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 1988 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_v2_updated.csv")
# verify that it only contains data from 1988
unique(data$year)
```

```
## [1] 88
```

```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] 34 21  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] 5 1
```

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

```
##        X   county     year    crime  probarr  probsen probconv   avgsen
##        0        0        0        0        0        0        0        0
##   police  density      tax     west  central    urban   pctmin  wagecon
##        0        0        0        0        0        0        0        0
```

```
## wagetuc  wagetrd  wagefir  wageser  wagemfg  wagefed  wagesta  wageloc
##       0        0        0        0        0        0        0        0
##     mix    ymale
##       0        0
```

The dataset contains 90 counties from North Carolina, all of which is collected in 1988. Out of the 90 counties, 34 are from western NC (out of which 5 is also urban), 21 are from central NC (out of which 1 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        crime         
## Min.   : 1.00   Min.   :  1.0   Min.   :88   Min.   :0.005533  
## 1st Qu.:23.25   1st Qu.: 51.5   1st Qu.:88   1st Qu.:0.020604  
## Median :45.50   Median :103.0   Median :88   Median :0.030002  
## Mean   :45.50   Mean   :100.6   Mean   :88   Mean   :0.033510  
## 3rd Qu.:67.75   3rd Qu.:150.5   3rd Qu.:88   3rd Qu.:0.040249  
## Max.   :90.00   Max.   :197.0   Max.   :88   Max.   :0.098966  
##    probarr          probsen          probconv          avgsen      
## Min.   :0.1500   Min.   :0.09277   Min.   :0.06838   Min.   : 5.380  
## 1st Qu.:0.3642   1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.: 7.375  
## Median :0.4222   Median :0.27146   Median :0.45170   Median : 9.110  
## Mean   :0.4106   Mean   :0.29524   Mean   :0.55086   Mean   : 9.689  
## 3rd Qu.:0.4576   3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:11.465  
## Max.   :0.6000   Max.   :1.09091   Max.   :2.12121   Max.   :20.700  
##     police             density            tax              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   :0.3778  
## 3rd Qu.:0.0018856   3rd Qu.:1.5693   3rd Qu.: 41.01   3rd Qu.:1.0000  
## Max.   :0.0090543   Max.   :8.8277   Max.   :119.76   Max.   :1.0000  
##    central           urban             pctmin          wagecon     
## 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   :0.2333   Mean   :0.08889   Mean   :25.713   Mean   :285.4  
## 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0  
## Max.   :1.0000   Max.   :1.00000   Max.   :64.348   Max.   :436.8  
##    wagetuc         wagetrd         wagefir         wageser      
## 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  
##    wagemfg         wagefed         wagesta         wageloc     
## 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              ymale        
## 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 *probarr* and *probconv*, which have probability values greater than 1.

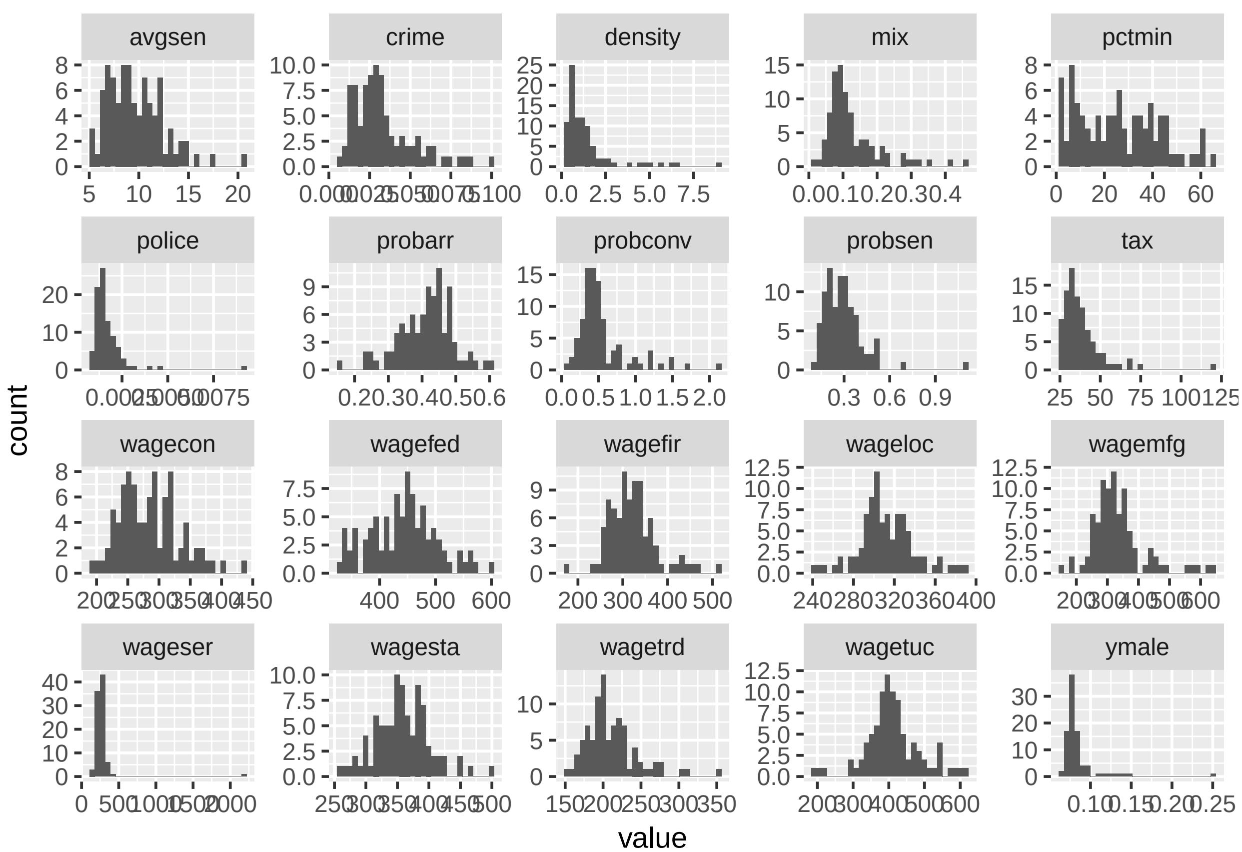
```r
# list number of probabilities (probarr, probconv, probsen, mix) that are not in range [0, 1]
c(sum(data$probarr < 0 | 1 < data$probarr), sum(data$probconv < 0 | 1 < data$probconv),
  sum(data$probsen < 0 | 1 < data$probsen), sum(data$mix < 0 | 1 < data$mix))
```

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

probconv and *probsen* contain 10 and 1 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)
```

```
##       crime     probarr     probsen    probconv      avgsen      police
##  1.28174888 -0.45254022  2.52529596  2.03950599  1.00116340  4.98348795
```

```
##     density         tax      pctmin     wagecon     wagetuc     wagetrd
##  2.65301071  3.29057447  0.36566169  0.60680223  0.06819768  1.46120657
##     wagefir     wageser     wagemfg     wagefed     wagesta     wageloc
##  0.82063145  8.69918165  1.42253166  0.13223761  0.36236826  0.29513808
##         mix       ymale
##  1.91657046  4.56069074
```

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)
```

```
##         crime       probarr       probsen      probconv        avgsen
##    0.09896590    0.15000001    1.09090996    2.12121010   20.70000076
##        police       density           tax        pctmin       wagecon
##    0.00905433    8.82765198  119.76145170   64.34819794  436.76663210
##       wagetuc       wagetrd       wagefir       wageser       wagemfg
##  187.61726380  354.67611690  509.46551510 2177.06811500  646.84997560
##       wagefed       wagesta       wageloc           mix         ymale
##  597.95001220  499.58999630  388.08999630    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$wageser)
```

```
##    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. . . .

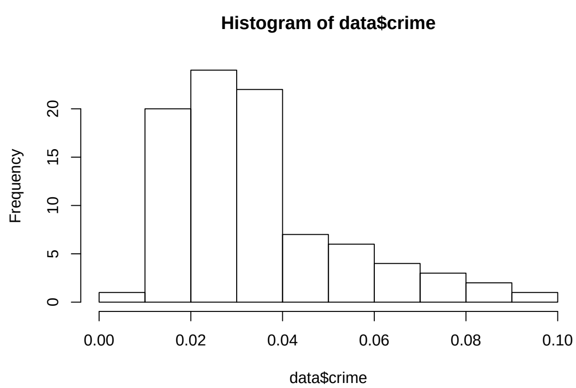
The belief was that by adopting a zero tolerance approach that enforced even the lowest level offenses, crime rates would subsequently go down. While New York City notably enforced this more stringent approach, San Francisco went the opposite direction of less strident law enforcement policies that reduced 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?

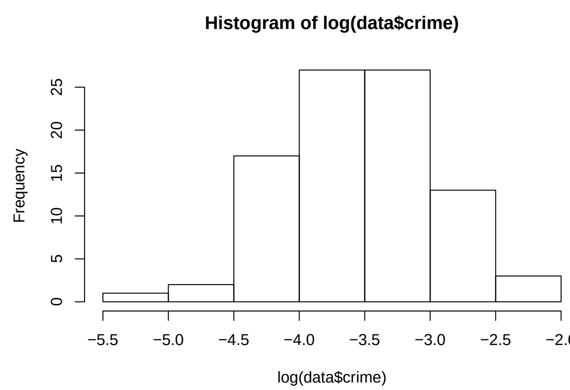
Model 1: only the explanatory variables of key interest

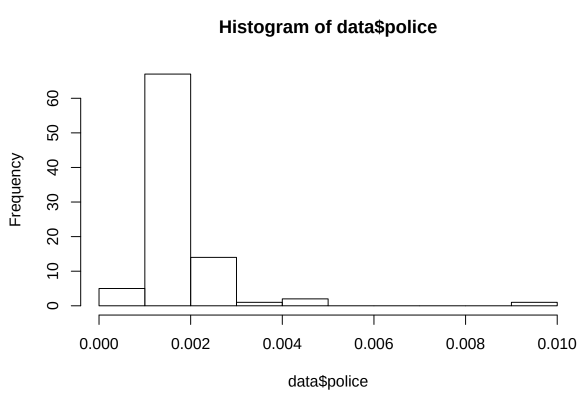
Based on the research question, our initial proposed model will include *crime* as the dependent variable and all variables related to stricter law enforcement policies: *probarr*, *probconv*, *probsen*, *avgsen*, and *police* as independent variables. Assuming the "broken windows theory" is valid, we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

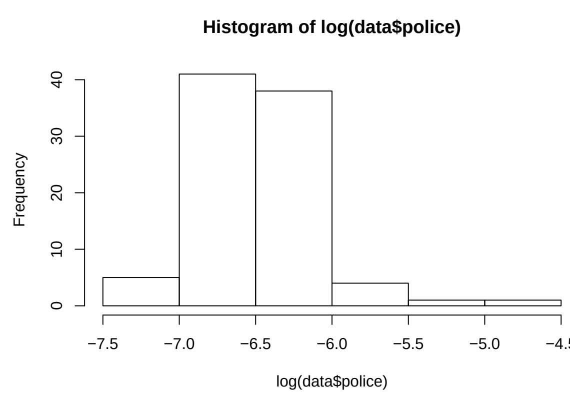
Given that the histogram of *crime* has a significant positive skew, we noted a log transformation may be suitable since its values are non-zero and positive. The same can be said about the independent variable *police* where its histogram is positively skewed and its values are non-zero and positive.

```
# before and after log transformation
hist(data$crime); hist(log(data$crime))
hist(data$police); hist(log(data$police))
```





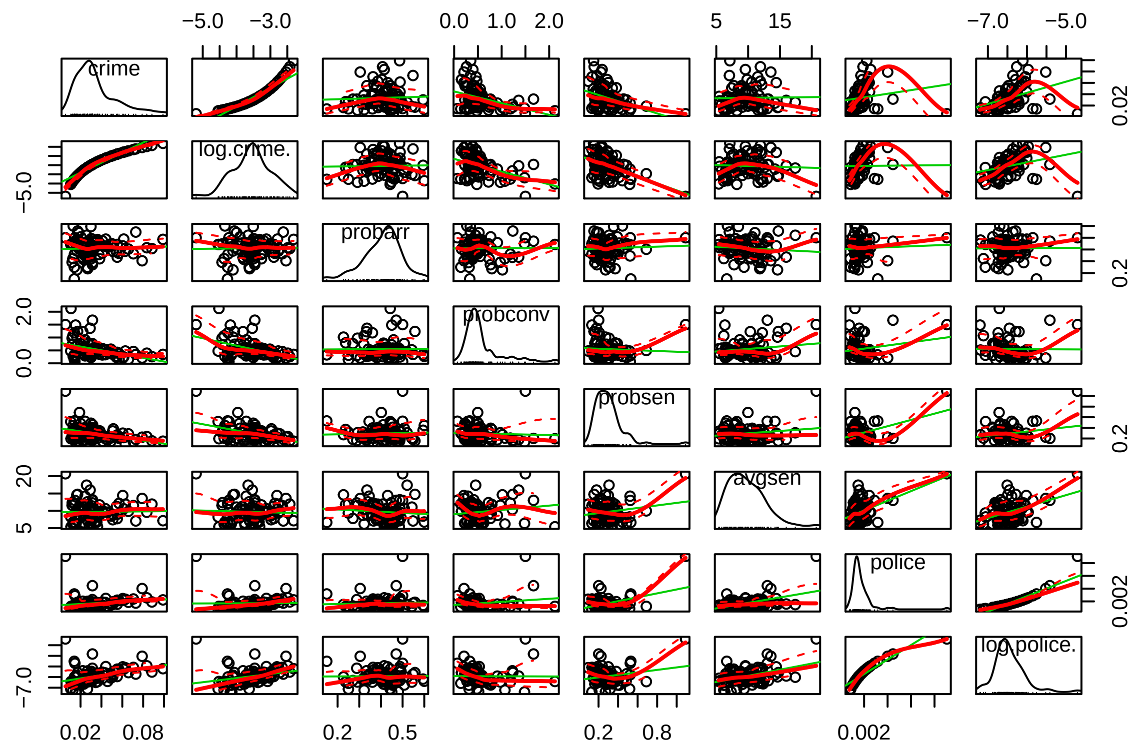




Though *probarr*, *probconv*, and *probsen* are positively skewed as well, we decided against taking the log of these variables because log transformations can make values between 0 and 1 more extreme. We also kept *avgsen* as is for easier interpretation.

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(~ crime + log(crime) +
                    probarr + probconv + probsen + avgsen + police + log(police), data = data)
```



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

Lastly, based on the exploratory data analysis, we should be careful when considering *probconv* and *probsen* as variables in the model with 10 and 1 datapoints respectively that have probabilties greater than 1. *probconv* is proxied by the ratio of convictions to arrest while *probsen* is proxied by the proportion of total convictions resulting in prison sentences. Although it is unlikely that an individual can be convicted without an arrest or sentenced without a conviction, we cannot rule out the possibility. Both of these variables are important in answering our research question and removing them will result in an omitted variable bias as we will demonstrate below.

Assuming we started out with a base model without *probconv* and *probsen*, we wanted to see what effects *probconv* and *probsen* respectively have on the other explanatory variables when we add them individually to the base model. We looked at the printout of their respective model coefficients to understand the effects. Based on the research question, we expect that higher conviction and higher sentencing will result in lower crime rate. And since the relationship of *probconv* and *probsen* are positive with the other explanatory variables as demonstrated by the correlation matrix, we expect negative bias overall.

```r
# demonstrate that probconv and probsen individually have positive relationship
# with the other explanatory variables: probarr, avgsen, police
ind.vars <- subset(data, select= c("probarr", "probconv", "probsen", "avgsen", "police"))
cor(ind.vars, ind.vars)
```

```
##              probarr    probconv     probsen      avgsen     police
## probarr   1.00000000  0.01102265  0.04583324 -0.09468083 0.04820783
## probconv  0.01102265  1.00000000 -0.05579621  0.15585232 0.17186514
## probsen   0.04583324 -0.05579621  1.00000000  0.17869425 0.42596480
## avgsen   -0.09468083  0.15585232  0.17869425  1.00000000 0.48815230
## police    0.04820783  0.17186514  0.42596480  0.48815230 1.00000000
```

```r
# test omitted variable bias by first creating a base model and a model for each omitted variable
m1.base <- lm(crime ~ probarr + avgsen + police, data=data)
m1.probconv <- lm(crime ~ probarr + probconv + avgsen + police, data=data)
m1.probsen <- lm(crime ~ probarr + probsen + avgsen + police, data=data)
# print out the model coefficients
(coef.base <- coeftest(m1.base, vcov=vcovHC))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  0.02885462  0.02976205  0.9695   0.3350
## probarr      0.00725067  0.03104952  0.2335   0.8159
## avgsen      -0.00050917  0.00082262 -0.6190   0.5376
## police       3.87085953 11.68258001  0.3313   0.7412
```

```r
(coef.probconv <- coeftest(m1.probconv, vcov=vcovHC))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  0.03638722  0.02803944  1.2977 0.197896
## probarr      0.00866278  0.03144264  0.2755 0.783593
## probconv    -0.02265313  0.00764791 -2.9620 0.003962 **
## avgsen      -0.00023551  0.00082518 -0.2854 0.776031
## police       4.87482880  9.63460820  0.5060 0.614187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(coef.probsen <- coeftest(m1.probsen, vcov=vcovHC))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  0.04399311  0.02316845  1.8988 0.060978 .
## probarr      0.01010303  0.02266319  0.4458 0.656882
## probsen     -0.07884692  0.02817947 -2.7980 0.006359 **
## avgsen      -0.00064360  0.00074499 -0.8639 0.390070
## police       8.71360915  6.17379949  1.4114 0.161781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the coefficients, it does appear that higher conviction (*probconv*) and higher sentencing (*probsen*) result in lower crime rate (*crime*) as seen by their negative sign in their respective coefficient. We also note that *probconv* and *probsen* are statistically significant when added to the base model. It appears that there is a negative omitted variable bias. For this reason, it would be best to include *probconv* and *probsen* in our Model 1 proposal.

As we will later discuss in section "Discussion of Causality", if the outliers happen to be a measurement error, it will result in our model being confounded by bias. Although that might be the case, there is also a likelihood that the measurement is valid, and we have demonstrated that not including *probconv* and *probsen* will most likely confound our model with omitted variable bias.

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

$$log(crime) = \beta_0 + \beta_1 \cdot probarr + \beta_2 \cdot probconv + \beta_3 \cdot probsen + \beta_4 \cdot avgsen + \beta_5 \cdot log(police) + u$$

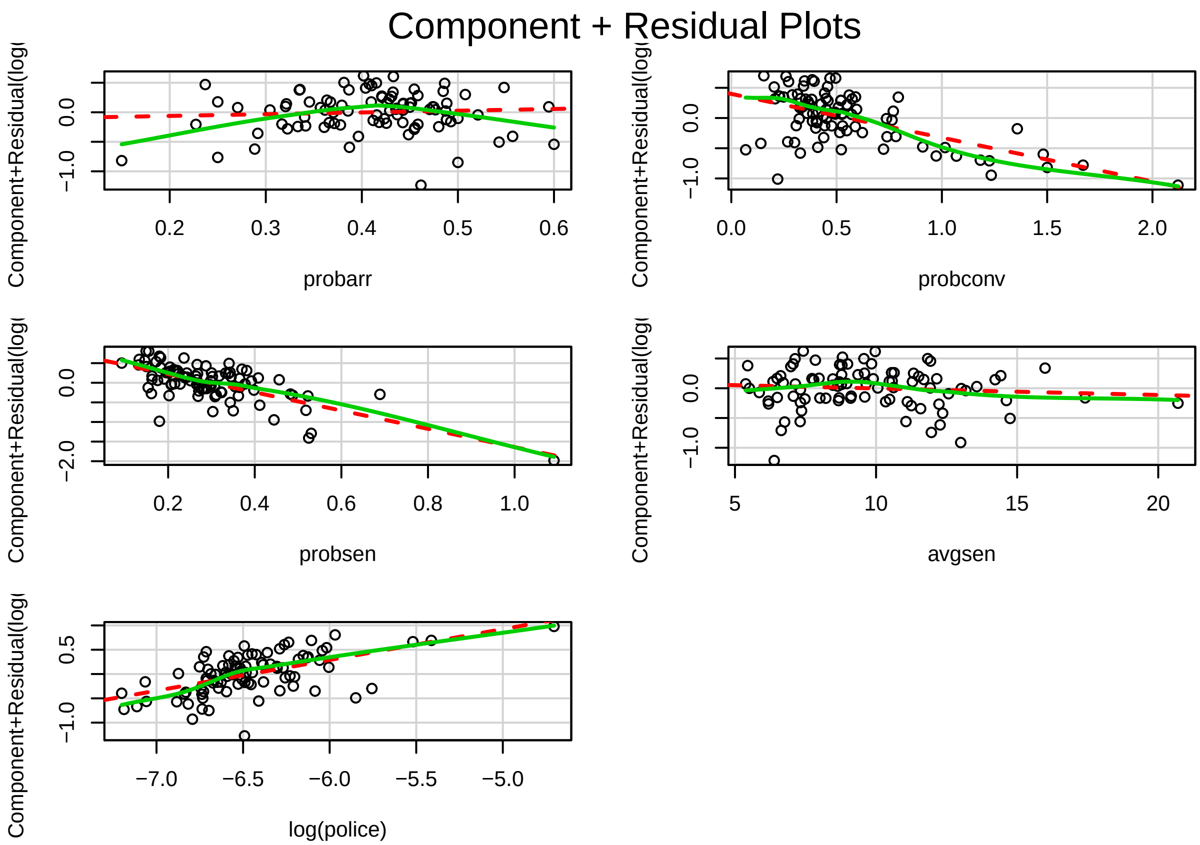
We will now run the model and test the validity of the 6 CLM assumptions to ensure that the OLS estimators are consistent, normally distributed, and BLUE (best linear unbiased estimator).

```
m1 <- lm(log(crime) ~ probarr + probconv + probsen + avgsen + log(police), 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 $log(crime)$. 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.

```
# 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 since the counties are subsets of North Carolina. 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):

```
# correlation matrix of explanatory variables
data$log.police <- log(data$police)
cor(data.matrix(
  subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "police", "log.police"))))
```

```
##                probarr     probconv     probsen      avgsen     police
## probarr     1.00000000  0.011022645  0.04583324 -0.09468083 0.04820783
## probconv    0.01102265  1.000000000 -0.05579621  0.15585232 0.17186514
## probsen     0.04583324 -0.055796206  1.00000000  0.17869425 0.42596480
## avgsen     -0.09468083  0.155852319  0.17869425  1.00000000 0.48815230
## police      0.04820783  0.171865142  0.42596480  0.48815230 1.00000000
## log.police  0.01041349 -0.007574593  0.21624362  0.43729326 0.90577332
##              log.police
## probarr     0.010413494
```

```
## probconv   -0.007574593
## probsen     0.216243619
## avgsen      0.437293263
## police      0.905773321
## log.police  1.000000000
```

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

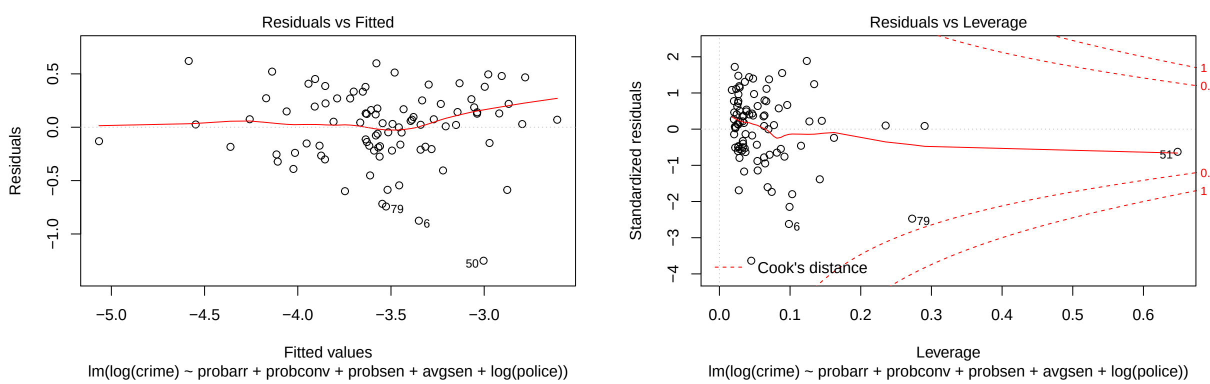
```
##     probarr    probconv     probsen      avgsen log(police)
##    1.016889    1.039388    1.068228    1.310152    1.277425
```

The explanatory variables (*probarr*, *probconv*, *prbpis*, *avgsen*, *log.police*) 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.

```r
# 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.

```r
# calculate the covariance for each independent variables with the model's residuals
lapply(subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "log.police")),
      function(var) cov(var, m1$residuals))
```

```
## $probarr
## [1] 0.000000000000000000008087166
##
## $probconv
## [1] 0.00000000000000000003500816
```

```
##
## $probsen
## [1] 0.0000000000000000001024187
##
## $avgsen
## [1] 0.0000000000000000008810665
##
## $log.police
## [1] -0.0000000000000000003502262
```

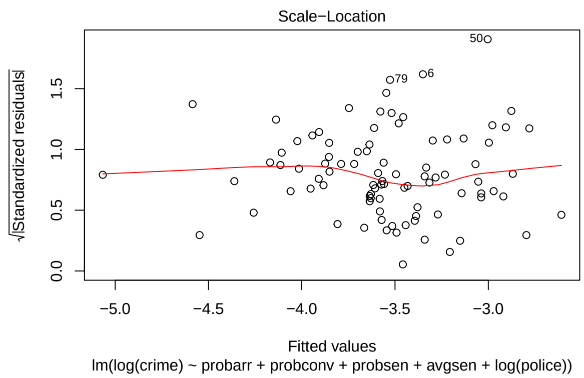
The covariances are very close to zero indicating the likelihood of being exogenous.

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.

```
# 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.

```
# 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.

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

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
```

```
## Chisquare = 1.496155    Df = 1     p = 0.2212639
```

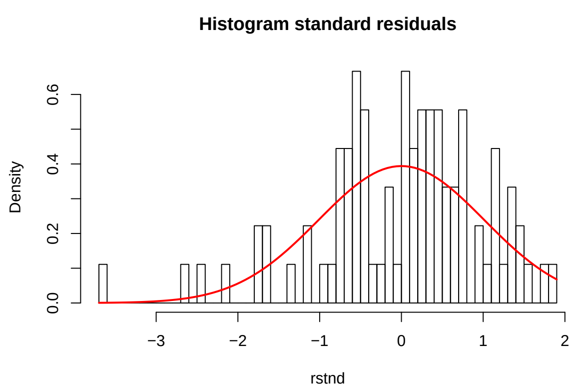
The Score-test also validates this assumption. Since the p-value is statistically not significant, we cannot reject the null hypothesis of constant error variance.

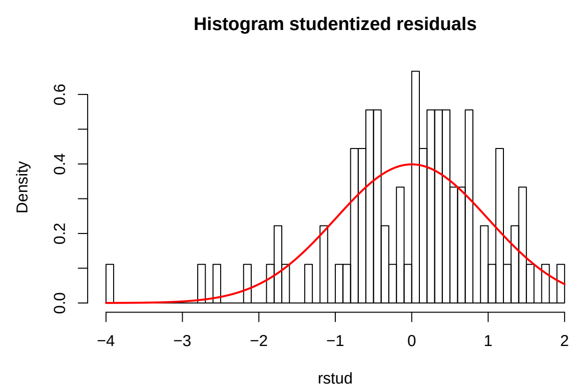
For this reason, the assumption of homoskedasticity is met.

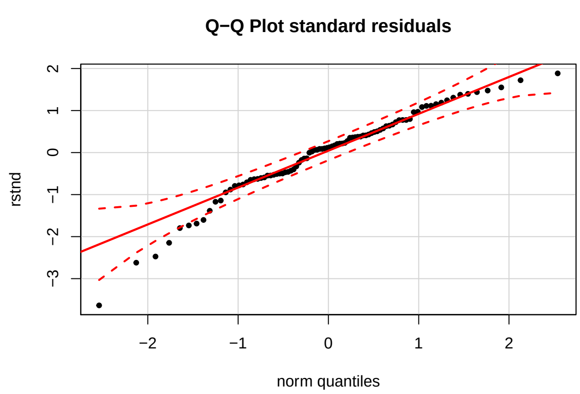
CLM 6 – Normality of residuals

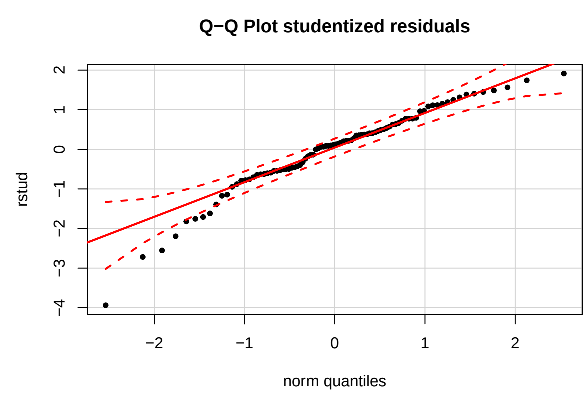
To determine whether there is normality of the residuals, we looked at the histogram and the Q-Q plot of the residuals and visually observe whether there is normality.

```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.

Since all six assumptions of the Classical Linear Model are met, we can assume that the OLS estimators are consistent, normally distributed and BLUE.

Model 2: add covariates that increase accuracy without bias

For Model 2, we decided to include variables that have an indirect impact to crime rate: *density*, *tax*, and *mix*.

We chose *density* based on the theory that the more densely populated an area is, the harder it is for individuals to commit crime, which in turn decreases the crime rate. We assumed that if we included *density* on top of Model 1, its coefficient will be positive since it has a negative relationship with *crime* and a negative relationship with most of the explanatory variables as shown below. Our thought process goes as follows: because we assume that densely populated area will lower crime rate, it will lower the probability of arrest, conviction and prison sentence and therefore have a negative relationship with them. On the other hand, we assume that an increased number of people per capita will reflect an increased number of police per capita and therefore *density* will have a positive relationship with *police*.

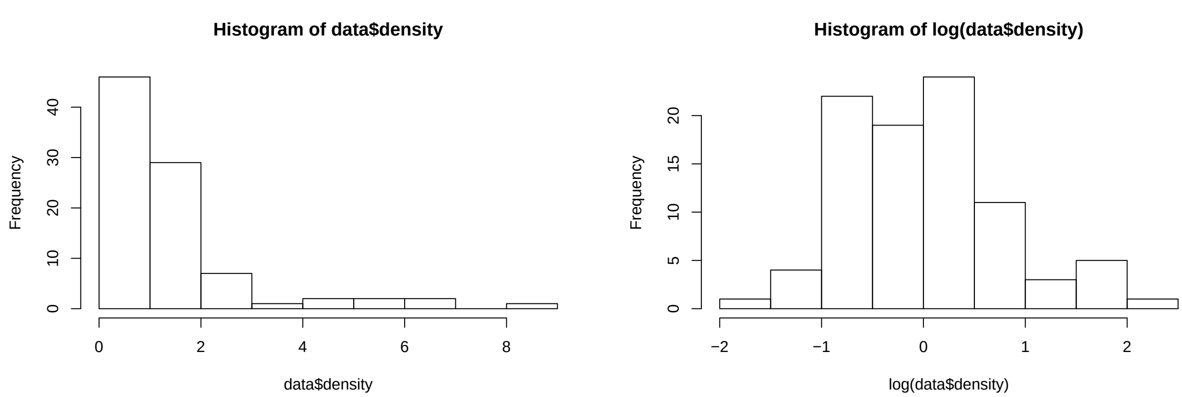
```r
# list the correlation between density and model 1's explanatory variables
cor(subset(data, select="density"),
    subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "log.police")))
```

```
##            probarr  probconv    probsen    avgsen log.police
## density 0.07260985 -0.227912 -0.3005332 0.0715956  0.3282668
```

The correlation matrix confirms our assumptions.

Since *density* is positively skewed, it will benefit with a log transformation because its values are non-zero and positive.

```r
# before and after log transformation
hist(data$density); hist(log(data$density))
```



Although *density* has a direct relationship with *urban*, we decided not to include *urban* because there are other factors, such as wage discrepancy between the wealthy and the poor that can potentially exist in urban counties which our current dataset does not have. In other words, by including *urban*, we might fall prey to the omitted variable bias because there are potentially multiple variables that influence *urban* which are not available in our current dataset.

We chose tax revenue per capita, *tax*, on a similar basis as *density* in that higher tax revenue usually equates to more funding for protection services and therefore lowers the rate of crime. *tax* also has a negative relationship with *crime* and a negative relationship with most of the explanatory variables in Model 1 except for *police*, which we assume will reflect in its coefficients being positive. Again, our reasonings are similar to *density* in that the more money a county has to pay for protection services, such as police, the less likely an individual will commit crime and therefore the lower the probability of arrest, conviction and sentencing is in that county.

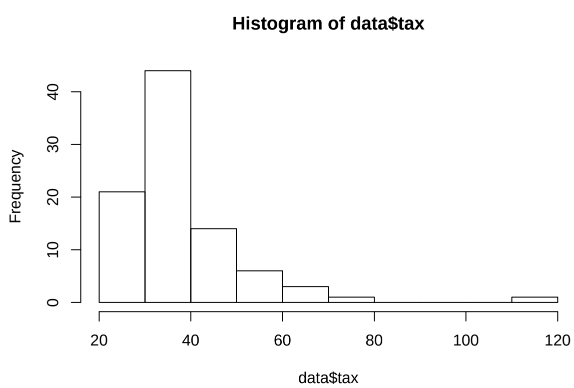
```
# list the correlation between tax and model 1's explanatory variables
cor(subset(data, select="tax"),
    subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "log.police")))
```

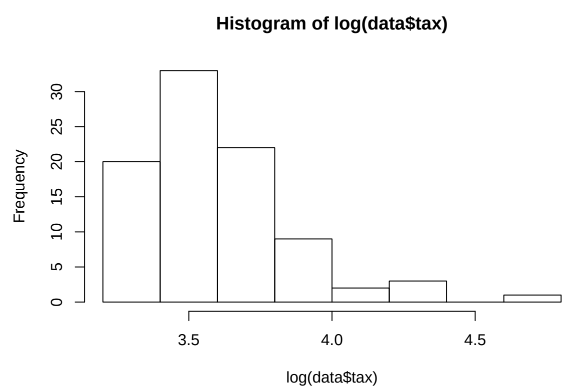
```
##         probarr   probconv   probsen     avgsen log.police
## tax -0.09236051 -0.1273896 -0.137191 0.08654323  0.4009476
```

The correlatin matrix again confirms our assumptions.

tax can also benefit with a log transformation because its distribution is skewed and its values are non-zero and positive.

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





Lastly, we chose *mix* as an indirect effect to *crime* because we are under the impression that the *mix* variable which reflects the ratio of face-to-face crime over all other crimes is a good indicator of violent crimes, and violent crimes have a direct impact to the average sentence a criminal will receive. We understand that it most likely has an effect on *probsen* and *avgsen* but we do not know what type of relationship it has. For this reason, we list the correlation between *mix* and the explanatory variables of Model 1 to get a better understanding.

```
# list the correlation between mix and model 1's explanatory variables
cor(subset(data, select="mix"),
    subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "log.police")))
```

```
##       probarr   probconv  probsen    avgsen log.police
## mix 0.1165888 -0.3042512 0.412898 -0.141705 0.03922583
```

Although *mix* is positively skewed as well, we decided against taking the log transformation of *mix* because it can makes its values between 0 and 1 more extreme.

Before we propose our second model, we would like to discuss the reasons why we decided not to include the other variables as covariates.

Neighborhood plays a central role in fostering the tendency of a person commiting a crime. Although we are given geographic locations such as *west* and *central* and neither *west* nor *central*, it does not inform us whether those counties are considered safe or unsafe. Knowing the unsafe rating of a county can give us a better picture of crime rates in those neighborhoods. Also, just having geographic locations do not inform us about laws enacted for safety in those particular regions. We will never know, for example, if counties in western region enact stricter laws than those in the central region. For this reason, we did not consider *west* and *central* as variables in Model 2.

In addition, including *ymale* and *pctmin* will introduce omitted variable bias in Model 2 because typically a county that has low education, high percentage of young male, and high percentage of minority will induce a high crime rate. Education plays a critical role when taking into consideration *ymale* and *pctmin*, and including *ymale* and *pctmin* without the education variable will open up to bias in the model.

Although wage is a good determinant of crime because those who are poor have a higher propensity to commit crime out of financial needs, the groupings of the wage variables do not provide us insight as to the different financial groups between the poor, the middle class, and the wealthy. For example, service industry is mostly thought of as jobs with high number of minimum wage workers, but as the outlier points out in *wageser*, it might be possible for someone to work in the service industry and earn a well-off paycheck if they worked, for example, in a five star hotel. Transportation industry can also be considered as mostly jobs with high number of minimum wage workers, but again, a pilot works in the transportation industry and gets paid well. It is for this reason that we did not include the wage variables in the model.

Hence, we can propose our second model as follows with all the indirect variables included:

$$\begin{aligned} log(crime) =& \beta_0 + \beta_1 \cdot probarr + \beta_2 \cdot probconv + \beta_3 \cdot probsen + \beta_4 \cdot avgsen + \\ & \beta_5 \cdot log(police) + \beta_6 \cdot log(density) + \beta_7 \cdot log(tax) + \beta_8 \cdot mix + u \end{aligned}$$

```
m2 <- lm(log(crime) ~ probarr + probconv + probsen + avgsen + log(police) +
         log(density) + log(tax) + mix, data=data)
```

CLM 1 - 6

1. The model is specified such that the dependent variable is a linear function of the explanatory variables.
2. As discussed in Model 1, we do not know how the survey is collected, but we assume that the variables are representative of the entire population distribution.
3. The variables are not perfectly correlated and the VIFs are low, so there is no perfect multicollinearity of the independent variables.

```
# verify VIFs are less than 10
vif(m2)
```

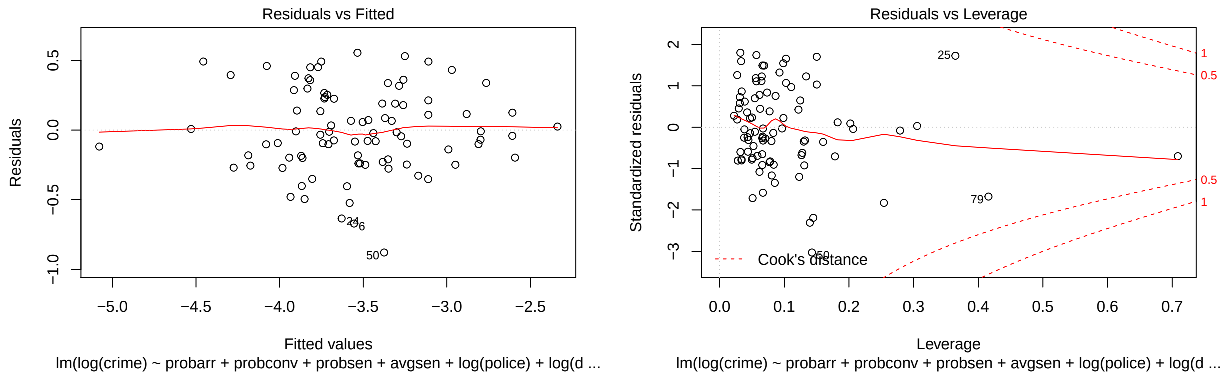
```
##      probarr     probconv      probsen       avgsen  log(police)
##     1.060959     1.359305     1.643384     1.365793     1.837204
## log(density)     log(tax)          mix
##     1.761691     1.292581     1.548026
```

4. Zero-conditional mean assumption holds because the spline line remains close to zero in the residual vs fitted plot, there is no outliers that have high influence, and the covariances are very close to zero indicating the likelihood of being exogenous.

```
data$log.density <- log(data$density)
data$log.tax <- log(data$tax)
# plot residual vs fitted plot & residual vs leverage plot
plot(m2, which=c(1, 5))
# calculate the covariance for each independent variables with the model's residuals
lapply(subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "log.police",
```

```r
                              "log.density", "log.tax", "mix")),
      function(var) cov(var, m2$residuals))
```

```
## $probarr
## [1] -0.000000000000000001030201
##
## $probconv
## [1] -0.000000000000000001414945
##
## $probsen
## [1] -0.000000000000000001206898
##
## $avgsen
## [1] -0.000000000000000008793001
##
## $log.police
## [1] 0.00000000000000001740289
##
## $log.density
## [1] 0.00000000000000001162179
##
## $log.tax
## [1] -0.000000000000000004366502
##
## $mix
## [1] 0.000000000000000001505435
```



5. The assumption of homoskedasticity is met because even though the residuals appear to spread in a downward curvature in the scale-location plot and the p-value is statistically significant based on the Breusch-pagan test, the error variance is constant because the p-value is not statistically significant in Score-test. Also, we can use the heteroscedasticity-robust standard errors for the hypothesis tests on the slope parameters.

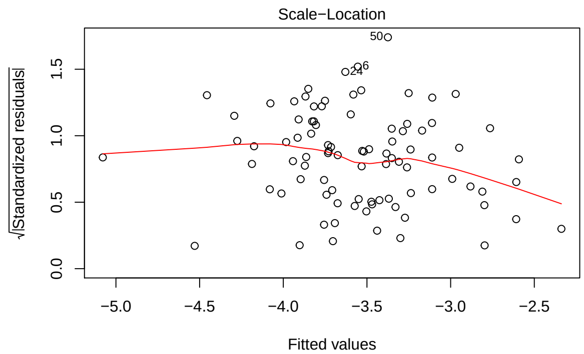
```r
# plot scale-location plot
plot(m2, which=3)
# Breusch-pagan test
bptest(m2)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  m2
## BP = 24.455, df = 8, p-value = 0.001922
```

```r
# Score-test for non-constant error variance
ncvTest(m2)
```

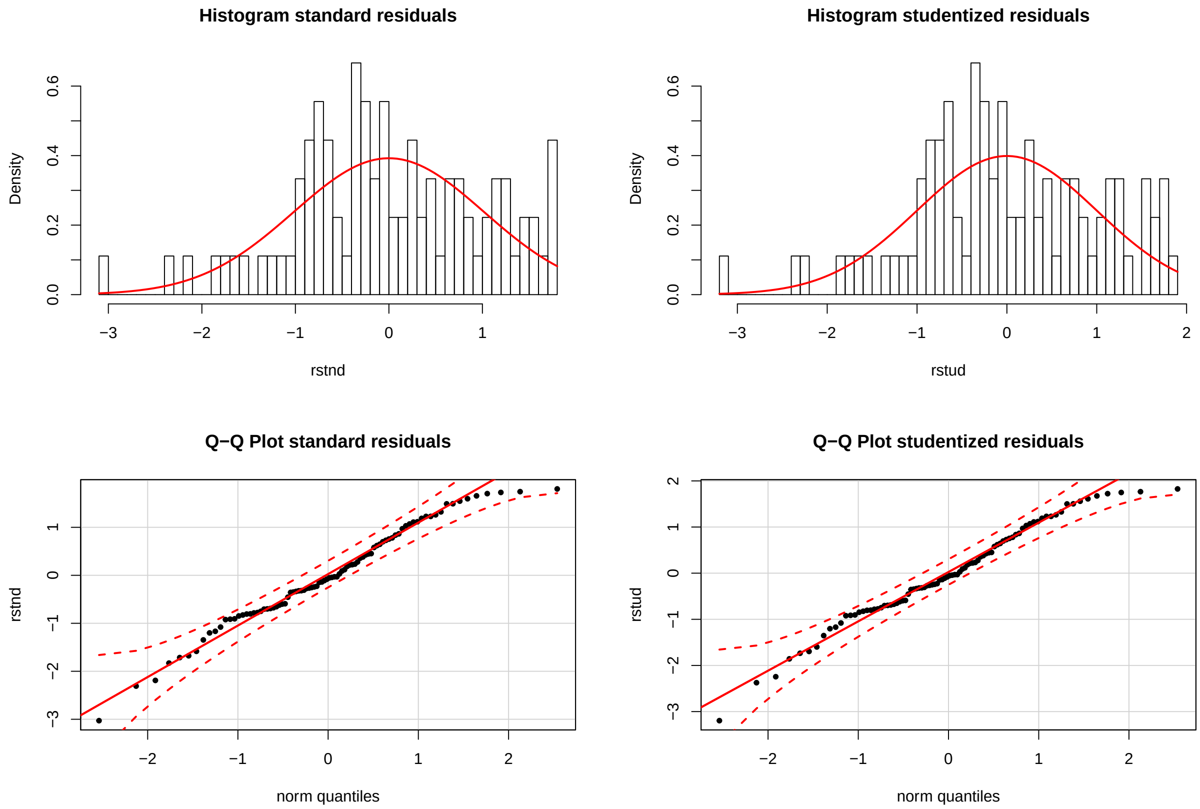
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.7415662    Df = 1     p = 0.3891596
```



6. 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.

```r
# normality of standard residuals
rstnd = rstandard(m2)
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(m2)
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)
# check sample size for model 2
nobs(m2)
```

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



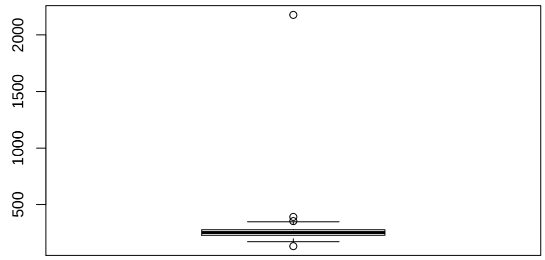
Since all six assumptions of the Classical Linear Model are met, we can assume that the OLS estimators areconsistent, normally distributed and BLUE.

Model 3: most, if not all, other covariates

In Model 3, we are going to include more variables from the dataset to try to control effects the other variables have on our dependent variable. Although this model might introduce more noises as it becomes over-specified, it will be able to explain more of the variances in the dependent variable than the previous models.

As discussed in the EDA, every variable seems to be in an expected range except for the outlier $2177.10 in *wageser*.

```
# Look into wageser with an unusual max of 2177.1
boxplot(data$wageser)
```



$2177.10 is clearly an outlier in the data and could be possibly due to a measurement error. In this case, we decide not to include this *wageser* to avoid confounding bias caused by potentially inaccurate data.

Moreover, let's look at how *west* and *central* distribute in the sample

```
unique(cbind(data$west, data$central))
```

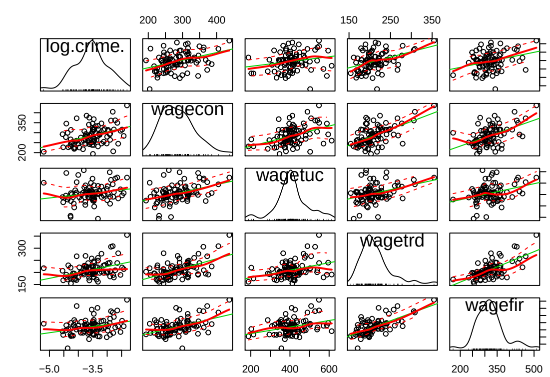
```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
## [3,]    0    0
```

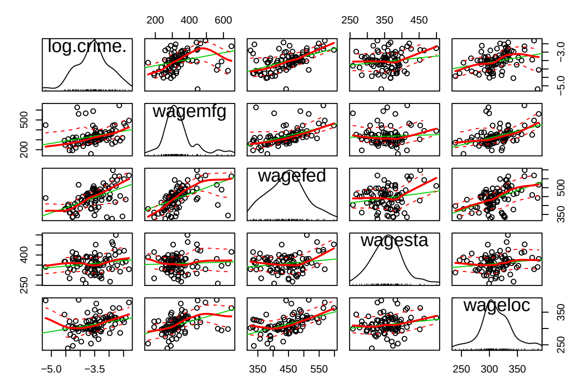
Note that although we don't have any counties that are both in west and central, as expected, we see some counties that are neither in west or central. In order to consider the effect of different regions, we will need to use both indicator variables in our model.

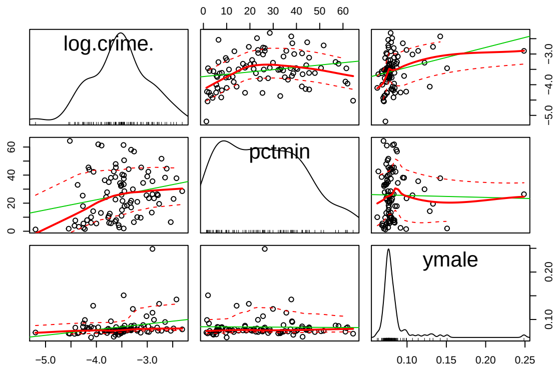
Out of all the variables that aren't included in Model 1 or Model 2, we have decided to include: 1) *west* & *central*, as indicator variables, to control for the regional effect on crime rate 2) *urban*, as an indicator variable, to control for the non-density impact of urbanization on crime crate 3) *wagecon*, *wagetuc*, *wagetrd*, *wagefir*, *wageser*, *wagemfg*, *wagefed*, *wagesta*, *wageloc* to control for the effects of wages in different industries have on our crime rate 4) *pctmin* and *ymale* to control for the demographic effect on crime rate

Now let's look the relationship between the selected variables and our dependent variable $log(crime)$

```
scatterplotMatrix(~ log(crime) + wagecon + wagetuc + wagetrd + wagefir, data = data)
scatterplotMatrix(~ log(crime) + wagemfg + wagefed + wagesta + wageloc, data = data)
scatterplotMatrix(~ log(crime) + pctmin + ymale, data = data)
```







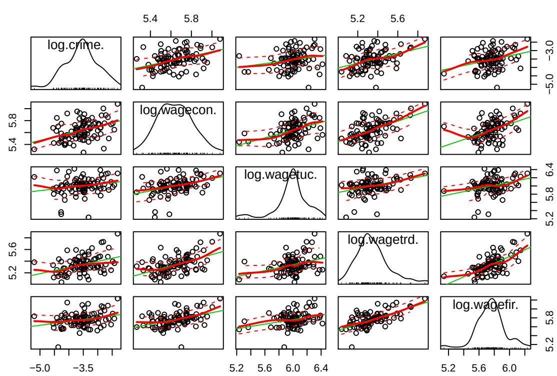
None of the variables shows strong evidence of non-linear relationship with the dependent variable $log(crmrte)$.

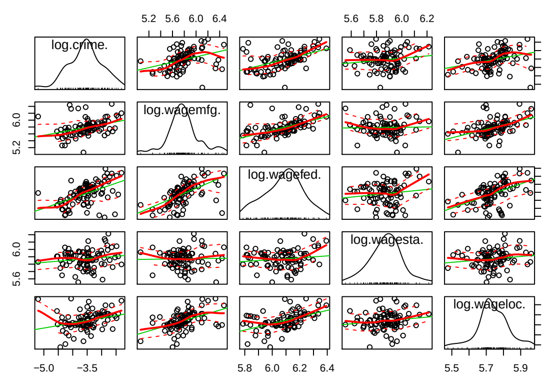
We notice that the distributions for the variables $wagecon$, $wagetrd$, $wagefir$, $wagemfg$, $wagesta$, and $wageloc$ are positively skewed, since they are all postive values, we can apply log transformations to all of them. For the ease of the intepretation of the model, we apply log tranformation to the other wage related variable $wagetuc$ as well.

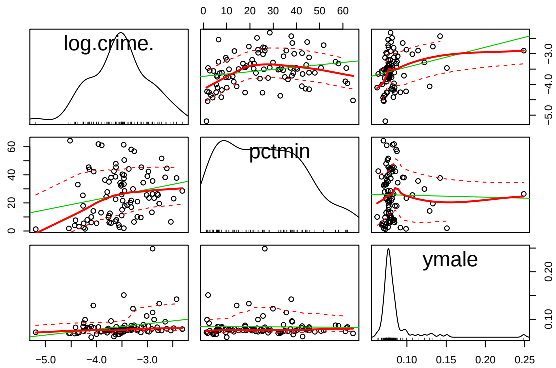
Although $ymale$ is also positively skewed and could benefit from a log transformation in terms of normality, it's hard to interpret the slope parameter of its log transformation. Hence, we decided to leave it as it is.

Let's double check the linearity of the relationship between our transformed variables and our dependent variable $log(crime)$

```
scatterplotMatrix(~ log(crime) + log(wagecon) + log(wagetuc) + log(wagetrd) + log(wagefir), data = data
scatterplotMatrix(~ log(crime) + log(wagemfg) + log(wagefed) + log(wagesta) + log(wageloc), data = data
scatterplotMatrix(~ log(crime) + pctmin + ymale, data = data)
```







We didn't see any strong violation agaisnt the linearity of the relationships. Hence, we will propose the following model:

```r
m3 <- lm(log(crime) ~ probarr + probconv + probsen + avgsen + log(police) +
         log(density) + log(tax) + mix +
         west + central + urban + ymale + pctmin +
         log(wagecon) + log(wagetuc) + log(wagetrd) + log(wagefir) +
         log(wagemfg) + log(wagefed) + log(wagesta) + log(wageloc), data=data)
```

CLM 1 - 6

1. The model is specified such that the dependent variable is a linear function of the explanatory variables.
2. As discussed in Model 1, we do not know how the survey is collected, but we assume that the variables are representative of the entire population distribution.
3. One typical concern when including variables that tend to be highly correlated, like wages, is that the multi-collinearity will inflate the variance of the OLS estimated parameters. Fortunately, we have determined that the variables are not perfectly correlated and the VIFs are low, so there is no perfect multicollinearity of the independent variables.

```r
# verify VIFs are less than 10
vif(m3)
```

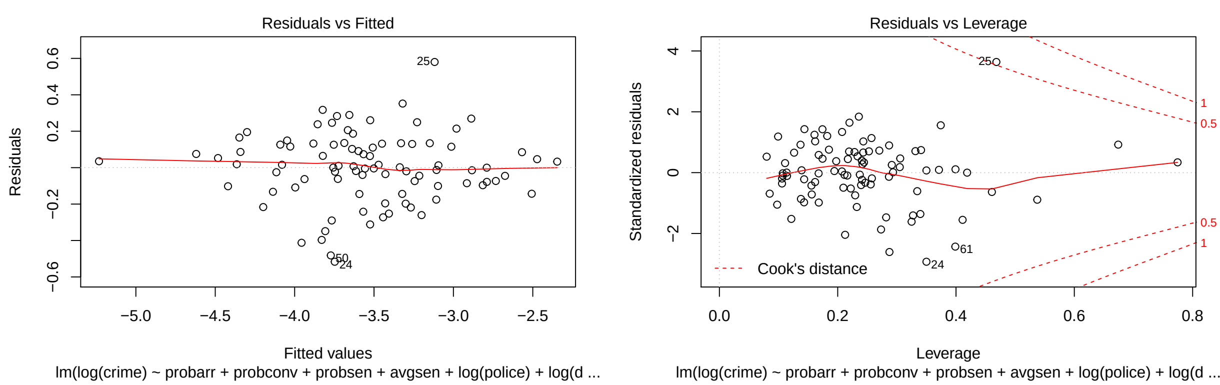
```
##      probarr     probconv      probsen       avgsen  log(police)
##     1.231873     1.677844     1.930592     1.829549     2.581464
## log(density)     log(tax)          mix         west      central
##     5.369341     2.093136     2.061905     2.184817     3.938538
##        urban        ymale       pctmin log(wagecon) log(wagetuc)
##     2.874248     1.541404     3.436532     2.138026     1.707329
## log(wagetrd) log(wagefir) log(wagemfg) log(wagefed) log(wagesta)
##     3.090761     2.666420     2.274164     3.297279     1.658977
## log(wageloc)
##     2.273518
```

4. Zero-conditional mean assumption holds because the spline line remains close to zero in the residual vs fitted plot, there is no outliers that have high influence, and the covariances are very close to zero indicating the likelihood of being exogenous.

```r
# plot residual vs fitted plot & residual vs leverage plot
plot(m3, which=c(1, 5))
# calculate the covariance for each independent variables with the model's residuals
lapply(subset(data, select=c("probarr", "probconv", "probsen", "avgsen", "log.police",
```

```
                              "log.density", "log.tax", "mix")),
      function(var) cov(var, m3$residuals))
```

```
## $probarr
## [1] -0.00000000000000000001265838
##
## $probconv
## [1] -0.00000000000000000002909986
##
## $probsen
## [1] 0.00000000000000000004238734
##
## $avgsen
## [1] -0.0000000000000000000261914
##
## $log.police
## [1] -0.00000000000000000007611876
##
## $log.density
## [1] -0.00000000000000000003151334
##
## $log.tax
## [1] 0.0000000000000000001022321
##
## $mix
## [1] 0.00000000000000000005488631
```



5. The assumption of homoskedasticity is met because even though the residuals appear to spread in a crescendo then a decrescendo motion in the scale-location plot and the p-value is statistically significant based on the Breusch-pagan test, the error variance is constant because the p-value is not statistically significant in Score-test. Also, we can use the heteroscedasticity-robust standard errors for the hypothesis tests on the slope parameters.

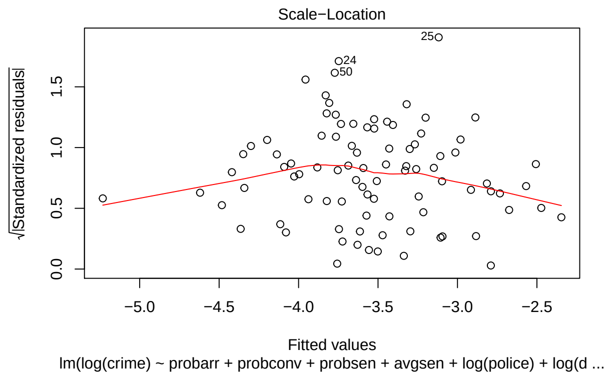
```
# plot scale-location plot
plot(m3, which=3)
# Breusch-pagan test
bptest(m3)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  m3
## BP = 48.447, df = 21, p-value = 0.0005975
```

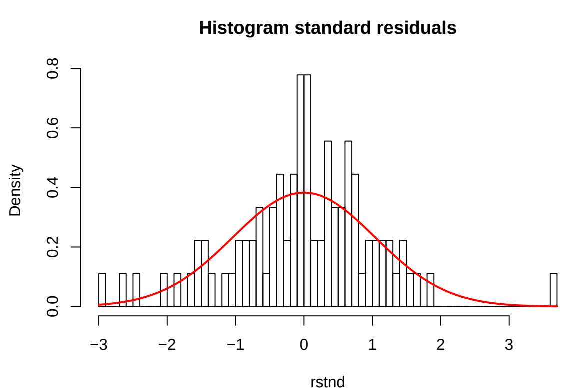
```
# Score-test for non-constant error variance
ncvTest(m3)
```

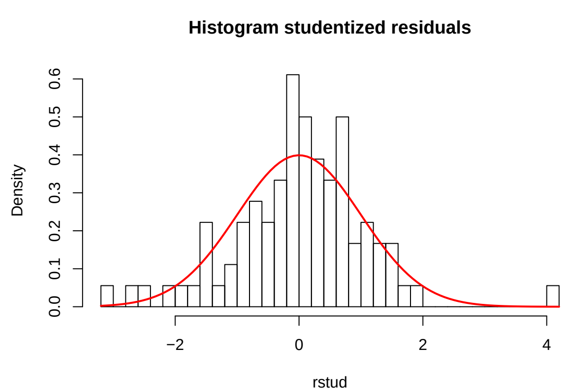
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1730897    Df = 1     p = 0.6773804
```

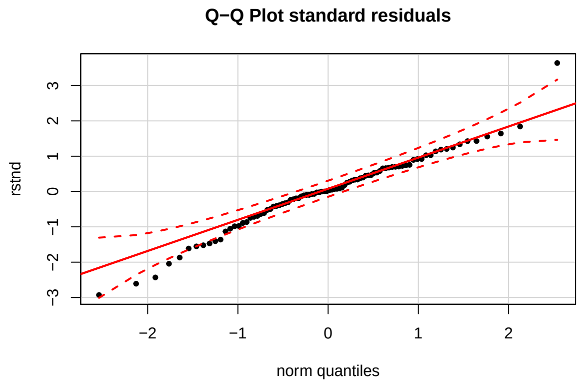


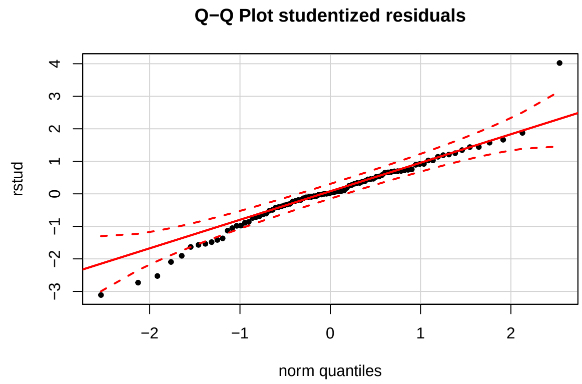
6. The distribution of the residuals appear to be approximately normal.

```
# normality of standard residuals
rstnd = rstandard(m3)
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(m3)
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)
```









Summary of Models

```
# calculate the standard error of each model
se.m1 <- sqrt(diag(vcovHC(m1)))
se.m2 <- sqrt(diag(vcovHC(m2)))
se.m3 <- sqrt(diag(vcovHC(m3)))
# report the results in a table format
stargazer(m1, m2, m3, type="text", omit.stat=c("f", "ser"),
          se=list(se.m1, se.m2, se.m3),
          star.cutoffs=c(0.05, 0.01, 0.001))
```

```
##
## ============================================
##                      Dependent variable:
##                  ---------------------------
##                          log(crime)
##                     (1)       (2)       (3)
## --------------------------------------------
## probarr           0.296     0.033    -0.108
##                  (0.727)   (0.592)   (0.373)
##
## probconv        -0.725*** -0.538*** -0.581***
##                  (0.098)   (0.131)   (0.128)
##
## probsen         -2.328*** -1.557*** -1.578***
##                  (0.366)   (0.392)   (0.260)
##
## avgsen           -0.011    -0.013    -0.011
##                  (0.019)   (0.016)   (0.015)
##
## log(police)      0.633***   0.374*    0.486**
##                  (0.137)   (0.171)   (0.183)
##
## log(density)               0.271***  0.269***
##                            (0.079)   (0.080)
##
## log(tax)                    0.147    -0.026
##                            (0.180)   (0.203)
##
```

```
##
## mix                        0.205     -0.119
##                           (0.785)    (0.663)
##
## west                                 -0.175*
##                                      (0.082)
##
## central                              -0.201
##                                      (0.139)
##
## urban                                -0.029
##                                      (0.149)
##
## ymale                                 0.751
##                                      (1.643)
##
## pctmin                                0.009**
##                                      (0.003)
##
## log(wagecon)                          0.169
##                                      (0.207)
##
## log(wagetuc)                          0.111
##                                      (0.256)
##
## log(wagetrd)                          0.209
##                                      (0.341)
##
## log(wagefir)                         -0.331
##                                      (0.318)
##
## log(wagemfg)                          0.006
##                                      (0.166)
##
## log(wagefed)                          0.271
##                                      (0.406)
##
## log(wagesta)                         -0.215
##                                      (0.292)
##
## log(wageloc)                          0.039
##                                      (0.605)
##
## Constant       1.617      -0.822     -1.025
##               (1.144)    (1.643)    (4.235)
##
## --------------------------------------------
## Observations     90         90         90
## R2             0.612      0.703      0.879
## Adjusted R2    0.589      0.674      0.841
## ============================================
## Note:          *p<0.05; **p<0.01; ***p<0.001
```

```r
# running AIC
AIC(m1)
```

```
## [1] 75.24534
```

```r
AIC(m2)
```

```
## [1] 57.13349
```

```r
AIC(m3)
```

```
## [1] 2.594278
```

What we have observed is as follows:

1. *probconv*, *probsen* and $log(police)$ come up as statistically significant in *Model 1*.

2. What is notable to point out is that *probarr* and $log(police)$ do not have negative coefficient. $log(police)$ makes sense because polices most likely occupy places with high activities of crime. As for the *probarr*, there might be underlying variables that are not yet considered in the model.

3. In *Model 2*, where direct and indirect variables to crime rate are added to the model, *probconv* and *probsen* remain statistically significant while $log(police)$ drops its statistically significance with the introduction of $log(density)$. What this tells us is that *police* and *density* are positively correlated. *density* has an independent effect on *police*. Then a regreession that omits *density* like in Model 1 will overstate the effect of *police*. What this also tells us is that there appears to be more crime in densely populated area because a 1% increase in people per square mile will reflect a 0.271% increase in crime rate.

4. When we dumped all the covariates in, like in the case of *Model 3*, *probconv*, *probsen*, $log(density)$ and $log(police)$ remain statistically significant, but other variables such as percentage of minority *pctmin* and *west* become statistically significant at the 0.01 and 0.05 level respectively. We were not expecting statistical significance of *pctmin* and *west*, so it has become an interesting finding that we need to consider.

5. The explanatory variables of key variables do not change much from Model 1 to Model 2 to Model 3 with the exception of *probarr*. Again we need to consider the inclusions of other variables when it comes to employing *probarr* in the model. But overall, the variables we have chosen seem to be good choices as explanatory variables of key interest

6. Model 3 includes the most variables and naturally has a higher R^2 value. One common concern for including many variables is that the multi-collinearity between the variables inflates the variances of the OLS estimates. But as shown in the analysis of Model 3, we did not encounter any high VIF for the variables. This could be a main reason why Model 3 is outperforming the other two models, as it explains more of the dependent variable without introducing much more noise.

7. Based on *Model 3*,

- *probconv*: keeping all other variables constant, a 0.1 increase in *probconv* results in a 5.81% decrease in crime rate, which is practically significant.
- *probsen*: keeping all other variables constant, a 0.1 increase in *probsen* results in a 15.78% decrease in crime rate, which again is practically significant.
- $log(police)$: keeping all other variables constant, a 1% increase in police per capita increases crime rate by 0.486%.
- $log(density)$: keeping all other variables constant, a 1% increase in density increases crime rate by 0.269%.
- *west*: keeping all other variables constant, being in west region decreases crime rate by 17.5%, which is practically significant.

Discussion of Causality

To make a causal model, we need to take into account for every factor except our x's and put them into the error term. As long as the error term doesn't change as we manipulate x, our model coefficents have a causal interpretation. In our case, causality cannot be fully determined since these models do not encompass all causal variables and may have three types of endogeneity bias: omitted variable bias, reverse causality and measurement error.

Our model does not include many factors that have been known to increase the risk of criminal behavior. We will list the factors below:

1. Income Inequality: There a few reasons why living in areas of high income inequality may lead to higher crime rates. Economic and social stress may affect parental practices (i.e. Households with lower incomes may have less attentive parents due to working multiple jobs to meet ends meet) leading to neglect and poor supervision of juveniles, hence increasing the risk of juvenile involvement in crime. Furthermore, areas with high income inequality may also bring those motivated to offend in close spatial contact with attractive targets for crime thus increasing the likelihood of criminal behavior. We reason that higher densely populated areas may suffer more from income inequality. Therefore given the positiv correlation between income inequality, density and crime rate, we would expect the omitted variable bias in this case to be positive.
2. Juvenile School Performance and Truancy: Studies have shown that juveniles with lower academic performance are more likely to offend, more likely to offend frequently, and more likely to persist in crime. Additionally, chronically truant juveniles, who end up dropping out all together, earn significantly less income during their lifetime, and subsequently are more likely to turn to crime. Given the varying education systems in different counties (west, central or neither), we believe such regional identification is endogenous and correlated with other factors contributing to crime rate such as juvenile school performance and truancy. While we expect a negative relationship between school perfromance and crime rate and a positive relationship between truancy and crime rate, we cannot conclude on a direction for this omitted variable bias without additional information on the actual systems within each region.

Reverse causality refers either to a direction of cause-and-effect contrary to a common presumption or to a two-way causal relationship in, as it were, a loop. It appears that police per capita variable may be subject to such a bias. As crime rate increases in an area, it makes sense that more law enforcement may be needed and thus police per capita may increase. In our model, keeping all other variables constant, a 1% increase in police per capita increases crime rate by 0.486%. This demonstrates a two-way causal relationship.

Lastly, there is the possibility of measurement error within our model, specifically around the proportion variables, *probconv* and *probsen*, with values greater than 1. Though we assumed these values are valid to avoid ommitted bias in our models, we recognize the possibility that the values may also have arised due to measurement procedures or calculations. However, without further details on how the data was collected, we will not be able to estimate the direction and size of this bias.

It is worth highlighting though that when we plot our independent variables against the model residuals, we find that the residuals appear unrelated to the values of the independent variables thus demonstrating exogenity. However, we understand that causality is not the same as exogeneity. Exogeneity is about whether OLS can correctly identify the beta coefficients while causality has stricter assumptions and is about whether manipulations to the explanatory variables do not infludence the error term.

Conclusion

Based on our best model, we have the following statiscally significant key variables of interest: probability of conviction, probability of sentencing, and police per capita. With all other variables controlled, an increase in the probability of conviction or probability of sentencing reduces the crime rate. Although we see that an increase in police per capita causes an increase in the crime rate, we suspect that there is a reverse causality bias, whereas in actuality, the cause is the crime rate and police per capita is the effect. With this potential bias, we concluded that police per capita is not a good determinant for crime rate.

Relating our analysis results to our research question: we see that strict incapacitations through conviction and imprisonment was effective in reducing crime rate in 1988 for the North Carolina counties. The use of this finding to support current policy suggestion is subject to a few limitations: 1) our analysis was based on data taken in North Carolina and it's not representative of the situation for United States, 2) our analysis was based on data taken in 1988 and our analysis might not be pertinent to be applied to the model for 2017 and 3) our analysis was based on one sample of 90 observations which limits our capability to find conclusive enough evidences.

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.