Lab 4

Shan He, Joanna Huang, Tiffany Jaya

17 December 2017

Introduction

This year, on October 18, 2017, Law Enforcement Leaders urged Attorney General Jeff Sessions to reconsider his stance on reverting back to "overly punitive" approaches of the 1980s and 1990s to reduce crime. Since President Trump believes that America is in the midst of a national crime wave, Sessions thought a more conservative approach of deterrence through arrests, incapacitation through imprisonment, harsh sentencing and higher police per capita would lead to lower crime rates overall. However, police chiefs who have first hand decades of experience on the front lines learned that these tactics are ineffective to reduce crime.

In this paper, we will explore whether the conservative approach to crime effectively reduce crime rates. We began by exploring North Carolina's crime dataset of 1987 when "overly punitive" approaches of the 1980s and 1990s would have taken place and analyzed the determinants of crime based on the research question: Does the conservative approach of deterrence through arrests, incapacitation through imprisonment, harsh sentencing and higher police per capita lead to lower crime rates? We will list out the limitations of our analysis, including any estimates that suffer from endogeneity bias, and generate policy suggestions based on our findings.

Exploratory Data Analysis

```r
# load the data
data <- read.csv("crime.csv")
# verify that it only contains data from 1987
unique(data$year)
```

```
## [1] 87
```

```r
# list number of counties
length(unique(data$county))
```

```
## [1] 90
```

```r
# list number of western, central, and urban counties
c(sum(data$west == 1), sum(data$central == 1), sum(data$urban == 1))
```

```
## [1] 21 34  8
```

```r
# list number of western & urban counties and central & urban counties
c(sum(data$west == 1 & data$urban == 1), sum(data$central == 1 & data$urban == 1))
```

```
## [1] 1 5
```

```r
# verify number of missing values
colSums(sapply(data, is.na))
```

```
##        X   county     year   crmrte   prbarr  prbconv  prbpris   avgsen
##        0        0        0        0        0        0        0        0
##    polpc  density    taxpc     west  central    urban pctmin80     wcon
##        0        0        0        0        0        0        0        0
```

```
##     wtuc     wtrd     wfir     wser     wmfg     wfed     wsta     wloc
##        0        0        0        0        0        0        0        0
##      mix  pctymle
##        0        0
```

The dataset contains 90 counties from North Carolina, all of which is collected in 1987. Out of the 90 counties, 21 are from western NC (out of which 1 is also urban), 34 are from central NC (out of which 5 is also urban), and 8 are considered urban counties. There are no missing values which will make our analysis easier.

```r
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
```

Most of the variables appear to be within a reasonable range, except for *prbarr* and *prbconv*, which have probability values greater than 1.

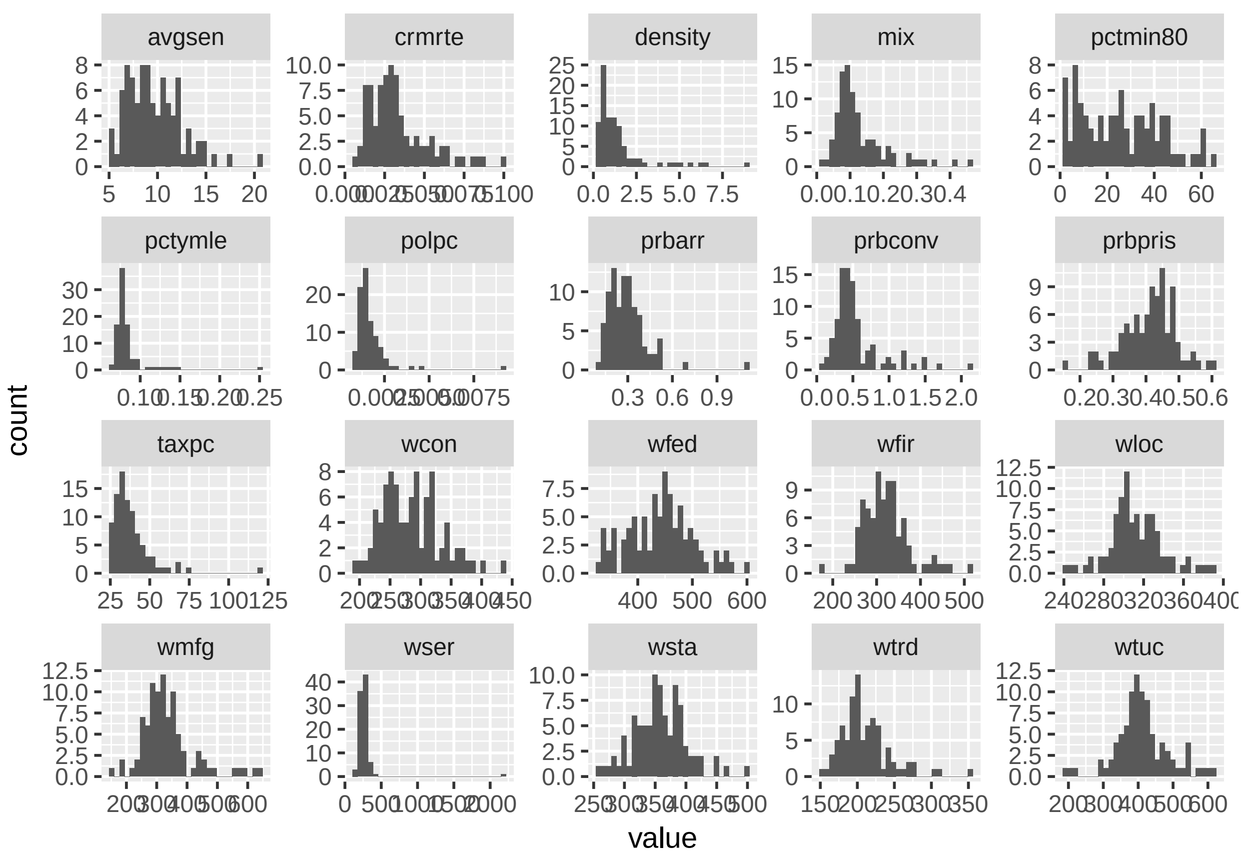
```r
# list number of probabilities (prbarr, prbconv, prbpris, mix) that are not in range [0, 1]
c(sum(data$prbarr < 0 | 1 < data$prbarr), sum(data$prbconv < 0 | 1 < data$prbconv),
  sum(data$prbpris < 0 | 1 < data$prbpris), sum(data$mix < 0 | 1 < data$mix))
```

```
## [1]  1 10  0  0
```

prbarr and *prbconv* contain 1 and 10 datapoints respectively that do not conform to the probability assumption. We will take these outliers into consideration when choosing variables for our models.

We then plot each numeric variable in a histogram to see its sample distribution.

```r
# plot every variable except X, county, year, west, central, urban
num.data <- data[!(names(data) %in% c("X", "county", "year", "west", "central", "urban"))]
ggplot(gather(num.data), aes(value)) +
       facet_wrap(~key, scales="free") +
       geom_histogram()
```




```r
skewness(num.data)
```

```
##     crmrte     prbarr    prbconv     prbpris     avgsen      polpc
## 1.28174888 2.52529596 2.03950599 -0.45254022 1.00116340 4.98348795
```

```
##     density       taxpc    pctmin80        wcon        wtuc        wtrd 
##  2.65301071  3.29057447  0.36566169  0.60680223  0.06819768  1.46120657 
##        wfir        wser        wmfg        wfed        wsta        wloc 
##  0.82063146  8.69918165  1.42253166  0.13223761  0.36236826  0.29513808 
##         mix     pctymle 
##  1.91657046  4.56069073
```

Most of the sample distributions appear to be positively skewed. When choosing the variables for our regression models, we will consider logarithmic transformations if the interpretations make sense.

From the histograms, we also see several notable outliers. We are under the impression that a county which has an outlier in one variable will likely have an outlier in another variable. For this reason, we have listed counties which have repeated outliers when we iterate through the entire numeric variables.

```r
# iterate through each numeric variable and list the outlier counties and their respective frequency
county.ids <- c()
for(var in num.data) {
  var.out <- boxplot.stats(var)$out
  county.ids <- c(county.ids, data[var %in% var.out, ]$county)
}
table(county.ids)
```

```
## county.ids
##   1   3   5   7  11  19  35  39  49  51  53  55  63  67  69  71  79  81 
##   1   1   1   1   2   4   2   2   1   3   1   3   5   1   3   2   1   2 
##  85  87  93  99 105 111 113 115 119 123 127 129 131 133 135 137 139 143 
##   1   1   1   2   1   1   1   5  10   1   2   3   1   1   2   2   1   2 
## 147 149 169 173 175 181 183 185 187 189 195 197 
##   1   1   1   4   1   2   4   2   1   1   2   1
```

```r
# list the most extreme outlier
outlier(num.data)
```

```
##       crmrte       prbarr      prbconv      prbpris       avgsen 
##   0.09896590   1.09090996   2.12121010   0.15000001  20.70000076 
##        polpc      density        taxpc     pctmin80         wcon 
##   0.00905433   8.82765198 119.76145172  64.34819794 436.76663208 
##         wtuc         wtrd         wfir         wser         wmfg 
## 187.61726379 354.67611694 509.46551514 2177.06811523 646.84997559 
##         wfed         wsta         wloc          mix      pctymle 
## 597.95001221 499.58999634 388.08999634   0.46511629   0.24871162
```

One outlier that is interesting to note is the weekly wage in the service industry for county with id 185, $2177.10.

```r
summary(data$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   133.0   229.3   253.1   275.3   277.6  2177.1
```

It is approximately eight times higher than the median. We do not know if the value is inputted incorrectly or if the county in general is making a weekly wage of $2177.10 in the service industry.

Research Question

James Q. Wilson and George Kelling's "broken windows theory" in 1982 led to a nation-wide movement for stricter crime-fighting policies between the 1980s and 1990s. The theory states:

if the first broken window in a building is not repaired, then people who like breaking windows will assume that no one cares about the building and more windows will be broken. Soon the building will have no windows....

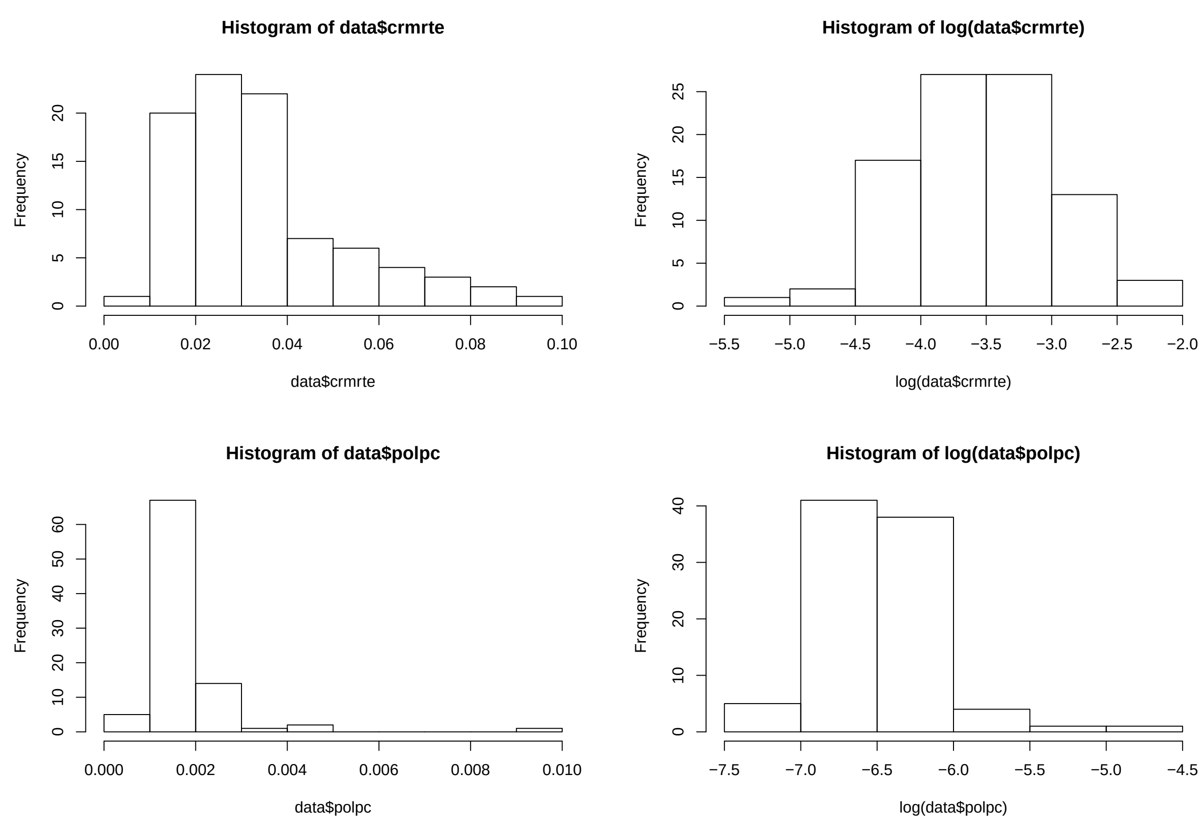
The belief was that by adopting a zero tolerance approach that enforced even the lowest level offenses, crime rates would subsequently go down. While New York City notably enforced this more stringent approach, San Francisco went the opposite direction of less strident law enforcement policies that reduced arrests, prosecutions and incarceration rates. Both sides experienced considerable declines in crime rates. Thus we hope to test the "broken windows theory" for the counties of South Carolina in 1987 and answer the question: Does the conservative approach of deterrence through arrests, incapacitation through imprisonment, harsh sentencing and higher police per capita lead to lower crime rates?

Model 1: only the explanatory variables of key interest

Based on the research question, our initial proposed model will include *crmrte* as the dependent variable and all variables related to stricter law enforcement policies: *prbarr*, *prbconv*, *prbpris*, *avgsen*, and *polpc* as independent variables. Assuming the "broken windows theory" is valid, we expect generally negative coefficients for all variables.

Given that the histogram of *crmrte* has a significant positive skew, we noted a log transformation may be suitable since its values are non-zero and positive. The same can be said about the independent variable *polpc* where its histogram is positively skewed and its values are non-zero and positive.

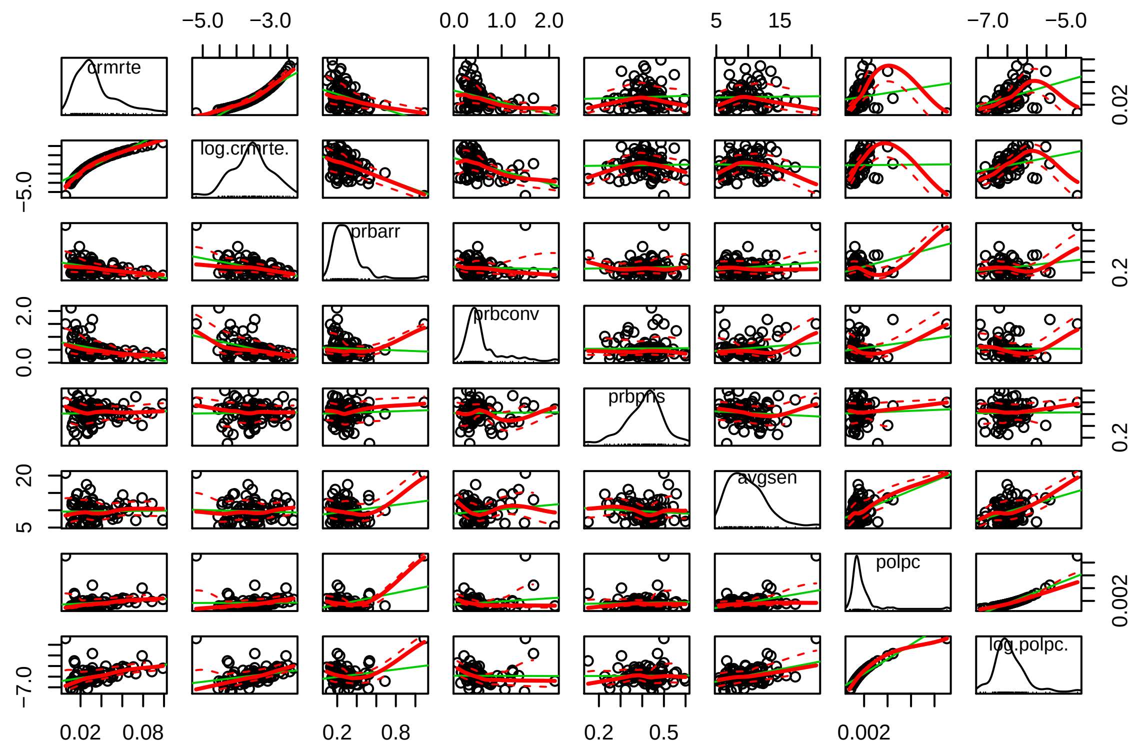
```
# before and after log transformation
hist(data$crmrte); hist(log(data$crmrte))
hist(data$polpc); hist(log(data$polpc))
```



Though *prbarr*, *prbconv*, and *prbpris* are positively skewed as well, we decided against taking the log of these variables because log transformations can make values between 0 and 1 more extreme. We also kept *avgsen* as is for easier interpretation.

Next, we want to check the relationships between the chosen independent variables and our dependent variable, before and after transformations. We want to ensure that we did not deviate any straight-line relationships between the independent variables and the dependent variable using the transformation.

```
scatterplotMatrix(~ crmrte + log(crmrte) + prbarr + prbconv + prbpris + avgsen + polpc + log(polpc), da
```



As we can see from the scatterplot matrix, it does not appear that the transformation drastically changed the relationship.

Hence, we propose our first model as follows which contains all explanatory variables of key interest:

$$log(crmrte) = \beta_0 + \beta_1 \cdot prbarr + \beta_2 \cdot prbconv + \beta_3 \cdot prbpris + \beta_4 \cdot avgsen + \beta_5 \cdot log(polpc) + u$$

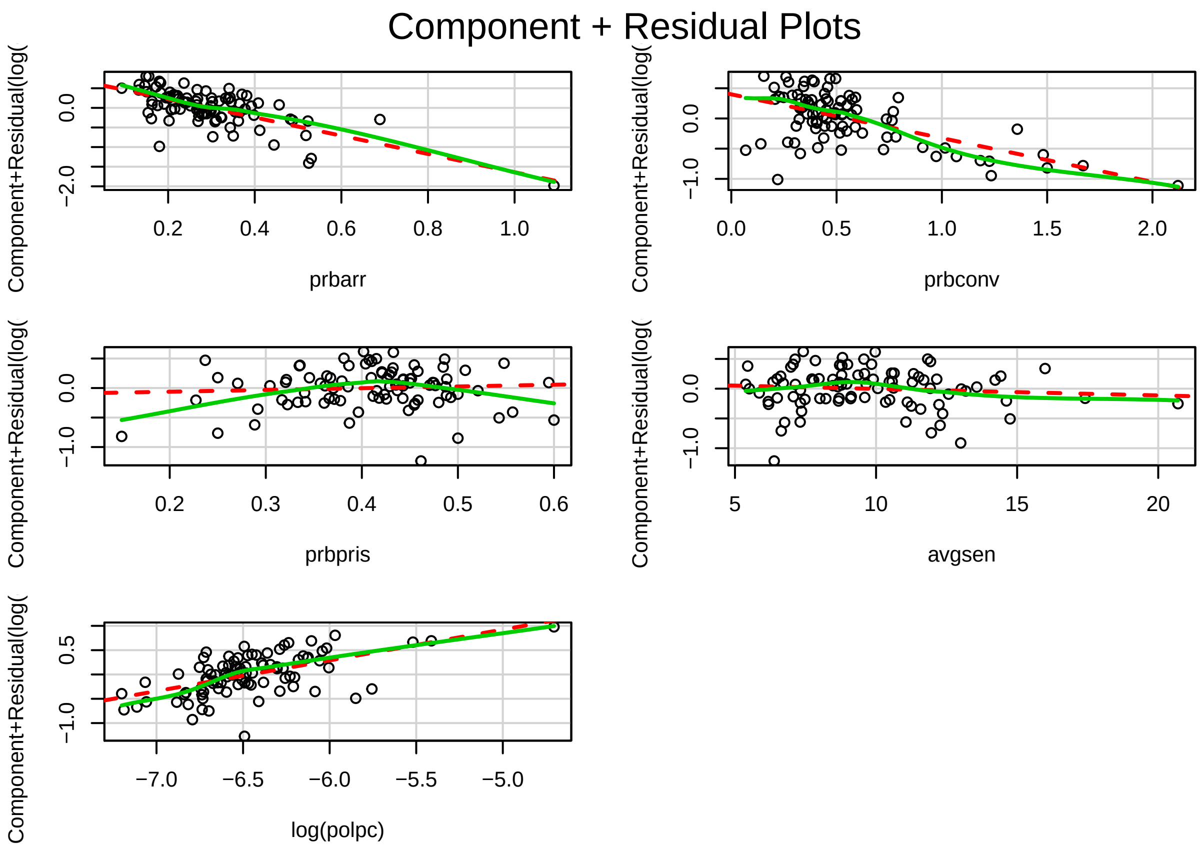
We will now run the model and test the validity of the 6 CLM assumptions:

```
m1 <- lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + log(polpc), data=data)
```

CLM 1 - A linear model

The model is specified such that the dependent variable is a linear function of the explanatory variables. As shown in the scatterplot matrix above, all of the dependent variables in the model seem to have a linear relationship with the independent variable $log(crmrte)$. We can verify further the linearity of the relationship using either component+residual plots (also called partial-residual plots) or the CERES plots. We have decided to do the former and note that for the most part, the relationships appear linear.

```r
# verify linearity of relationships using component+residual plots
crPlots(m1)
```



CLM 2 - Random Sampling

We do not know how the survey is collected. We assume that the variables are representative of the entire population distribution, but we cannot assess this assumption perfectly. Since there is a possibility that the individuals who collect the survey reach out to only one municipal police department instead of the county police department, the data collected this way are not representative of the county. There is nothing we can do to correct this, so we note this as a potential weakness in the analysis.

CLM 3 - Multicollinearity

As a quick test of the multicollinearity condition, we check the correlation of the explanatory variables and their Variance Inflation Factors (VIF):

```r
# correlation matrix of explanatory variables
data$log.polpc <- log(data$polpc)
cor(data.matrix(subset(data, select=c("prbarr", "prbconv", "prbpris", "avgsen", "polpc", "log.polpc"))))
```

```
##                prbarr      prbconv     prbpris      avgsen      polpc
## prbarr     1.00000000 -0.055796206  0.04583324  0.17869425 0.42596481
## prbconv   -0.05579621  1.000000000  0.01102265  0.15585232 0.17186516
```

```
## prbpris    0.04583324  0.011022645  1.00000000 -0.09468083 0.04820783
## avgsen     0.17869425  0.155852319 -0.09468083  1.00000000 0.48815230
## polpc      0.42596481  0.171865155  0.04820783  0.48815230 1.00000000
## log.polpc  0.21624362 -0.007574581  0.01041348  0.43729326 0.90577332
##              log.polpc
## prbarr     0.216243622
## prbconv   -0.007574581
## prbpris    0.010413481
## avgsen     0.437293258
## polpc      0.905773320
## log.polpc  1.000000000
```

```
# verify VIFs are less than 10
vif(m1)
```

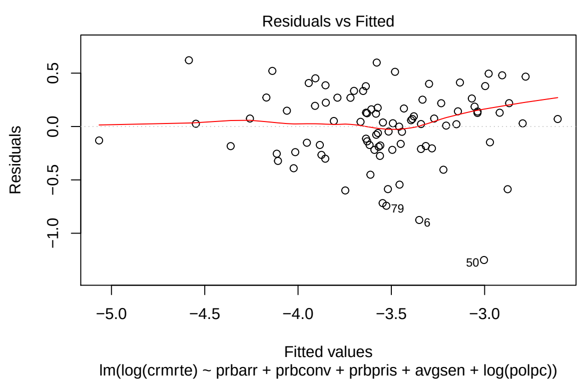
```
##     prbarr    prbconv    prbpris     avgsen log(polpc)
##   1.068228   1.039388   1.016889   1.310152   1.277425
```

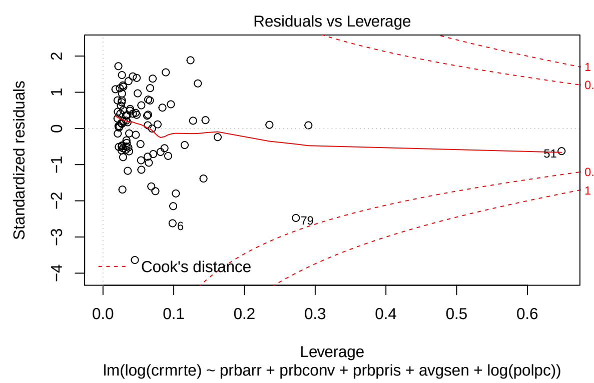
The explanatory variables (*prbarr*, *prbconv*, *prbpis*, *avgsen*, *log.polpc*) are not perfectly correlated and the VIFs are low (i.e. less than 10), so there is no perfect multicollinearity of the independent variables.

CLM 4 – Zero-Conditional Mean

To see whether there is a zero-conditional mean across all x's, we will plot the residuals against the fitted values.

```
# plot residual vs fitted plot & residual vs leverage plot
plot(m1, which=c(1, 5))
```





The residual vs fitted plot indicates little evidence that the zero-conditional mean assumption does not hold since the red spline line remains close to zero despite its slight dip and rise at both ends due to fewer observations.

Furthermore, it does not appear that the outliers have undue influence on the model fit. Based on the residual vs leverage plot, none of the outliers have a leverage that exceeds a Cook's distance of 1 on the regression model.

We have also taken a look at the covariances of the independent variables with the residuals to see if the variables we chose are likely to be exogenous.

```
# calculate the covariance for each independent variables with the model's residuals
lapply(subset(data, select=c("prbarr", "prbconv", "prbpris", "avgsen", "log.polpc")),
      function(var) cov(var, m1$residuals))
```

```
## $prbarr
## [1] -0.000000000000000001322951
##
## $prbconv
## [1] -0.0000000000000000006040772
##
## $prbpris
## [1] -0.0000000000000000009032893
##
## $avgsen
## [1] 0.0000000000000000129483
##
## $log.polpc
## [1] 0.00000000000000001204546
```

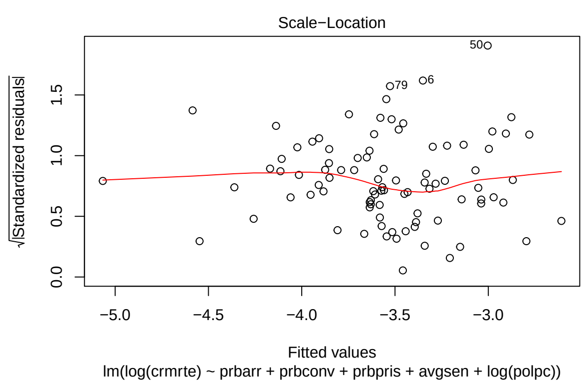
The covariances are very close to zero indicating the likelihood of being exogenous.

Because of the substantial sample size and the results of the verifications we have performed above, there is little evidence that the zero-conditional mean assumption is invalid.

CLM 5 - Homoscedasticity

To determine whether the variance of u is fixed for all x's, we look at the scale-location plot to see if residuals are spread equally along the ranges of the explanatory variables.

```r
# plot scale-location plot
plot(m1, which=3)
```



The residuals appear randomly spread; therefore we can assume that the variance is equal.

To further verify this assumption, we run Breusch-Pagan and the Score-test for non-constant error variance.

```r
# Breusch-pagan test
bptest(m1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 6.1759, df = 5, p-value = 0.2895
```

The Breusch-pagan test validates our assumption of homoskedasticity. Since the p-value is statistically not significant, we cannot reject the null hyothesis of homoskedasticity.

```r
# Score-test for non-constant error variance
ncvTest(m1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.496156    Df = 1     p = 0.2212638
```

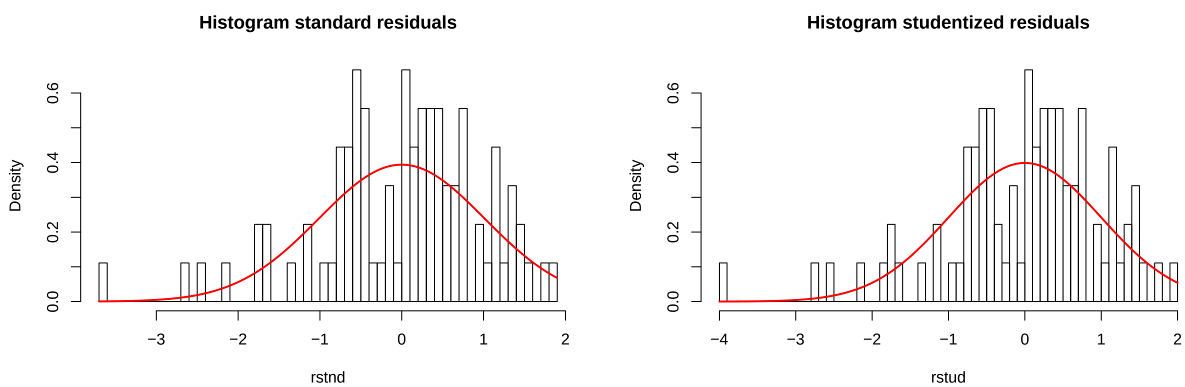
The Score-test also validates this assumption. Since the p-value is statistically not significant, we cannot reject the null hypothesis of constant error variance.

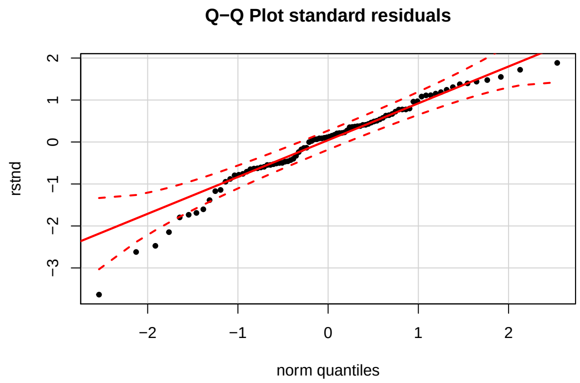
For this reason, the assumption of homoskedasticity is met.

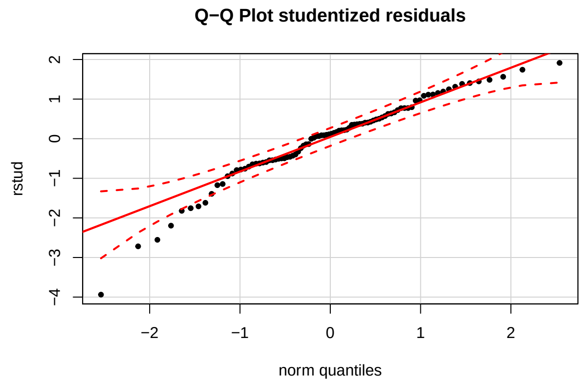
CLM 6 – Normality of residuals

To determine whether there is normality of the residuals, we looked at the histogram and the QQ-plot of the residuals and visually observe whether there is normality.

```r
# normality of standard residuals
rstnd = rstandard(m1)
hist(rstnd, main="Histogram standard residuals", breaks=50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(rstnd)), col="red", lwd=2, add=TRUE)
# normality of studentized residuals
rstud = rstudent(m1)
hist(rstud, main="Histogram studentized residuals", breaks=50, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
# Q-Q plot standard residuals
qqPlot(rstnd, distribution="norm", pch=20, main="Q-Q Plot standard residuals")
qqline(rstnd, col="red", lwd=2)
# Q-Q plot studentized residuals
qqPlot(rstud, distribution="norm", pch=20, main="Q-Q Plot studentized residuals")
qqline(rstud, col="red", lwd=2)
```







The histograms appear to be negatively skewed. The Q-Q plots further supports it with a fat negative tail.

```r
#check sample size for model 1
nobs(m1)
```

```
## [1] 90
```

Although the assumption is not met, given the substantial sample size, we can be confident that due to OLS asymptotics the distribution of the residuals will be approximately normal.

Conclusion:

TODO: These are notes for policy suggestions

Read more on http://lawenforcementleaders.org/wp-content/uploads/2017/10/Law-Enforcement-Letter-to-President-Trump-ar pdf

this effort is undercut by a diffuse focus. Attorney General Sessions' regular statements encouraging law enforcement to focus on drug and nonviolent offenders divert officers away from that vital mission. Law enforcement resources are limited. Focusing on low-level non-violent offenders means less time to stop and bring to justice the most dangerous offenders.

From our experience, we do not believe that always seeking the longest possible sentence will make our country safer. More than 25 percent of the Justice Department's budget is consumed by federal prisons.[iv] Every unnecessary dollar spent on prisons is a dollar not spent on policing. And often, the best way to prevent recidivism is through treatment, not prison. Responsibly reducing incarceration will free funding and time for our officers to focus on targeting and preventing violent crime, making our streets safer.[v]

References:

"Shattering"Broken Windows": An Analysis of San Francisco's Alternative Crime Policies", CENTER ON JUVENILE AND CRIMINAL JUSTICE, October 1999 http://www.cjcj.org/uploads/cjcj/documents/shattering.pdf

Jackman, Tom. "Nation's top cops, prosecutors urge Trump not to roll back successful crime policies." The Washington Post, WP Company, 18 Oct. 2017, www.washingtonpost.com/news/true-crime/wp/2017/10/18/nations-top-cops-prosecutors-urge-trump-not-to-roll-back-successful-crime-policies/?utm_term=.53fb295eac1e.