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.

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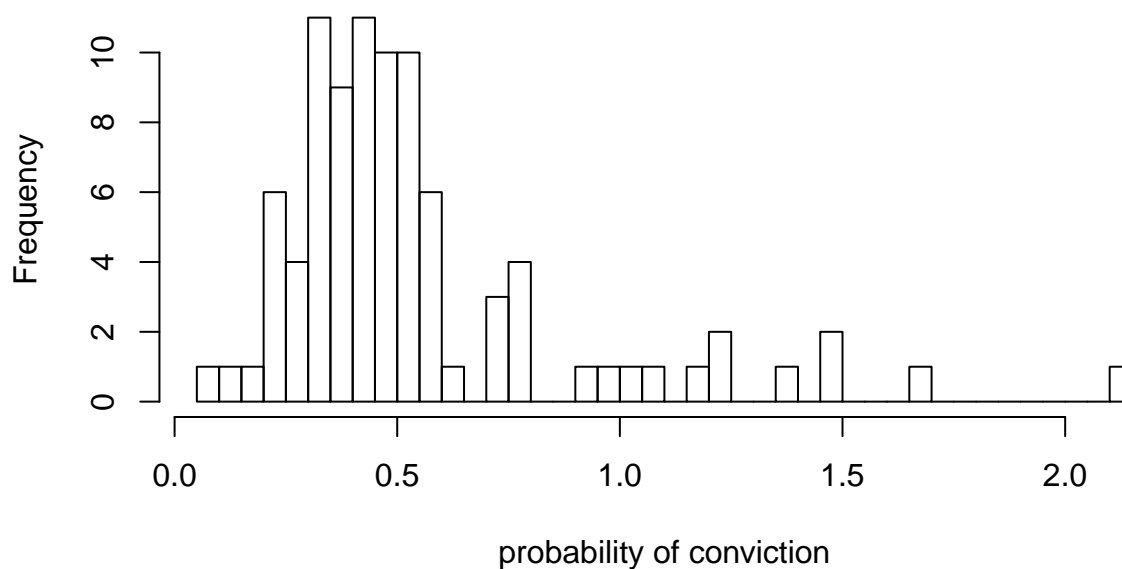
Crime Rate Hist
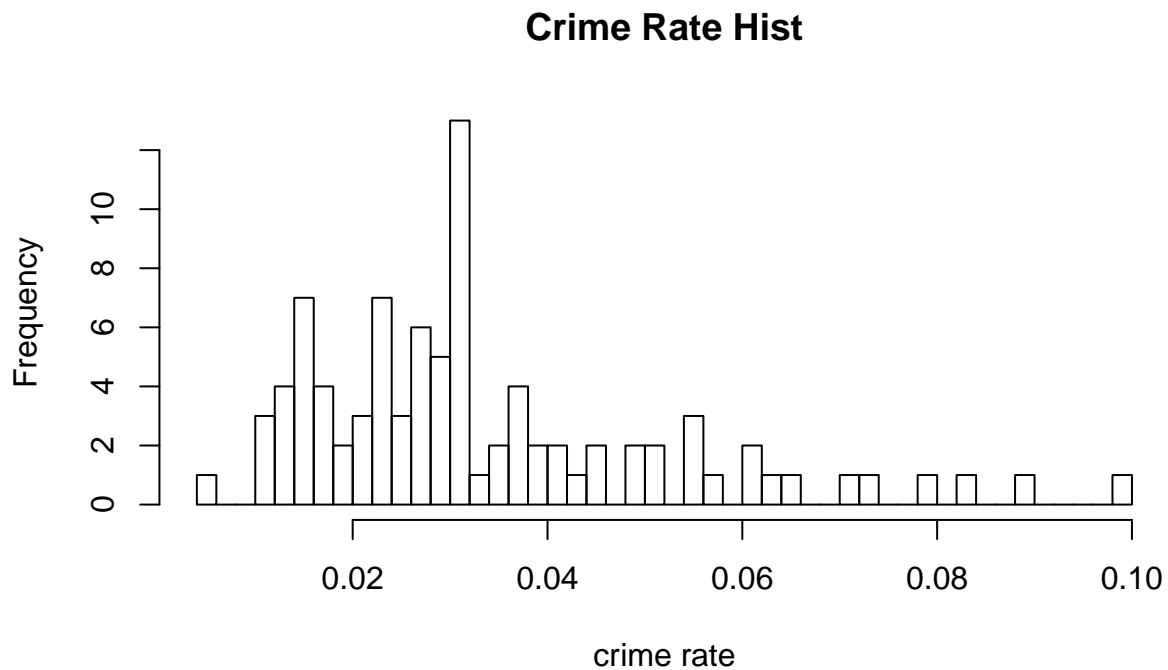

```
hist(crime$prbarr, breaks = 50, xlab = "probability of arrest", ylab = "Frequency", main = "Probability
```
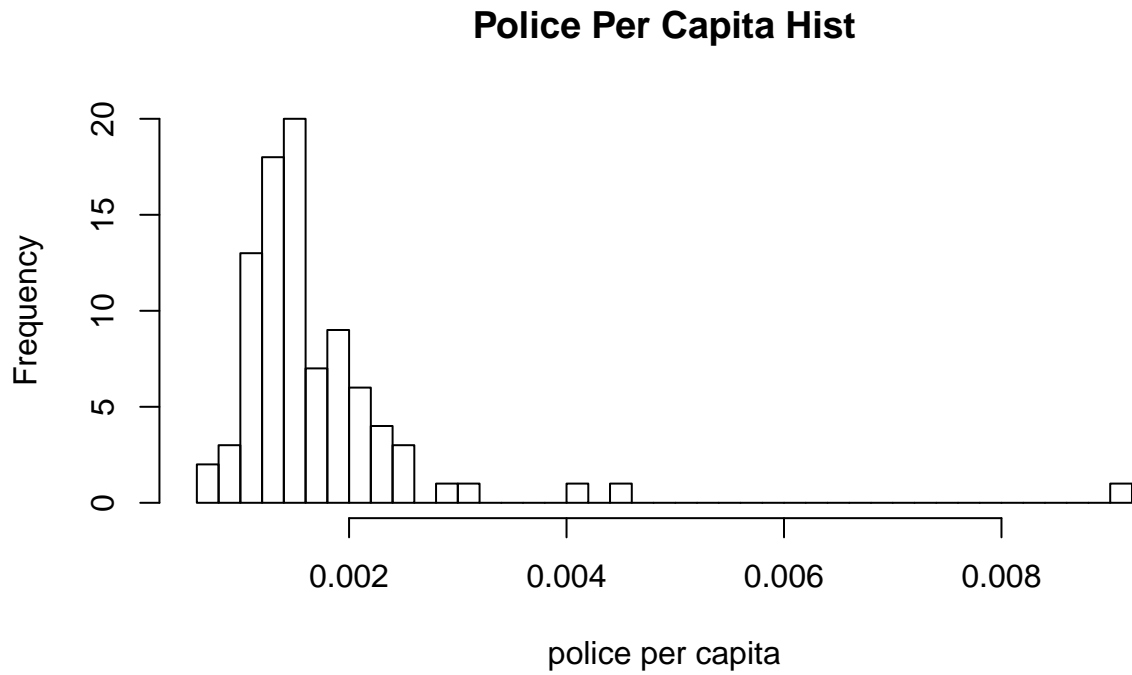
Probability of Arrest Hist
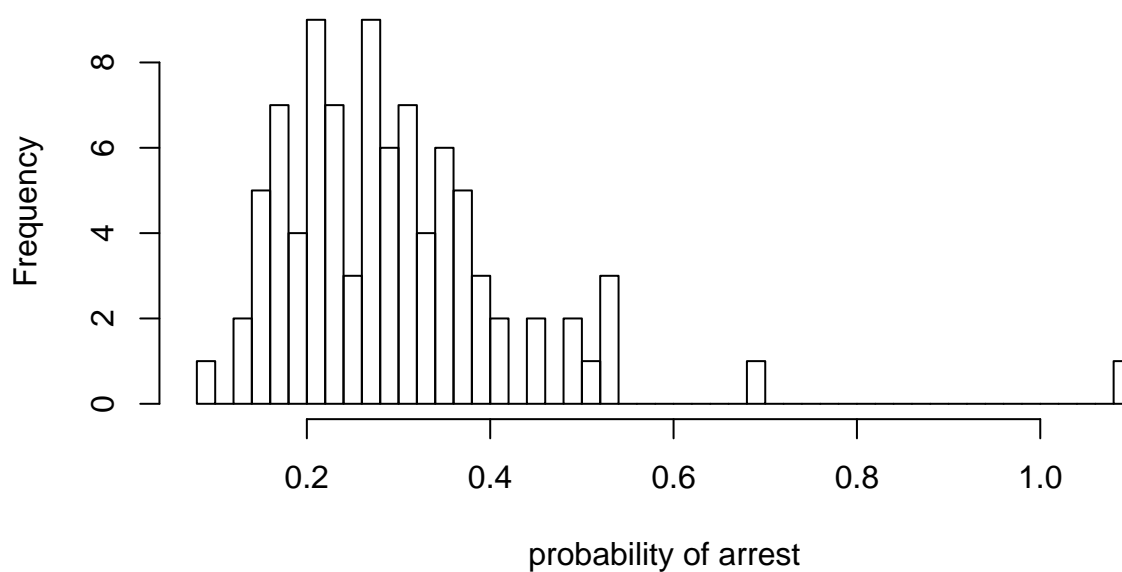

```r
hist(crime$prbconv, breaks = 50, xlab = "probability of conviction", ylab = "Frequency", main = "Probab
```
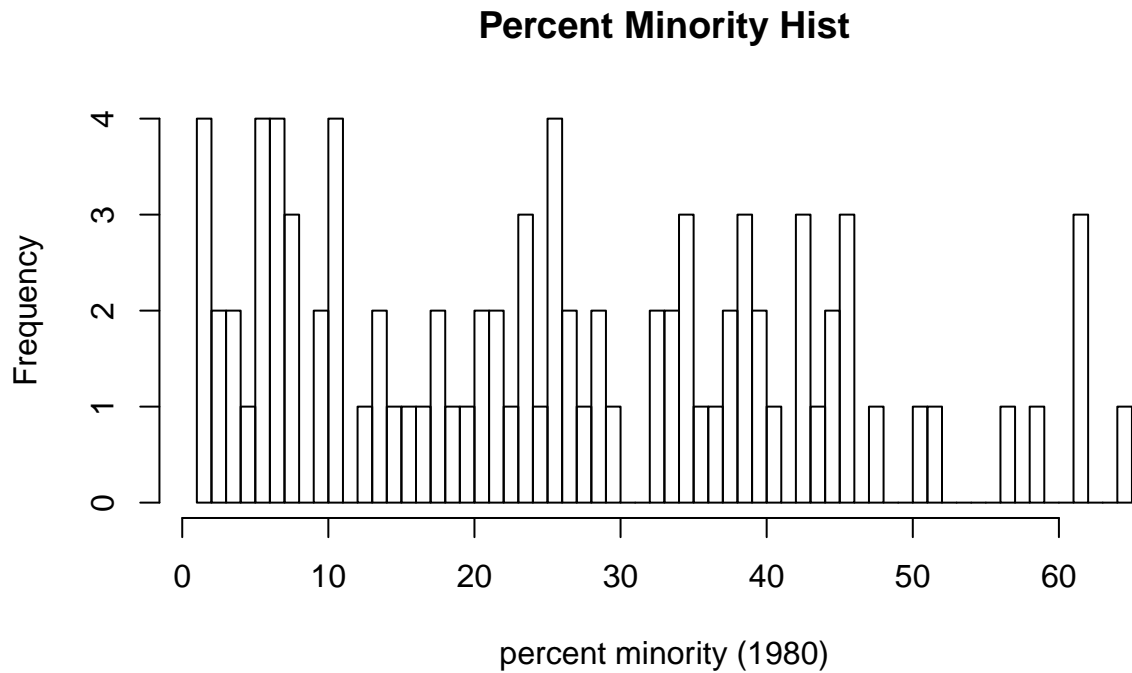
Probabiity of Conviction Hist


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

Police Per Capita Hist


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

Percent Minority Hist

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.
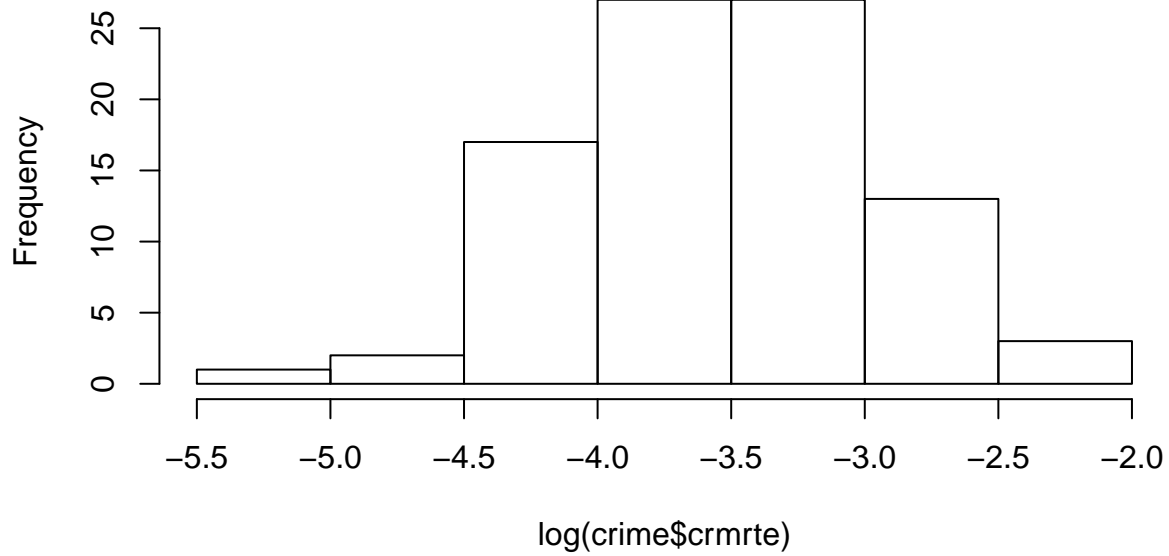
```
hist(crime$wser)
```

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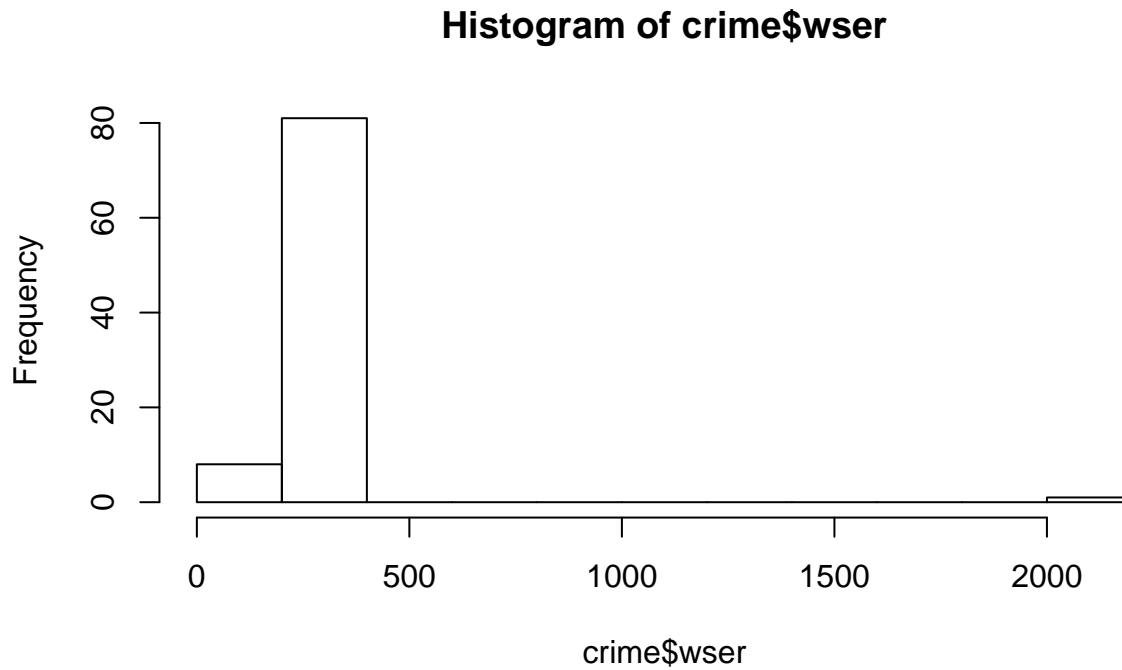
For some of the most skewed variables, polpc, density, and crmrte, we also look at the log transform.

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

Histogram of log(crime$crmrte)


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

Histogram of log(crime$density)


```r
hist(log(crime$polpc))
```
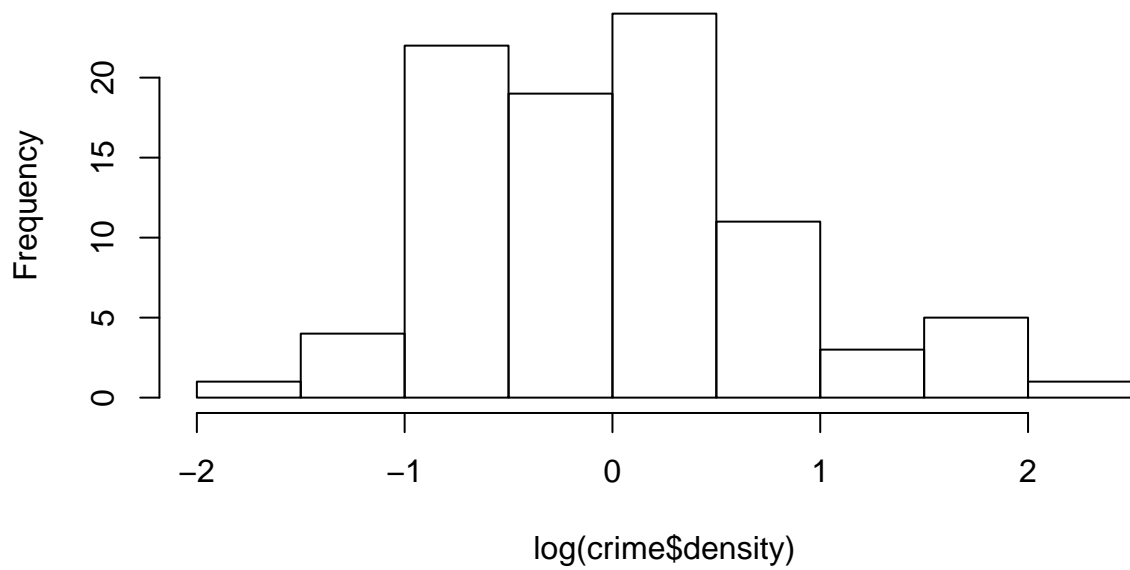
Histogram of log(crime$polpc)
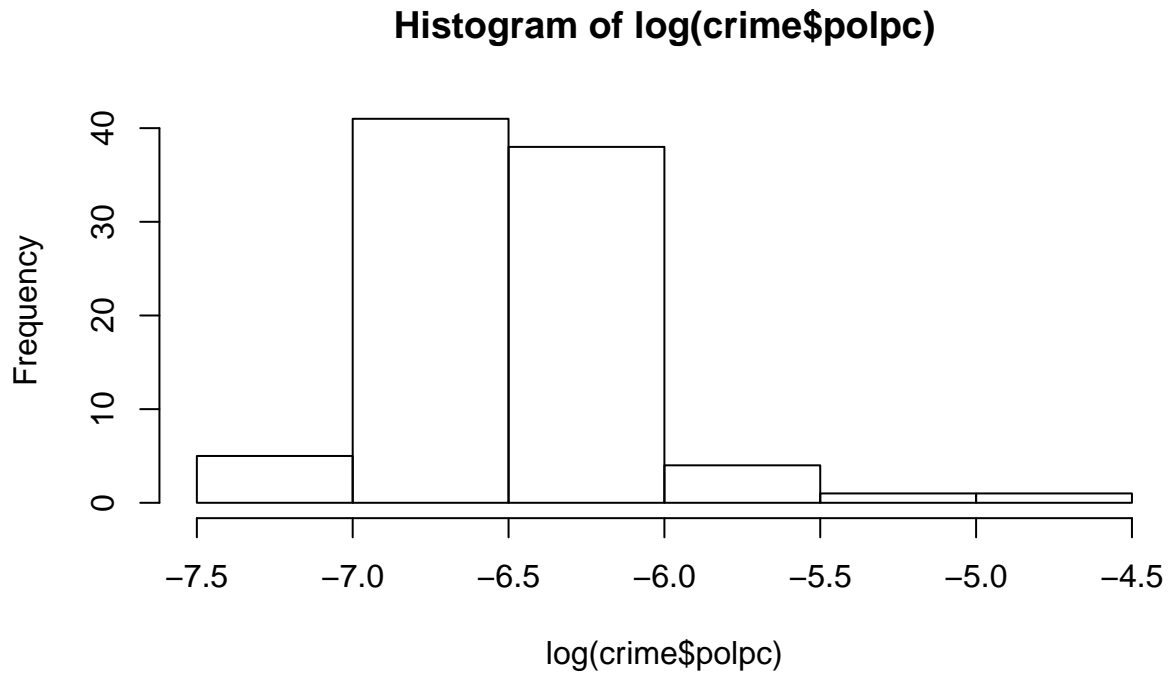
Here, we see that these variables take on a near normal distribution under a log transform.

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

```
scatterplotMatrix(~crmrte + prbarr + prbconv + prbpris + avgsen, data=crime)
```
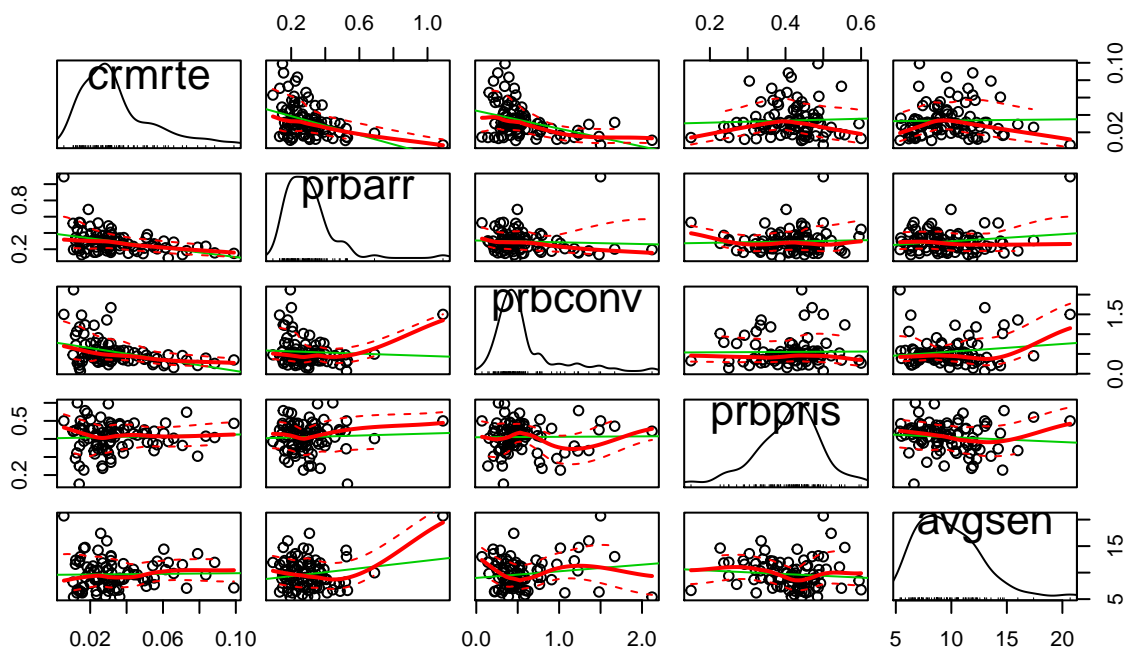
```r
scatterplotMatrix(~crmrte + polpc + density + taxpc, data=crime)
```
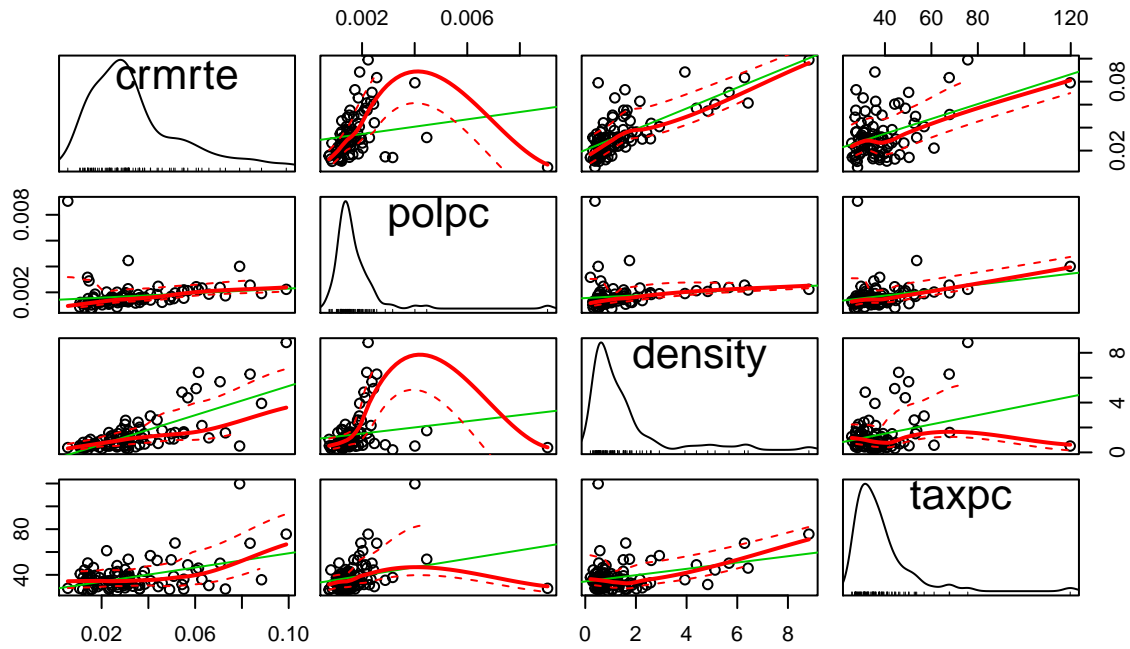


```r
scatterplotMatrix(~crmrte + pctmin80 + pctymle, data=crime)
```
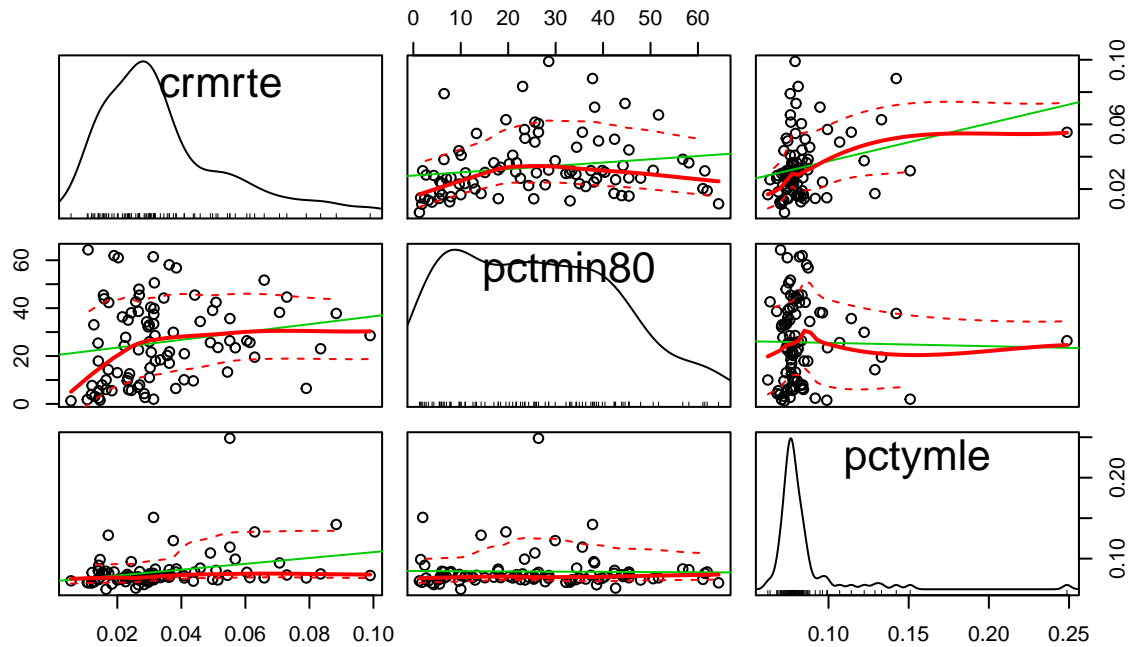
```
scatterplotMatrix(~polpc + density + taxpc + pctmin80 + pctymle, data=crime)
```
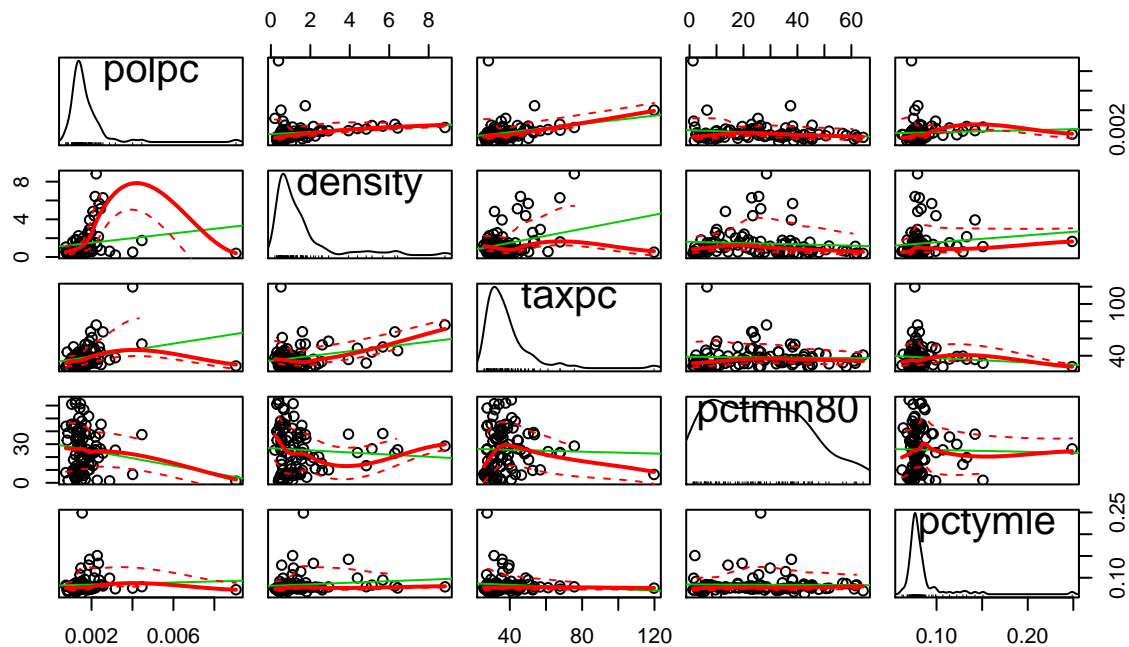

Finally, we added to this by analyzing bivariate and multivariate relationships between a number of variables and combinations of variables. We looked at the R squared values, adjusted R squared values, p values, and correlation values for significance. Several of these are included below.

```
plot(crime$polpc, crime$crmrte)
model3 <- lm(crmrte~polpc, data = crime)
model3
```

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

Coefficients:
(Intercept)        polpc
    0.02806      3.18839
```

```
abline(model3)
```

```
summary(model3)
```
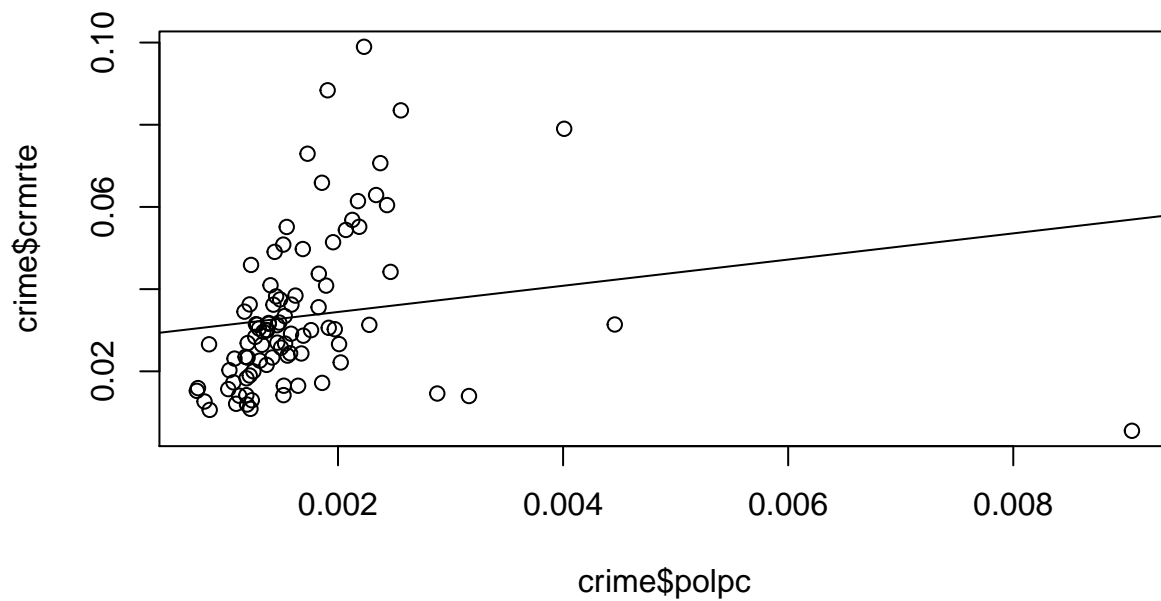
```
Call:
lm(formula = crmrte ~ polpc, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.051400 -0.011799 -0.003837  0.006455  0.063787

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02806    0.00395   7.105 2.99e-10 ***
polpc        3.18839    2.00318   1.592    0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01873 on 88 degrees of freedom
Multiple R-squared:  0.02798,   Adjusted R-squared:  0.01694
F-statistic: 2.533 on 1 and 88 DF,  p-value: 0.115
```

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

```
[1] 0.1672816
```

```
plot(crime$density, crime$crmrte)
model4 <- lm(crmrte~density, data = crime)
model4
```

```
Call:
```

```
lm(formula = crmrte ~ density, data = crime)

Coefficients:
(Intercept)      density
   0.020503     0.009046
```
```r
abline(model4)
```
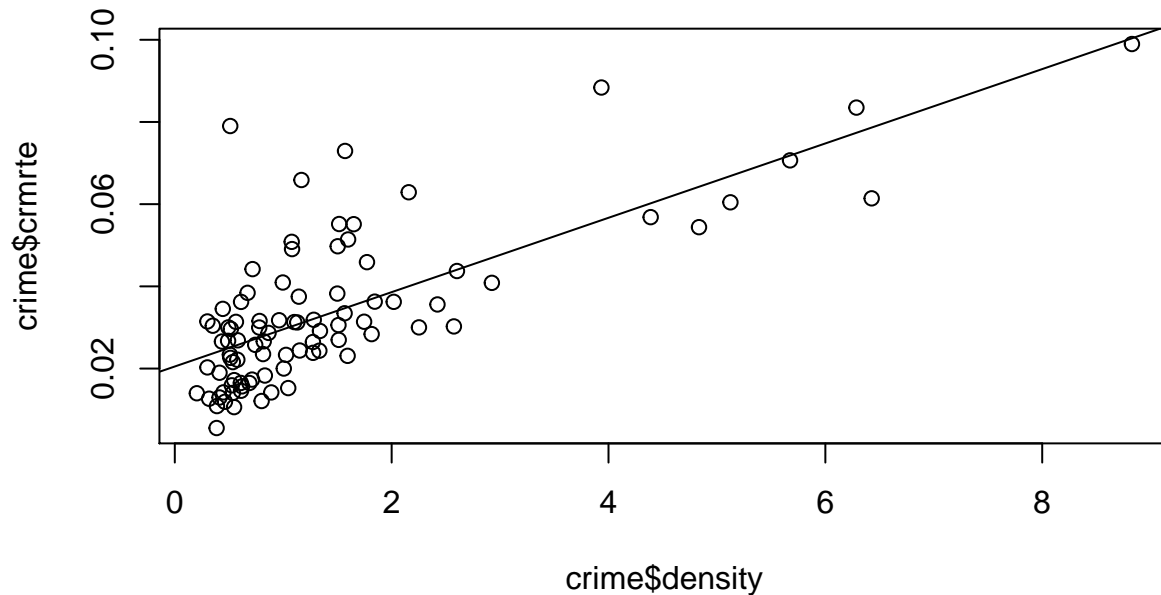


```r
summary(model4)
```

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

```
[1] 0.7277783
```

```
plot(crime$taxpc, crime$crmrte)
model5 <- lm(crmrte~taxpc, data = crime)
model5
```

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

Coefficients:
(Intercept)        taxpc
  0.0088444    0.0006464
```

```
abline(model5)
```
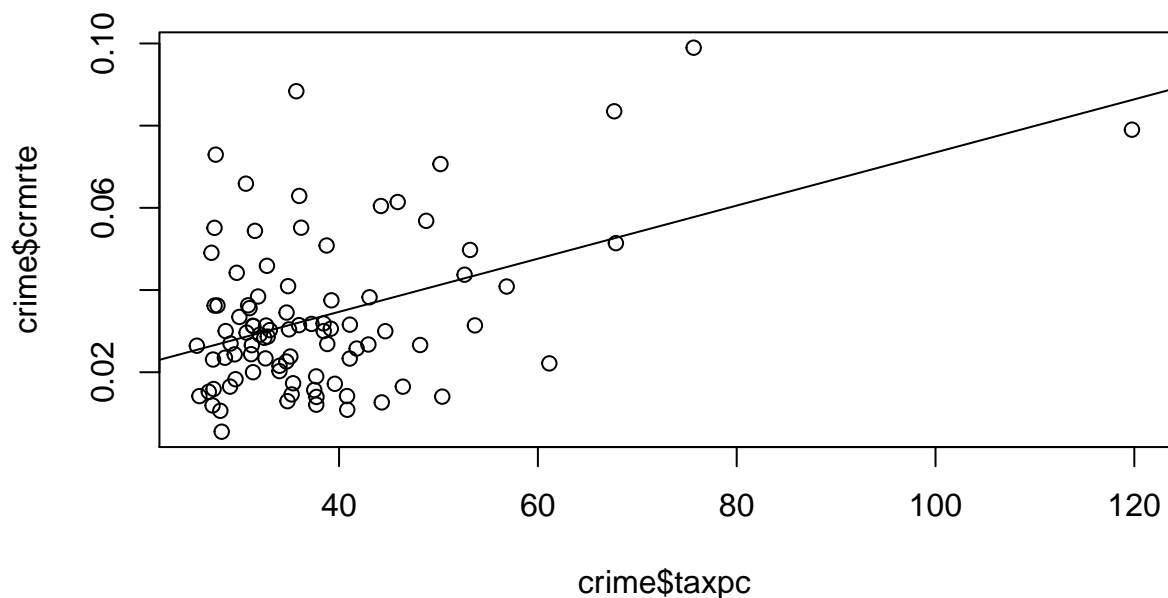


```
summary(model5)
```

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

Residuals:
     Min        1Q    Median        3Q       Max
-0.027343 -0.010886 -0.002133  0.006679  0.056466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0088444  0.0055339   1.598    0.114
taxpc       0.0006464  0.0001372   4.710 9.18e-06 ***
---
```

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

Residual standard error: 0.01697 on 88 degrees of freedom
Multiple R-squared:  0.2013,    Adjusted R-squared:  0.1923
F-statistic: 22.19 on 1 and 88 DF,  p-value: 9.181e-06
```

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

```
[1] 0.4487151
```

```r
plot(crime$pctmin80, crime$crmrte)
model6 <- lm(crmrte~pctmin80, data = crime)
model6
```

```
Call:
lm(formula = crmrte ~ pctmin80, data = crime)

Coefficients:
(Intercept)      pctmin80
   0.028316      0.000202
```
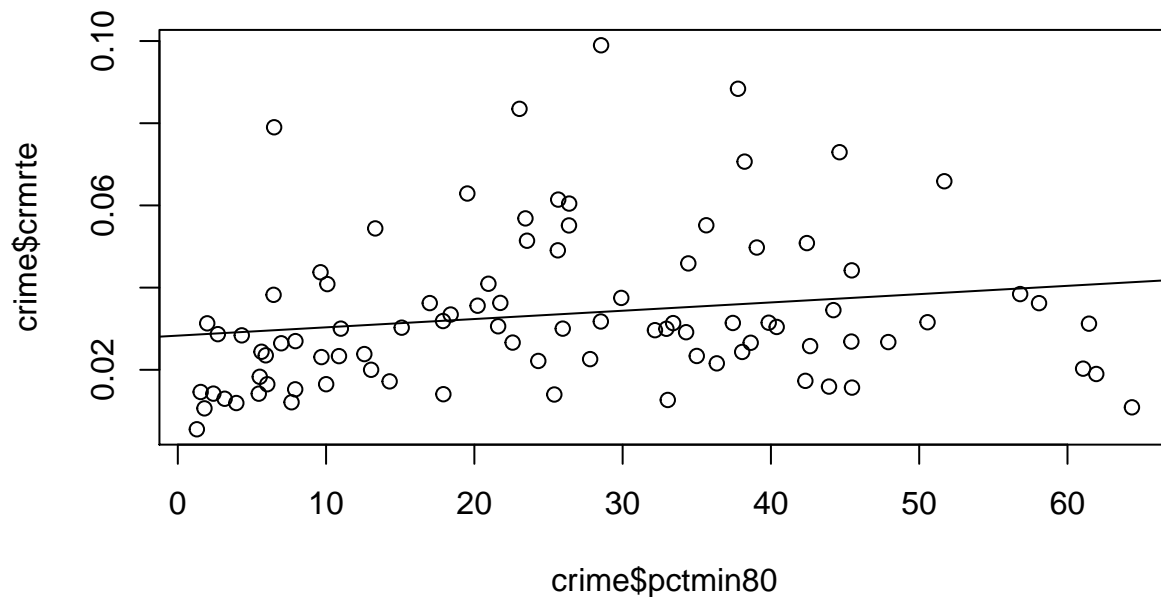
```r
abline(model6)
```



```r
summary(model6)
```

```
Call:
lm(formula = crmrte ~ pctmin80, data = crime)

Residuals:
     Min       1Q    Median       3Q       Max
```

13

```
-0.030444 -0.011928 -0.004692  0.007969  0.064884
```

Coefficients:
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0283161  0.0035861   7.896 7.55e-12 ***
pctmin80    0.0002020  0.0001166   1.733   0.0866 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.01868 on 88 degrees of freedom
Multiple R-squared:  0.033,	Adjusted R-squared:  0.02201
F-statistic: 3.003 on 1 and 88 DF,  p-value: 0.08662
```

```r
cor(crime$crmrte, crime$pctmin80, use="pairwise.complete.obs")
```

```
[1] 0.1816506
```

```r
model21 <- lm(crmrte~polpc + density + taxpc, data = crime)
summary(model21)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.016453 -0.007955 -0.002938  0.003917  0.042047

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0087751  0.0042403   2.069  0.04150 *
polpc       -0.1086169  1.3896572  -0.078  0.93788
density      0.0080949  0.0009183   8.816 1.17e-13 ***
taxpc        0.0003480  0.0001094   3.181  0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.01243 on 86 degrees of freedom
Multiple R-squared:  0.5814,	Adjusted R-squared:  0.5668
F-statistic: 39.82 on 3 and 86 DF,  p-value: 3.122e-16
```

```r
model22 <- lm(crmrte~polpc + density, data = crime)
summary(model22)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.025716 -0.009161 -0.002382  0.004857  0.051552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0189880  0.0029109   6.523 4.38e-09 ***
polpc       0.9731434  1.4162336   0.687    0.494
```

```
density      0.0089433   0.0009235    9.684 1.79e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01307 on 87 degrees of freedom
Multiple R-squared:  0.5322,    Adjusted R-squared:  0.5214
F-statistic: 49.49 on 2 and 87 DF,  p-value: 4.442e-15
```

```
model29 <- lm(crmrte~pctmin80 + pctymle + density, data = crime)
summary(model29)
```

```
Call:
lm(formula = crmrte ~ pctmin80 + pctymle + density, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.021145 -0.006220 -0.001704  0.003676  0.060281

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.920e-04  5.072e-03   -0.117  0.90735
pctmin80     2.665e-04  7.314e-05    3.643  0.00046 ***
pctymle      1.709e-01  5.318e-02    3.213  0.00185 **
density      8.966e-03  8.228e-04   10.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01169 on 86 degrees of freedom
Multiple R-squared:  0.6301,    Adjusted R-squared:  0.6172
F-statistic: 48.83 on 3 and 86 DF,  p-value: < 2.2e-16
```

```
model31 <- lm(crmrte~pctmin80 + density, data = crime)
summary(model31)
```

```
Call:
lm(formula = crmrte ~ pctmin80 + density, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.023115 -0.007062 -0.002932  0.003894  0.059163

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.340e-02  2.737e-03    4.896 4.47e-06 ***
pctmin80    2.639e-04  7.695e-05    3.430 0.000926 ***
density     9.266e-03  8.601e-04   10.773  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0123 on 87 degrees of freedom
Multiple R-squared:  0.5857,    Adjusted R-squared:  0.5762
F-statistic: 61.49 on 2 and 87 DF,  p-value: < 2.2e-16
```

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.

```
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
```
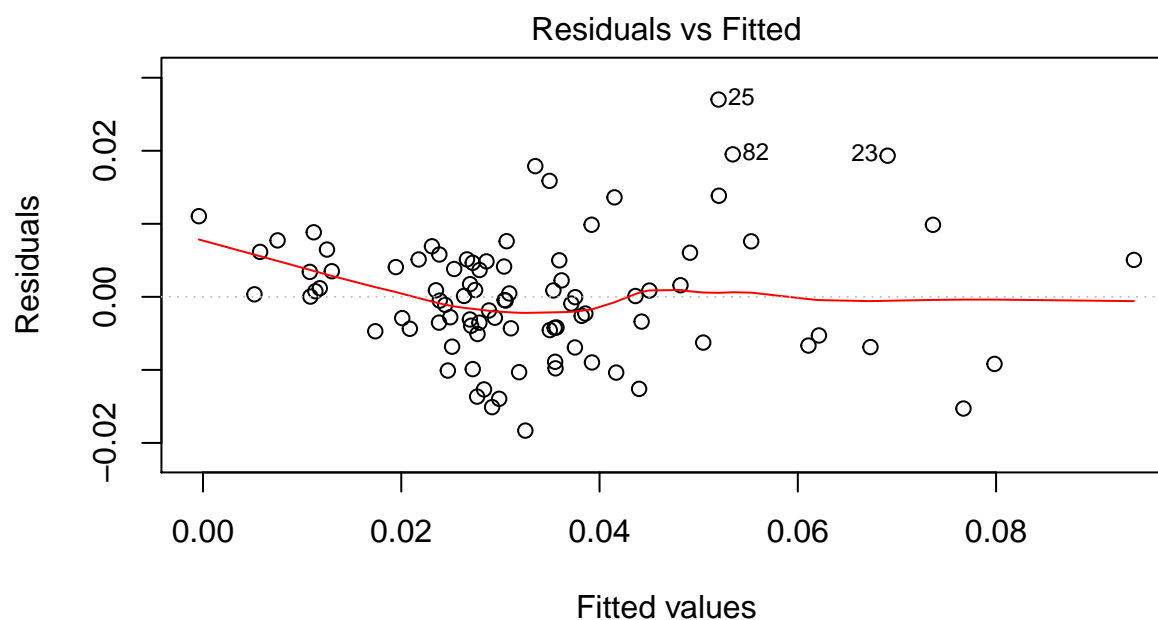
```
AIC(model_3)
```

```
[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)
```


Residuals vs Fitted

lm(crime$crmrte ~ crime$density + crime$prbarr + crime$prbconv + crime$polp ...

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

Fitted values
lm(crmrte ~ density + prbarr + prbconv + polpc + pctmin)

Normal Q–Q

Theoretical Quantiles
lm(crmrte ~ density + prbarr + prbconv + polpc + pctmin)

Scale−Location

√|Standardized residuals|

Fitted values
lm(crmrte ~ density + prbarr + prbconv + polpc + pctmin)

Residuals vs Leverage

Standardized residuals

Cook's distance

Leverage
lm(crmrte ~ density + prbarr + prbconv + polpc + pctmin)

```
summary(model_4)
```
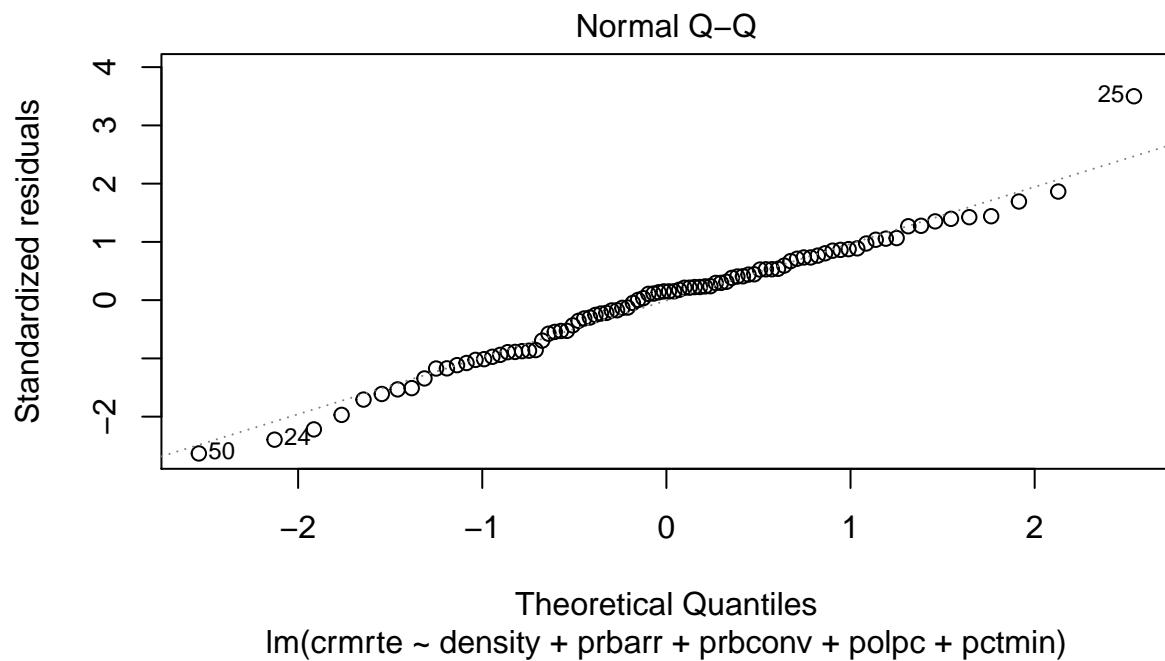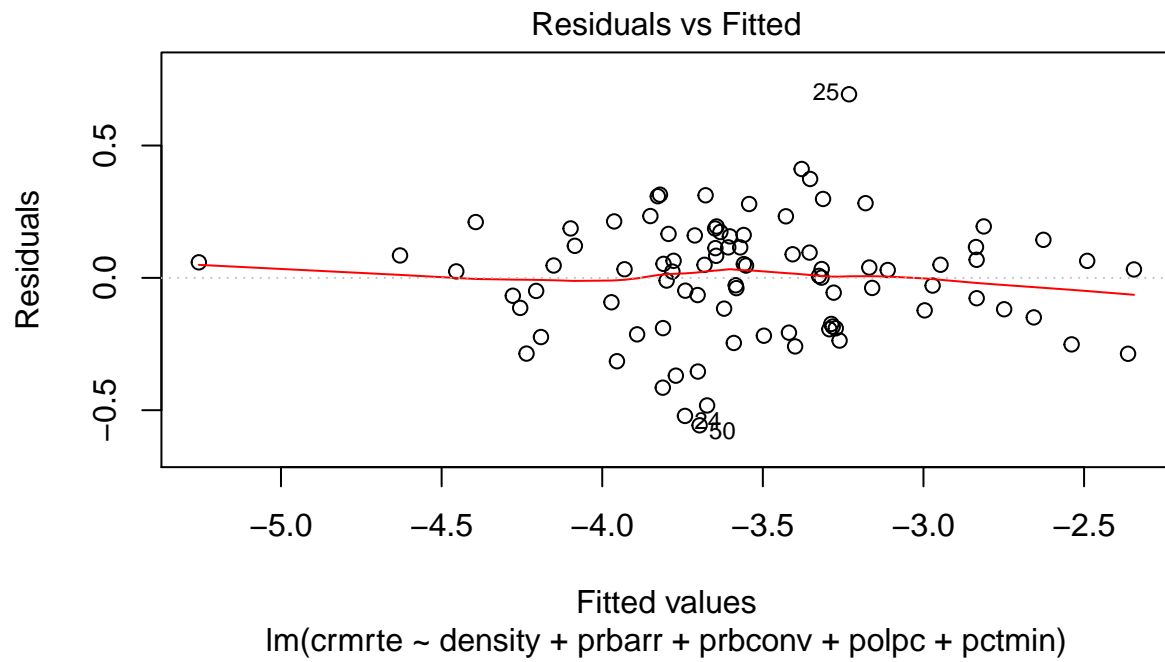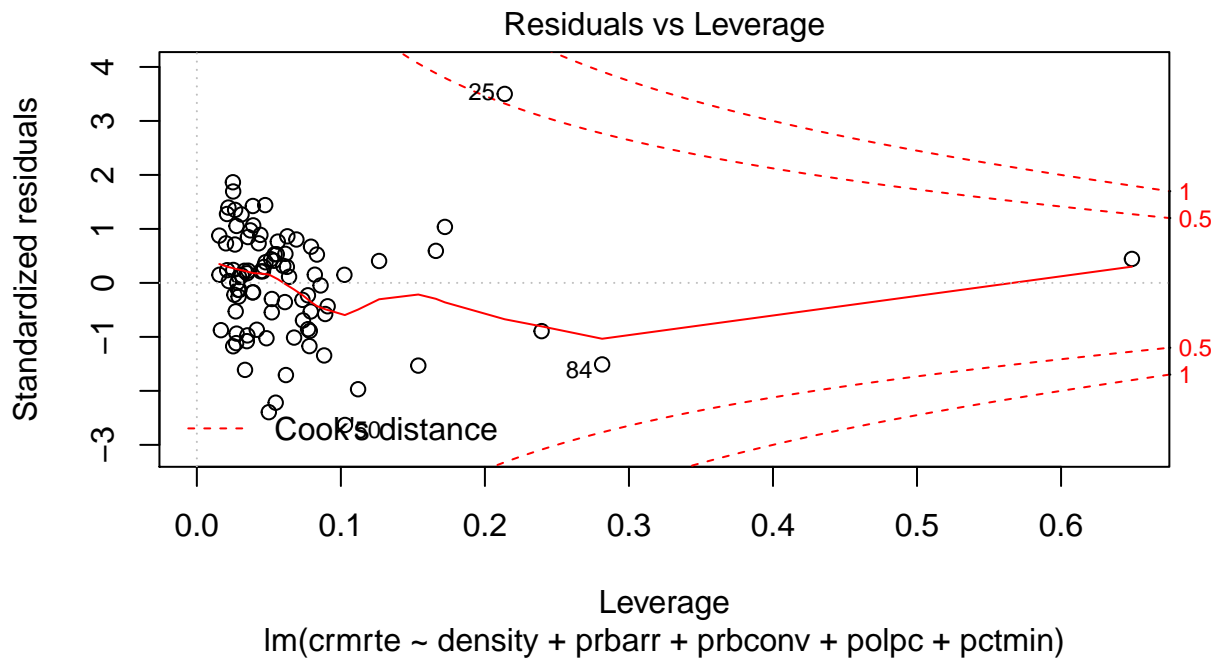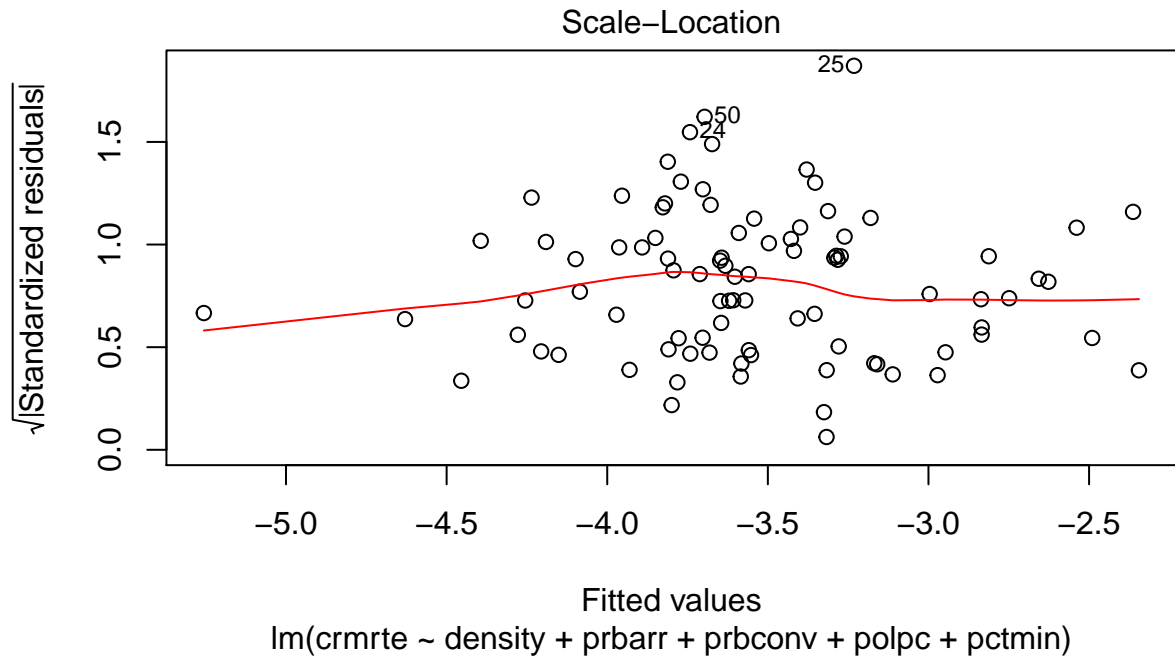
```
Call:
lm(formula = crmrte ~ density + prbarr + prbconv + polpc + pctmin)
```

```
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 would truly 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 f 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
```

```r
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
```

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

```
[1] 0.03350992
```

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

```
[1] 1.43793
```

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

```
[1] 1.519548
```

```r
(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)))
```

```
[1] 0.0009909212
(m_pctmin80 <- mean(pctmin))
```

```
[1] 25.71285
(sd_pctmin80 <- sqrt(var(pctmin)))
```

```
[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 these variables. 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 log transfrom three variables - crime rate, population density, and police per capita.

- Causality is hard to intepret from the model directly because several potentially important factors hat may affect crim 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.