
Wald test

Model 1: crmrte ~ density
Model 2: crmrte ~ density + wcon + wtuc + wtrd + wfir + wser + wmfg +
    wfed + wsta + wloc
  Res.Df Df      F Pr(>F)
1     88
2     79  9 1.3379 0.2312
```

Key observations from correlations and plots :

- From the correlations, density is the most strongly correlated with crime rate.
- Urban also looks very correlated with crime rate, but doesn't show up as significant once density is included in the model (model b vs model a)
- probability of arrest and conviction show strong negative correlation with crime rate and show up as significant even after including density. (model c vs model a above)
- several of the taxpc and wage variables show some positive correlation with crimerate. This is counterintuitive, as higher tax and wages should correlate with higher income and hence lower crime, not higher crime. The wage variables don't show up as significant once density is included in the model, either singly, or jointly (as seen in the F-test for a model including all the wage variables (modeld) and density vs a pure density model(modela))

Building a Model

To build the model, we use a backwards approach. We build 3 seperate models and with each iteration we remove variables. We first build a model that includes all the data we are given then remove the data with the least explanatory power. Throughout, we explain why the information for the variables removed is already incorporated into the model and thus why it is excluded from the final model.

```r
model_1<-lm(crime$crmrte~crime$prbarr+crime$prbconv+crime$prbpris
                    +crime$avgsen+crime$polpc+crime$density+crime$taxpc
                    +crime$west+crime$central+crime$urban+crime$pctmin80
                    +crime$wcon+crime$wtuc+crime$wtrd+crime$wfir+crime$wser
                    +crime$wmfg+crime$wfed+crime$wsta+crime$wloc+crime$mix
                    +crime$pctymle)

summary(model_1)
```

```
Call:
lm(formula = crime$crmrte ~ crime$prbarr + crime$prbconv + crime$prbpris + 
    crime$avgsen + crime$polpc + crime$density + crime$taxpc + 
    crime$west + crime$central + crime$urban + crime$pctmin80 + 
    crime$wcon + crime$wtuc + crime$wtrd + crime$wfir + crime$wser + 
    crime$wmfg + crime$wfed + crime$wsta + crime$wloc + crime$mix + 
    crime$pctymle)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0168836 -0.0039309 -0.0004161  0.0046227  0.0228050 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.333e-02  1.972e-02   0.676 0.501164    
crime$prbarr   -5.135e-02  9.919e-03  -5.177 2.24e-06 ***
crime$prbconv  -1.854e-02  3.770e-03  -4.917 5.97e-06 ***
crime$prbpris   4.159e-03  1.209e-02   0.344 0.731917    
crime$avgsen   -3.958e-04  4.241e-04  -0.933 0.354003    
crime$polpc     6.918e+00  1.546e+00   4.476 3.03e-05 ***
crime$density   5.156e-03  1.400e-03   3.682 0.000464 ***
crime$taxpc     1.676e-04  9.530e-05   1.759 0.083168 .  
crime$west     -2.416e-03  4.190e-03  -0.577 0.566193    
crime$central  -4.163e-03  2.869e-03  -1.451 0.151468    
crime$urban     5.814e-04  6.382e-03   0.091 0.927681    
crime$pctmin80  3.277e-04  9.886e-05   3.315 0.001484 ** 
crime$wcon      2.406e-05  2.794e-05   0.861 0.392189    
crime$wtuc      5.257e-06  1.511e-05   0.348 0.729007    
crime$wtrd      2.896e-05  4.641e-05   0.624 0.534745    
crime$wfir     -3.482e-05  2.749e-05  -1.267 0.209657    
crime$wser     -1.887e-06  5.678e-06  -0.332 0.740741    
crime$wmfg     -8.792e-06  1.435e-05  -0.613 0.542111    
crime$wfed      2.981e-05  2.562e-05   1.164 0.248655    
crime$wsta     -2.326e-05  2.597e-05  -0.895 0.373764    
crime$wloc      1.337e-05  4.897e-05   0.273 0.785627    
crime$mix      -1.936e-02  1.472e-02  -1.315 0.192895    
crime$pctymle   1.035e-01  4.522e-02   2.288 0.025298 *  
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008317 on 67 degrees of freedom
Multiple R-squared:  0.854, Adjusted R-squared:  0.8061
F-statistic: 17.81 on 22 and 67 DF,  p-value: < 2.2e-16
```

From here we trim the variables with the least explanatory power to produce model 2 or the trimmed model. We note that while the r squared drops slightly, the adjusted r squared actually goes up.

```
model_2<-lm(crime$crmrte~crime$density+crime$prbarr+crime$prbconv+
             crime$polpc+crime$pctmin80+crime$taxpc+crime$pctymle)
summary(model_2)
```

```

Call:
lm(formula = crime$crmrte ~ crime$density + crime$prbarr + crime$prbconv + 
    crime$polpc + crime$pctmin80 + crime$taxpc + crime$pctymle)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0181805 -0.0052377 -0.0000983  0.0047007  0.0240021 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.761e-02  6.561e-03   2.683  0.00882 ** 
crime$density   5.462e-03  6.955e-04   7.854 1.36e-11 ***
crime$prbarr   -5.579e-02  8.955e-03  -6.230 1.90e-08 ***
crime$prbconv  -1.882e-02  2.919e-03  -6.448 7.37e-09 ***
crime$polpc     6.524e+00  1.269e+00   5.140 1.83e-06 ***
crime$pctmin80  3.560e-04  5.379e-05   6.618 3.50e-09 ***
crime$taxpc     1.874e-04  7.876e-05   2.380  0.01964 *  
crime$pctymle   8.855e-02  4.089e-02   2.165  0.03326 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00829 on 82 degrees of freedom
Multiple R-squared:  0.8225,    Adjusted R-squared:  0.8074 
F-statistic: 54.28 on 7 and 82 DF,  p-value: < 2.2e-16
```

Some of these variables can we discarded simply on the grounds that they have very little correlation with the dependant variable.

```
dat_1<-data.frame(crime$crmrte,crime$avgsen,crime$prbpris)
cor(dat_1)
```

```
              crime.crmrte crime.avgsen crime.prbpris
crime.crmrte    1.00000000   0.01979653    0.04799540
crime.avgsen    0.01979653   1.00000000   -0.09468083
crime.prbpris   0.04799540  -0.09468083    1.00000000
```

For wages it seems even alone they have little predictive power. It may be that case that what we really want to measure is not wages but unemployment as it may be that case that even if one doesn't have much money, they are at least employed and therefore will commit less crimes.

```
summary(lm(crime$crmrte~crime$wcon+crime$wtuc+crime$wtrd+crime$wfir
           +crime$wser+crime$wmfg+crime$wfed+crime$wsta+crime$wloc))
```

```
Call:
lm(formula = crime$crmrte ~ crime$wcon + crime$wtuc + crime$wtrd +
    crime$wfir + crime$wser + crime$wmfg + crime$wfed + crime$wsta +
    crime$wloc)

Residuals:
      Min        1Q    Median        3Q       Max
-0.035348 -0.009720 -0.003703  0.006302  0.052214

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.898e-02  2.390e-02  -2.887  0.00501 **
crime$wcon   6.737e-05  4.800e-05   1.404  0.16431
crime$wtuc  -8.665e-07  2.747e-05  -0.032  0.97492
crime$wtrd   1.245e-04  8.289e-05   1.501  0.13718
crime$wfir  -6.460e-05  5.016e-05  -1.288  0.20150
crime$wser  -5.261e-06  8.428e-06  -0.624  0.53424
crime$wmfg   3.333e-05  2.573e-05   1.295  0.19889
crime$wfed   7.975e-05  4.379e-05   1.821  0.07230 .
crime$wsta   8.239e-05  4.497e-05   1.832  0.07062 .
crime$wloc   1.162e-05  8.493e-05   0.137  0.89156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01636 on 80 degrees of freedom
Multiple R-squared:  0.3253,    Adjusted R-squared:  0.2494
F-statistic: 4.285 on 9 and 80 DF,  p-value: 0.0001451
```

To address why none of the flag variables, urban, west, and central, were not included in the trimmed model, we see that all three variables are highly correlated with values that are also in the model, but have higher correlations with the dependant variable. Or in the case of west, it is does have a higher correlation than pctmin80, but after removing the effects of including density into the model, it loses most of it's value.

```
dat_2<-data.frame(crime$crmrte,crime$density,crime$pctmin80,
                  crime$prbconv,crime$urban, crime$west, crime$central)
cor(dat_2)
```

```
               crime.crmrte crime.density crime.pctmin80 crime.prbconv
crime.crmrte      1.0000000    0.72777835     0.18165059   -0.38596559
crime.density     0.7277783    1.00000000    -0.07470698   -0.22791204
crime.pctmin80    0.1816506   -0.07470698     1.00000000    0.06249824
crime.prbconv    -0.3859656   -0.22791204     0.06249824    1.00000000
crime.urban       0.6150631    0.82068254     0.01619569   -0.19709186
crime.west       -0.3803387   -0.19458889    -0.62451443    0.07198455
crime.central     0.1658803    0.35682850    -0.05487554   -0.04640007
               crime.urban  crime.west crime.central
crime.crmrte    0.61506307 -0.38033874    0.16588032
crime.density   0.82068254 -0.19458889    0.35682850
crime.pctmin80  0.01619569 -0.62451443   -0.05487554
crime.prbconv  -0.19709186  0.07198455   -0.04640007
crime.urban     1.00000000 -0.08000341    0.15927023
crime.west     -0.08000341  1.00000000   -0.42986348
crime.central   0.15927023 -0.42986348    1.00000000
```

Finally we trim off the two variables with the lowest significance to produce our final model. Taxpc and pctymle do add value to model but at the cost of brevity and understanding.

```
model_3<-lm(crime$crmrte~crime$density+crime$prbarr+crime$prbconv+
              crime$polpc+crime$pctmin80)
summary(model_3)
```

```

Call:
lm(formula = crime$crmrte ~ crime$density + crime$prbarr + crime$prbconv + 
    crime$polpc + crime$pctmin80)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0183081 -0.0046706 -0.0000045  0.0049657  0.0270131 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.345e-02  3.541e-03   9.447 7.47e-15 ***
crime$density   5.551e-03  7.157e-04   7.756 1.86e-11 ***
crime$prbarr   -6.614e-02  8.497e-03  -7.783 1.65e-11 ***
crime$prbconv  -2.162e-02  2.849e-03  -7.588 4.02e-11 ***
crime$polpc     8.139e+00  1.176e+00   6.923 8.21e-10 ***
crime$pctmin80  3.739e-04  5.538e-05   6.752 1.77e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008593 on 84 degrees of freedom
Multiple R-squared:  0.8046,	Adjusted R-squared:  0.793 
F-statistic: 69.19 on 5 and 84 DF,  p-value: < 2.2e-16
```

```
AIC(model_2)
```

```
[1] -597.6594
```

```
AIC(model_3)
```

```
[1] -593.0191
```

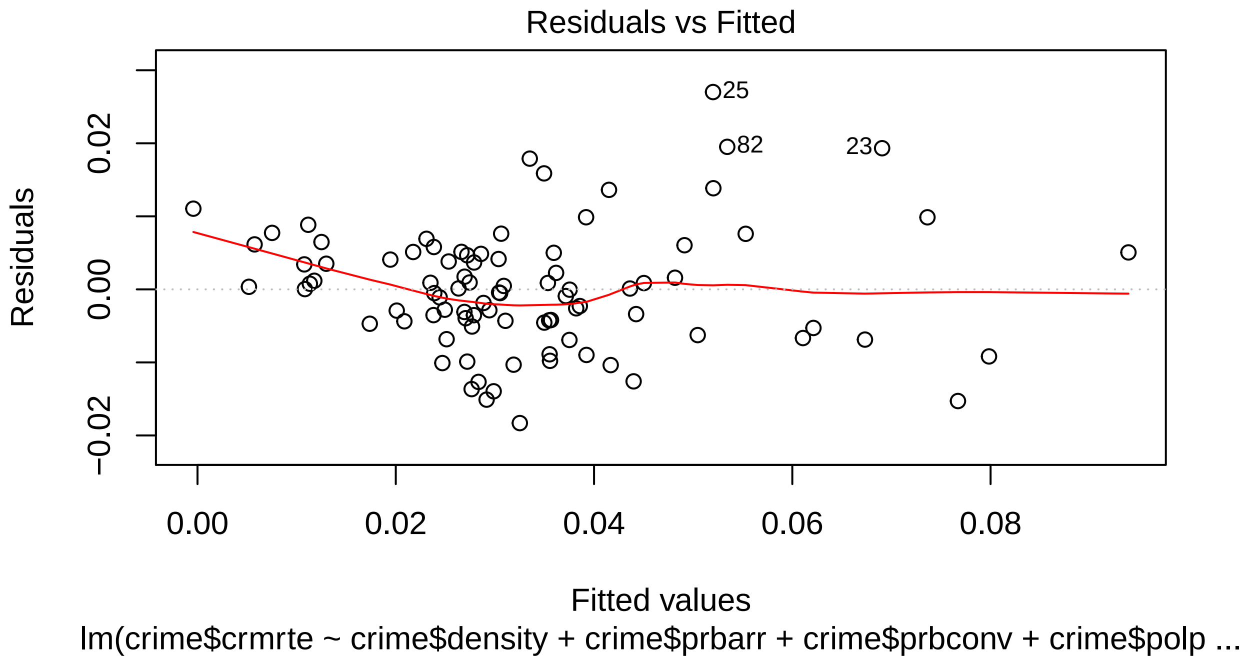
Verify Assumptions

Here we verify the the six assumptions of our model:

1) Linearity of the Parameters
2) Random Sampling
3) No Perfect Multicollinearity
4) Zero Conditional Mean
5) Homoskedasticity
6) Normality of Residuals

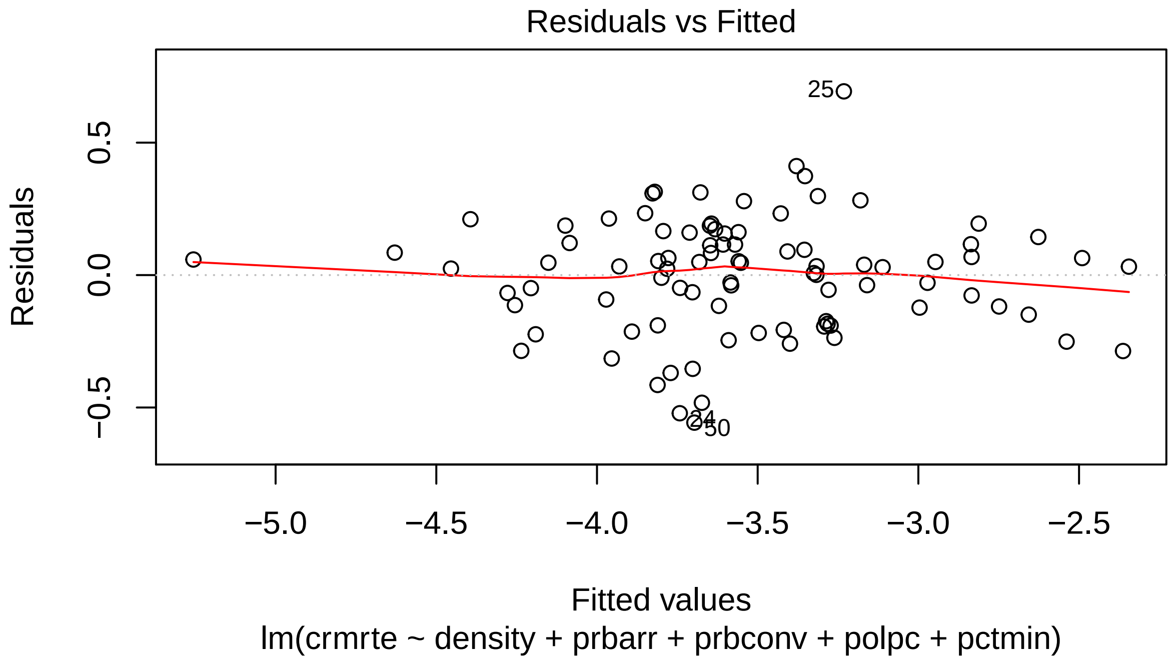
First we check for linearity by looking at the Residuals vs Fitted plot.

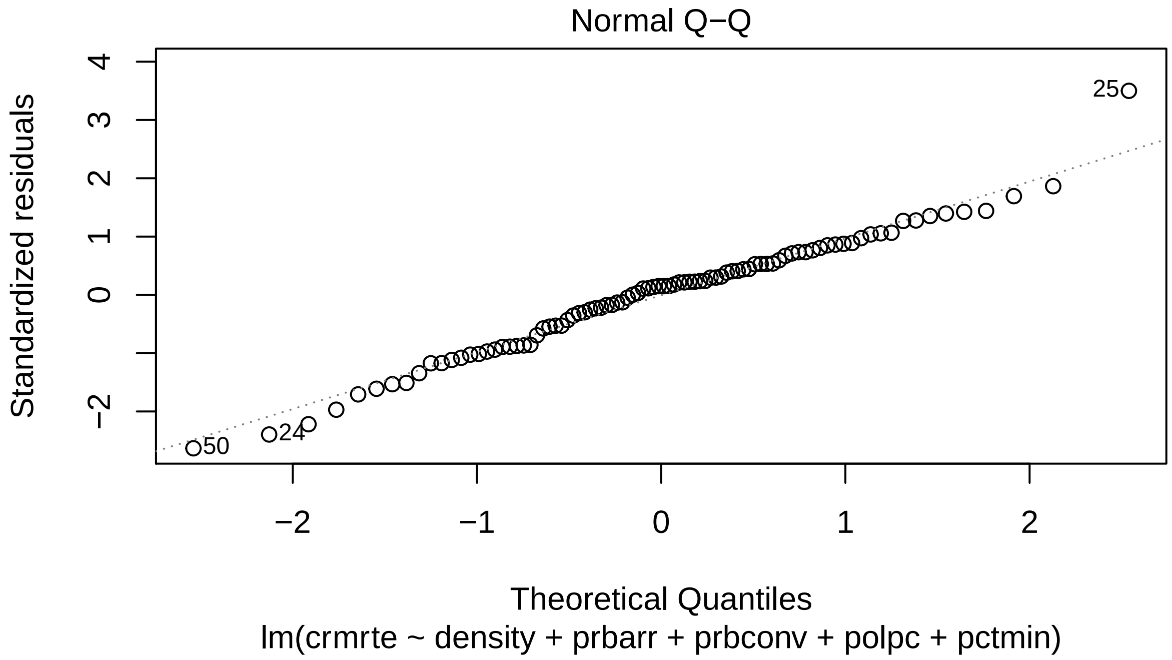
```
plot(model_3, which=1)
```

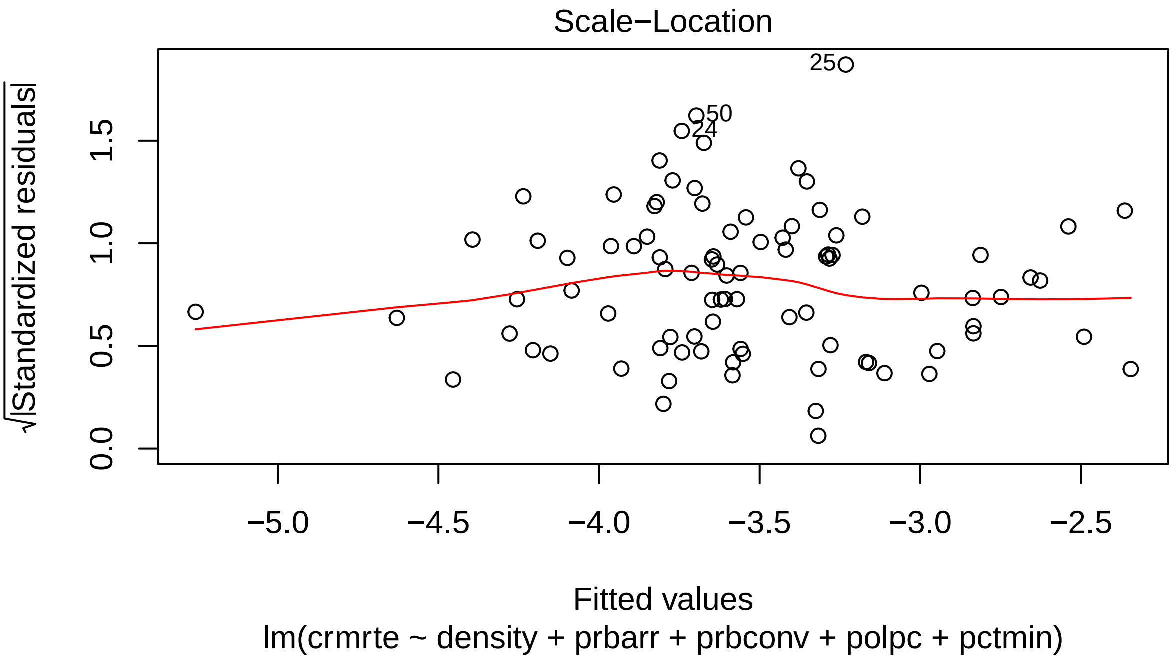


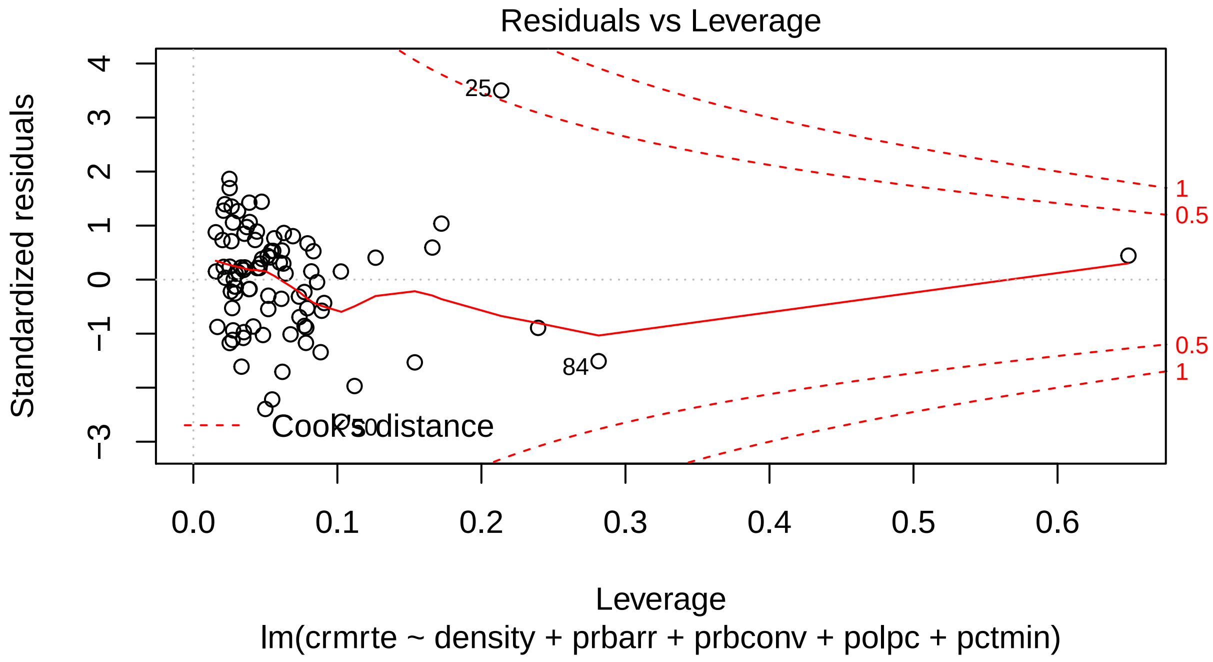
Here we see evidence of nonlinear relationship at the lower end of the range of our dependant variable. We address this by looking at the log-log relationship with respect to crimerte, density, and polpc, our variables that show the strongest evidence of skew.

```
crmrte<-log(crime$crmrte)
density<-log(crime$density)
prbarr<-crime$prbarr
prbconv<-crime$prbconv
polpc<-log(crime$polpc)
pctmin<-crime$pctmin80
model_4<-lm(crmrte~density+prbarr+prbconv+polpc+pctmin)
plot(model_4)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
25
24
50
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
Fitted values
lm(crmrte ~ density + prbarr + prbconv + polpc + pctmin)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−2
25
24
50
−2
−1
0
1
2
Theoretical Quantiles
lm(crmrte ~ density + prbarr + prbconv + polpc + pctmin)





```
summary(model_4)
```

```

Call:
lm(formula = crmrte ~ density + prbarr + prbconv + polpc + pctmin)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.55708 -0.14276  0.03213  0.13805  0.69364 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.105615   0.499340  -0.212    0.833    
density      0.291715   0.038075   7.662 2.87e-11 ***
prbarr      -1.678919   0.205049  -8.188 2.55e-12 ***
prbconv     -0.612226   0.070437  -8.692 2.48e-13 ***
polpc        0.454625   0.072273   6.290 1.37e-08 ***
pctmin       0.012764   0.001424   8.960 7.13e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2235 on 84 degrees of freedom
Multiple R-squared:  0.8435,    Adjusted R-squared:  0.8341 
F-statistic: 90.52 on 5 and 84 DF,  p-value: < 2.2e-16
```

This plot is very strong evidence that the log-log transform takes care of the linearity assumption and the zero conditional mean assumption. Additionally, the normal QQ plot fits extremely well so we can safely assume we have normality of our residuals.

```r
dat_3<-data.frame(log(crime$density),log(crime$polpc),crime$prbarr,
                  crime$prbconv,crime$pctmin80)
vif(dat_3)
```

```
           Variables      VIF
1 log.crime.density. 1.578154
2   log.crime.polpc. 1.303623
3       crime.prbarr 1.420060
4      crime.prbconv 1.109106
5     crime.pctmin80 1.043042
```

To check for multicollinearity we use the measured variance inflation factors shown above. These values are sufficiently low for each of our independant variables so there is very little evidence of multicollinearity.

For the assumption of a random sample, we have to assume that the person gathering the data for the model took proper precautions to gather a truly random sample. We could gather a second sample and compare the distributions of the two samples and see how similar they are, but this would likely be costly and time consuming. For the purposes of this study, we assume that the person gathering the information used due diligence to gather a sample that is representative of the larger population of counties that the candidate intends to represent.

Referring back to the Residuals vs Fitted plot, we see strong evidence of heteroskedasticity and will use robust standard errors when assessing our model from here.

```r
coeftest(model_4, vcov=vcovHC)
```

```

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -0.1056145  0.9616514 -0.1098 0.9128092    
density      0.2917154  0.0639652  4.5605 1.722e-05 ***
prbarr      -1.6789191  0.2852187 -5.8864 7.848e-08 ***
```

```
prbconv     -0.6122263  0.1029381 -5.9475 6.040e-08 ***
polpc        0.4546253  0.1297689  3.5033 0.0007391 ***
pctmin       0.0127636  0.0015183  8.4066 9.274e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
se.model = coeftest(model_4, vcov = vcovHC)[ , "Std. Error"]
stargazer(model_4, type="text", se=list(se.model))
```

```

===============================================
                        Dependent variable:    
                    ---------------------------
                              crmrte           
-----------------------------------------------
density                      0.292***          
                              (0.064)          
                                               
prbarr                       -1.679***         
                              (0.285)          
                                               
prbconv                      -0.612***         
                              (0.103)          
                                               
polpc                        0.455***          
                              (0.130)          
                                               
pctmin                       0.013***          
                              (0.002)          
                                               
Constant                      -0.106           
                              (0.962)          
                                               
-----------------------------------------------
Observations                    90             
R2                             0.843           
Adjusted R2                    0.834           
Residual Std. Error       0.223 (df = 84)      
F Statistic           90.516*** (df = 5; 84)   
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Practical significance.

```
(m_crmrte <- mean(crime$crmrte))
```

```
[1] 0.03350992
```

```
(m_density <- mean(crime$density))
```

```
[1] 1.43793
```

```
(sd_density <- sqrt(var(crime$density))) # standard deviation of density
```

```
[1] 1.519548
```

```
(m_prbarr <-mean(prbarr))
```

```
[1] 0.2952375
```

```
(sd_prbarr <- sqrt(var(prbarr))) # standard deviation of probability of arrest
```

```
[1] 0.137673
```

```
(m_prbconv <- mean(prbconv))
```

```
[1] 0.5508611
```

```
(sd_prbconv <- sqrt(var(prbconv)))  # standard deviation of probability of conviction
```

```
[1] 0.3541942
```

```
(m_polpc <- mean(crime$polpc))
```

```
[1] 0.001707973
```

```
(sd_polpc <- sqrt(var(crime$polpc))) # standard deviation of police per capita
```

```
[1] 0.0009909212
```

```
(m_pctmin80 <- mean(pctmin))
```

```
[1] 25.71285
```

```
(sd_pctmin80 <- sqrt(var(pctmin))) # standard deviation of percent minority, 1980
```

```
[1] 16.98474
```

```
(model_4$coefficients[2]*sd_density) # % change in crime rate with 1 standard deviation % change in den
```

```
  density
0.4432755
```

```
(100*model_4$coefficients[3]*sd_prbarr) # % change in crime rate with 1 standard deviation change in pr
```

```
   prbarr
-23.11418
```

```
(100*model_4$coefficients[4]*sd_prbconv) # % change in crime rate with 1 standard deviation change in p
```

```
 prbconv
-21.6847
```

```
(model_4$coefficients[5]*sd_polpc)  # % change in crime rate with 1 standard deviation % change in polp
```

```
       polpc
0.0004504978
```

```
(100*model_4$coefficients[6]*sd_pctmin80)  # % change in crime rate with 1 standard deviation change in
```

```
  pctmin
21.67867
```

```
(cor(crime$prbarr,crime$polpc))
```

```
[1] 0.4259648
```

Discussion of practical significance.

In order to look at practical significance, we looked at the change in crime rate with a 1 standard deviation change in the statistically significant variables. We observe that density, probability of arrest (prbarr), probability of conviction(prbconv) and percent minority are all practically significant. However, police per capita is not practically significant once the other variables have been accounted for. Of course, since prbarr is correlated with polpc, it's effect might have been picked up by this variable. Intuitively, in order for probability of arrest to be high, police per cap has to be sufficiently high.

Discussion of causality

The clear statistically significant variables showing correlation with crimerate are : population density, probability of arrest, probability of conviction, police per capita and percent minority. There are a few observations wrt these variables that are key to consider before we make policy decisions.

a. In order to make policy changes based on these decisons, we would need an idea of causality, not just correlations. However, it is hard to interpret causality here since several key variables that could affect crime rate are missing here e.g unemployment rate, poverty rate. So it would be very helpful to get data on these variables for doing analysis on causality

b. With current data, we can make an attempt to apply judgement to the relationships that show up as significant to then make policy choices. In this context :

1. While population density and percent minority are significant varibles, there is very little we can do from a policy point of view to change these. However, it is important to include these variables in the model to be able to see the effect of the other variables net of these.

2. Both probabilty of arrest and probability of conviction show significant negative correlation with crime rate. It suggests that the higher the likelihood of being arrested and convicted for a crime, the lesser the likelihood of a crime. ie arrest and conviction are significant deterrants to crime. Hence policy changes that lead to an increase in catching the criminals would be good to reduce crime.

3. Interestingly, average sentence, and probability of prison sentence did not show up as significant, even though they are not very correlated with the other two variables that show the adverse consequence of crime for the criminal ie probability of arrest, and probability of conviction. This suggests that whether the criminal is caught and convicted is more important than the magnitude or nature of the sentence if convicted.
So policy should focus on law enforcement initiatives that lead to greater success in arrest and conviction.
4. Police per capita has a positive correlation with crime rate. Clearly, increasing police cannot lead to increase in crime rate.The magnitude of the effect is not practically significant. In addition, since police per cap is correlated with probability of arrest (correlation = 0.43), it is possible that the effect of police per cap is picked up in the probability of arrest.

Since we want to increase the probability of arrest and conviction, one way to do that is to increase the police force, not reduce it.

Note that it is possible that the desired outcomes can also be achieved by focusing the existing police force on arresting criminals vs other activities like traffic etc.

Conclusions

- While we have 25 variables available to predict crime rate, only 4 were statistically and practically significant ie density, probability of arrest, probability of conviction, percent minority. Police per cap was statistically significant but not practically significant. It's sign is also counterintuitive.
- In order to have a model that can give us unbiased estimates (MLR1-4), and to reduce heteroskedasticity, we had to take the log transform for three variables - crime rate, population density, and police per capita.
- Causality is hard to intepret from the model directly because several potentially important factors that may affect crime rate like unemployment, poverty level are missing from the data. Collecting this data may lead to better causality models.
- Population density is an important variable amongst those already collected in predicting crime per cap. Since this is not under policy control, we can't use this to drive policy. However, it is important to include this variable in the model so that we can see the effect of the other variables controlling for population density.
- The significant variables for which we can do something via policy are probability of arrest, and probability of conviction. Interestingly, length of sentence did not matter once criminals were arrested and convicted.
- So policy change should be directed to increasing the likelihood of a criminal being caught and convicted irrespective of the size of the crime. This is similar to the "no broken windows" policy implemented in New York in the 90's - where every broken window / suspicious activity was investigated and this led to a dramatic reduction in crime in the city.