Lab 3 - Reducing Crime [FINAL]

Linda Dong, Shahbakht Hamdani, and Arturo Esquerra

3/28/2018

Introduction

Providing a safe environment for American citizens and their families is one of the foremost obligations of any public servant, especially local authorities. With this in mind, it is imperative to have a working understanding of which factors are drivers for crime rates in the State of North Carolina. Once these determinants of crime are identified, quantified, and ordered in importance via statistic modelling, it will be possible to effect significant changes to improve security by creating policies to address these factors. This project will study a wide variety of economic, judicial, and demographic variables, devoting a special focus on the variables that are most actionable from a public publicy perspective.

Exploratory Data Analysis

The *crime_data* dataset contains information at the county level for the State of North Carolina.

```
summary(crime_data)
```

```
##      county           year         crmrte              prbarr      
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277  
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568  
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095  
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492  
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438  
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091  
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6        
##         prbconv      prbpris           avgsen           polpc        
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746  
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231  
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485  
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702  
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877  
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054  
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6         
##     density             taxpc             west           central       
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000  
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736  
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000  
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000  
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6       
##      urban            pctmin80           wcon            wtuc       
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6  
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6  
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5  
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7  
```

```
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd             wfir             wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed             wsta             wloc             mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

We observe that there are missing values, which upon closer inspection are found to be caused by the csv file itself which had 6 extra rows which are removed as they don't have any value.

```
crime_data = crime_data[1:91,] #Removed the unnecesary values
```

In total we have observations of 91 counties for the year of 1987. For each county we have measured a total of 25 variables which include demographic, location, law-enforcement, and economic variables gathered from different sources.

```
#Structure of the data
str(crime_data)
```

```
## 'data.frame':    91 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
```

```
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

Given that these variables were read from a csv file (doesn't hold meta-data about the variable types), it is necessary to recode some of them to their appropriate variable type.

```
#Changing the variable types to the correct ones
crime_data$county = as.factor(crime_data$county)
crime_data$prbconv = as.numeric(levels(crime_data$prbconv))[crime_data$prbconv]
```

```
## Warning: NAs introduced by coercion
```

```
crime_data$west = as.factor(crime_data$west)
crime_data$central = as.factor(crime_data$central)
crime_data$urban = as.factor(crime_data$urban)
```

We will use the **Crime Rate** (crmte) as the dependent variable in this study and will attempt to predict it based on a series of regressors which will be classified into three buckets based on each variable's objective in the current study. These buckets are based on prior knowledge of the subject as well as the literature (Cornwell 1994 p. 360), which states that the two most important crime deterrants are law enforcement and economic variables such as the labor market:

Law Enforcement: Probability of arrest (prbarr), Probability of conviction (prbconv), Probability of prision sentence (prbpris), Average sentence (avgsen), and Police per capita (polpc)

Economic factors: Weekly (average) wage construction (wcon), Weekly (average) wage transport, util, and commun (wtuc), Weekly (average) wage wholesale, retail, and trade (wtrd), Weekly (average) wage financial institutions and real estate (wfir), Weekly (average) wage service industry (wser), Weekly (average) wage manufacturing (wmfg), Weekly (average) wage federal employees (wfed), Weekly (average) wage state employees (wsta), Weekly (average) wage local government employees (wloc), Tax revenue per capita (taxpc)

Other (demographic covariates): Density per square mile (density), Location (west, central), Urban Metropolitan Statistical Area (urban), Percent of the population that are minority, 1980 (pctmin80), Offense mix: ratio of crimes involving "face-to-face" contact (e.g., robbery) to those that do not (mix), Percent young male (pctymle)

Transformations of the Dependent Variable

As can be observed in the previous sections, the dependent variable (crime rate) has very low decimal values with the highest being 0.0989659. This has several important consequences for the linear model: first, estimates of the regressors will have very small values, since we are not applying a model specifically designed to allocate variables that are limited to values between [0,1] it is possible to obtain predictions smaller than 0, and finally interpretation of the model estimates becomes complicated due to their expected small values.

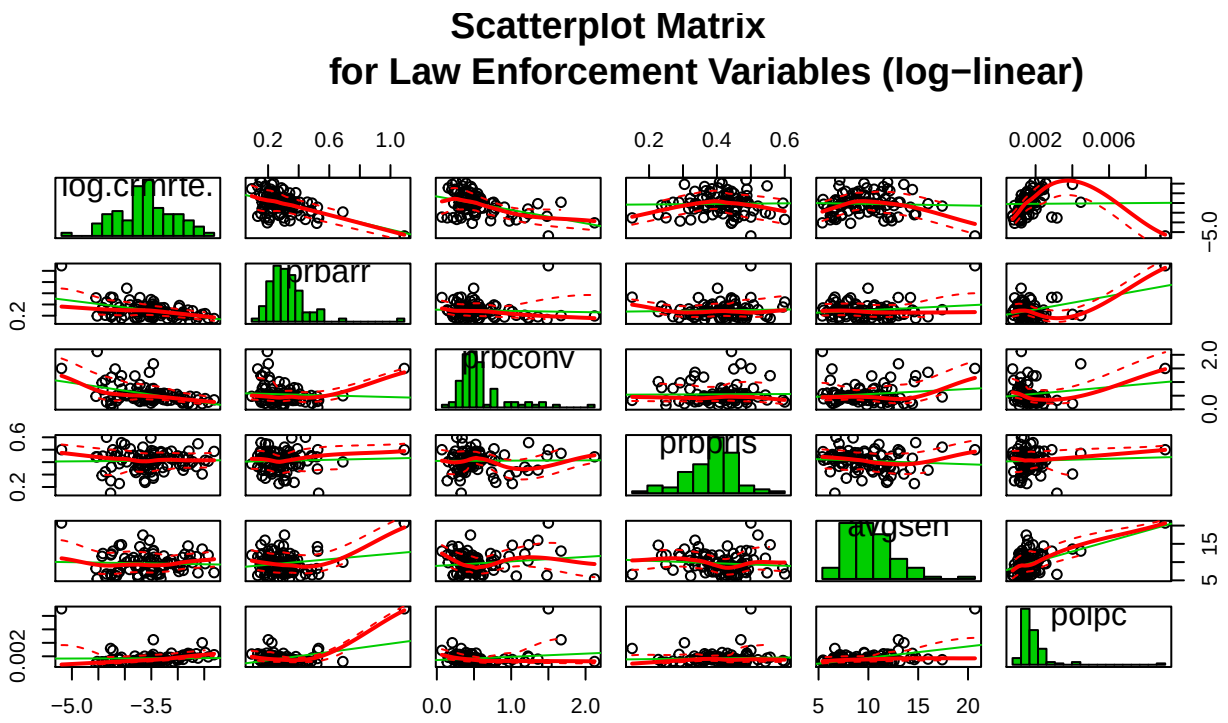
With this in mind, and given that one of the primary objectives of this study is the interpretability of the results (together with the predicting power), we propose the analysis of log-linear models which are interpreted as semi-elasticities. That is, the model we will be considering will be of the form:

$$ln(y_i) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u\text{: log-linear model}$$

Scatterplot Matrices

Given the large amount of regressors, we will split the scatterplot matrix into three buckets: one for each of the dimensions described before.

```
scatterplotMatrix(~  log(crmrte) + prbarr + prbconv + prbpris + avgsen + polpc, data = crime_data,
                  diagonal="histogram", main = "Scatterplot Matrix
                  for Law Enforcement Variables (log-linear)")
```



In the histograms we observe the pressence of large values for the *prbarr* and *prbconv* variables.

```
crime_data[crime_data$prbarr<=0 | crime_data$prbconv<=0 | crime_data$prbpris<=0,1:6]
```

```
## [1] county  year    crmrte  prbarr  prbconv prbpris
## <0 rows> (or 0-length row.names)
```

```
crime_data[crime_data$prbarr>=1 | crime_data$prbconv>=1 | crime_data$prbpris>=1,1:6]
```

```
##    county year    crmrte   prbarr prbconv  prbpris
## 2       3   87 0.0152532 0.132029 1.48148 0.450000
## 10     19   87 0.0221567 0.162860 1.22561 0.333333
## 44     99   87 0.0171865 0.153846 1.23438 0.556962
## 51    115   87 0.0055332 1.090910 1.50000 0.500000
## 56    127   87 0.0291496 0.179616 1.35814 0.335616
## 61    137   87 0.0126662 0.207143 1.06897 0.322581
## 67    149   87 0.0164987 0.271967 1.01538 0.227273
## 84    185   87 0.0108703 0.195266 2.12121 0.442857
## 90    195   87 0.0313973 0.201397 1.67052 0.470588
## 91    197   87 0.0141928 0.207595 1.18293 0.360825
```

We observe that several probabilities are above one which is an impossibility given the strict definition of this metric. This could be caused by errors in the data, inconsistencies between the numerator and denominator, or double-counting cases in the numerator. In any case, having probabilities above one can disturb the interpretability of our model, therefore we will trust that these are high probabilities but will simply recode

them as one, which means that phenomenon will happen with certainty. It is important to mention that these issues only arise with the prbconv and prbarr variables, being the former the one with the most inconsistencies. The counties with issues are 3, 19, 99, 115, 127, 137, 149, 185, 195, 197 (we note that observation 51 which correspond to county 115 is the only case where more than one probability is larger than one, we will keep special attention on this county).

```
crime_data[crime_data$prbarr>=1, 4] = 1
crime_data[crime_data$prbconv>=1, 5] = 1
```

```
head(sort(crime_data$polpc, decreasing = TRUE))
```

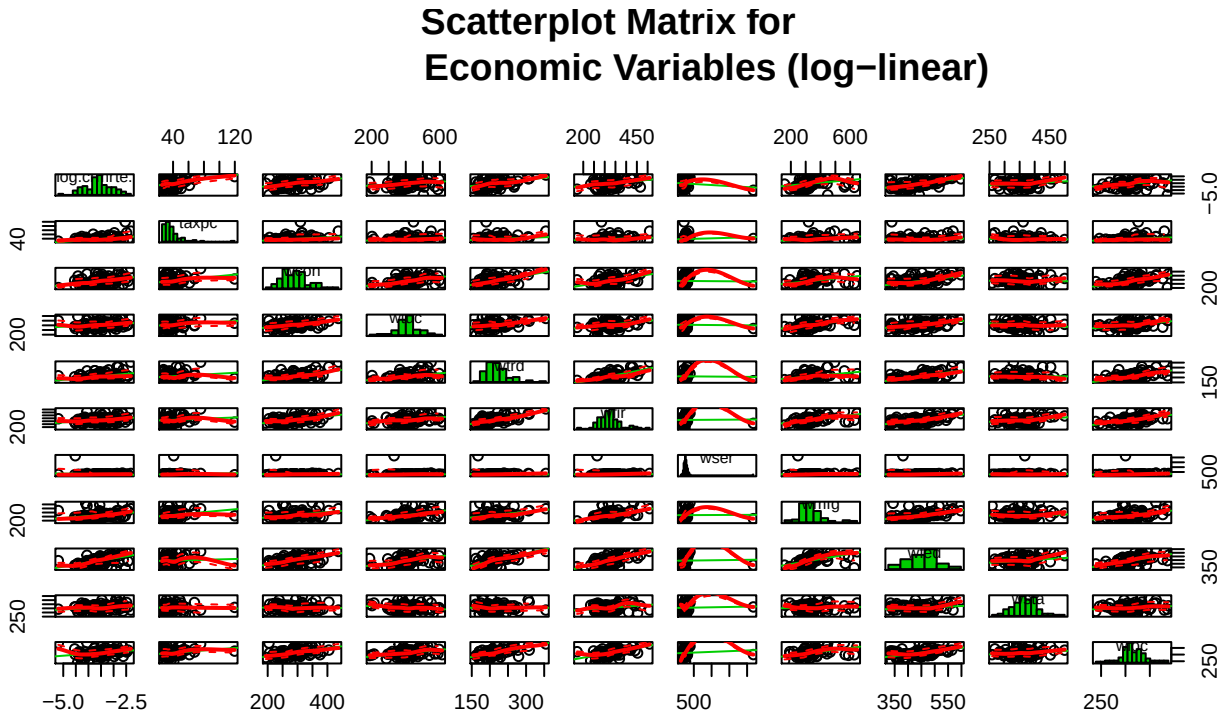
```
## [1] 0.00905433 0.00445923 0.00400962 0.00316379 0.00288203 0.00255849
```

```
crime_data[crime_data$polpc>0.005,]
```

```
##    county year    crmrte prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332      1        1     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

We also note a large outlier in the police per capita variable for county 115, which incidentally also had issues with other law-enforcement variables. One possible explanation for having so many outlier values in this dimension could be that the crime in this county is very distinct from the others in North Carolina and is being driven by special characteristics of the county.

```
scatterplotMatrix(~  log(crmrte) + taxpc + wcon + wtuc + wtrd + wfir +
                    wser + wmfg + wfed + wsta + wloc,
                  data = crime_data,
                  diagonal="histogram", main = "Scatterplot Matrix for
                  Economic Variables (log-linear)")
```



```r
head(sort(crime_data$wser, decreasing = TRUE))
```

```
## [1] 2177.0681  391.3081  354.3007  348.2754  347.6609  320.1325
```

```r
crime_data[crime_data$wser>1000,]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 84    185   87 0.0108703 0.195266       1 0.442857   5.38 0.0012221
##      density    taxpc west central urban pctmin80     wcon    wtuc
## 84 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 84 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944
##       pctymle
## 84 0.07008217
```

```r
crime_data[84,c(1,15:23)]
```

```
##    county     wcon    wtuc     wtrd     wfir     wser   wmfg   wfed   wsta
## 84    185 226.8245 331.565 167.3726 264.4231 2177.068 247.72 381.33 367.25
##      wloc
## 84 300.13
```

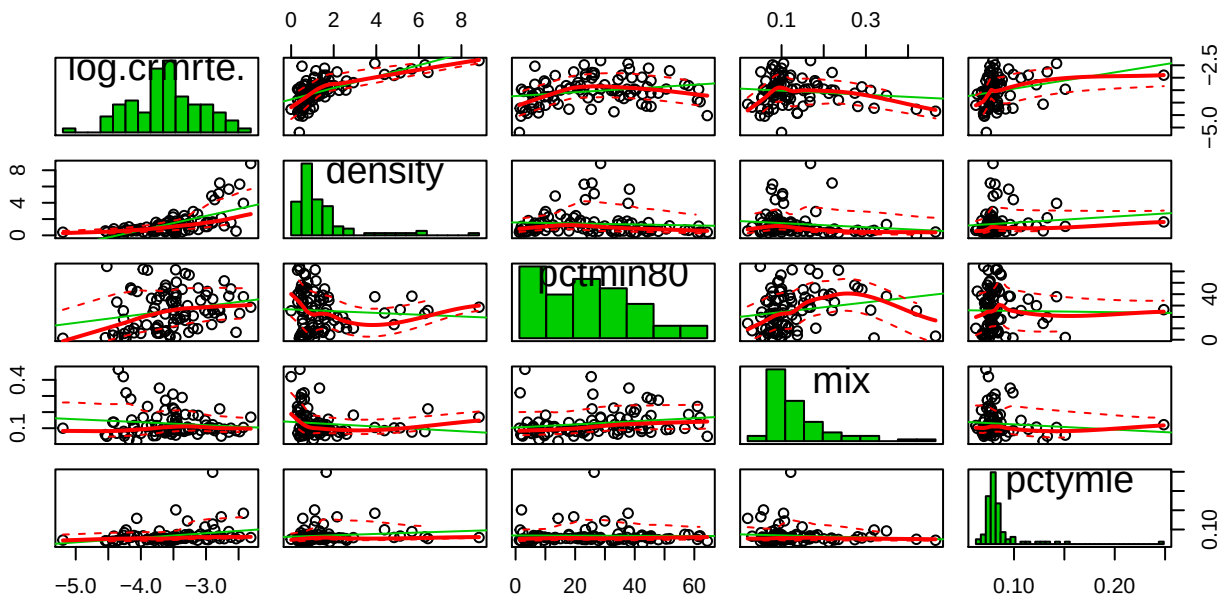
We note a large atypical value in the weekly wage for the service industry. Upon further inspection, we note that Entry 84 (county #187 in the county identifier variable) has an atypically large weekly salary. If the county code is the county's FIPS code, it would be Warrenton NC which isn't a particularly wealthy county. This salary implies a yearly income above USD $110,000 which would be extremely uncommon for this particular industry. Furthermore, looking at the "Employment, Hours, and Earnings Vol. II" report from the US Department of Labor from 1909 - 1990 we find no indication that such a weekly salary existed in 1987 in the US for this industry. Furthermore, we see that none of the weekly wages for any of the other industries were anyqhere near the value for this county. With this in consideration and the fact that this wage will likely be a very influential observation on a critical variable, we will acknowledge that this county might have a large weekly wage for the service industry and will set its value to the second largest weekly wage for this industry in North Carolina.

```r
head(sort(crime_data$wser, decreasing = TRUE))
```

```
## [1] 2177.0681  391.3081  354.3007  348.2754  347.6609  320.1325
```

```r
crime_data[crime_data$county==185, 19] = head(sort(crime_data$wser, decreasing = TRUE))[2]
```

```r
scatterplotMatrix(~  log(crmrte) + density + pctmin80 + mix + pctymle,
                  data = crime_data,
                  diagonal="histogram", main = "Scatterplot Matrix for
                  Other Covariates (log-linear)")
```



Among the demographic variables we notice some outliers on the people per square mile density and percent male variables.

Overall looking at all the exploratory data analysis plots, we observe that there seems to be a negative linear relationship between the crime rate and the probability of arrest. Furthermore, we also note a negative relationship between crime rate and the probability of conviction, however, this relationship doesn't seem to be as linear with a lot of different crime rates located at counties with low probability of conviction. It is important to note that there also seems to be a very tenuous relationship between crime rate and the variables probability of prision sentence and average sentence. We also note a positive linear relationship between crime rate and police per capita.

Moreover, we observe a non-linear positive relationship between the crime rate and the tax revenue per capita which is a proxy for economic level of the county. Finally, we note a non-linear relatioship between most of the wage variables (with the exception of the weekly wage for the service industry) and the crime rate.

```
round(cor(crime_data[c("crmrte", "prbarr", "prbconv", "prbpris",
                       "avgsen", "polpc", "taxpc", "wcon",
                       "wtuc", "wtrd", "wfir", "wser", "wmfg",
                       "wfed", "wsta", "wloc")]), digits = 3)
```

```
##         crmrte prbarr prbconv prbpris avgsen  polpc  taxpc   wcon   wtuc
## crmrte   1.000 -0.399  -0.417   0.047  0.027  0.170  0.451  0.392  0.229
## prbarr  -0.399  1.000  -0.109   0.039  0.158  0.387 -0.135 -0.249 -0.084
## prbconv -0.417 -0.109   1.000  -0.048  0.138  0.043 -0.207 -0.071  0.020
## prbpris  0.047  0.039  -0.048   1.000 -0.096  0.047 -0.093 -0.060  0.126
## avgsen   0.027  0.158   0.138  -0.096  1.000  0.490  0.096 -0.030  0.215
## polpc    0.170  0.387   0.043   0.047  0.490  1.000  0.284 -0.024  0.166
## taxpc    0.451 -0.135  -0.207  -0.093  0.096  0.284  1.000  0.263  0.163
## wcon     0.392 -0.249  -0.071  -0.060 -0.030 -0.024  0.263  1.000  0.408
## wtuc     0.229 -0.084   0.020   0.126  0.215  0.166  0.163  0.408  1.000
## wtrd     0.410 -0.108  -0.142   0.140  0.081  0.112  0.167  0.555  0.361
## wfir     0.330 -0.184   0.049   0.034  0.164  0.189  0.124  0.487  0.333
## wser     0.280 -0.285   0.031  -0.001 -0.030  0.134  0.244  0.472  0.365
```

```
## wmfg     0.354 -0.168   0.026   0.009  0.116  0.272  0.261  0.347  0.462
## wfed     0.486 -0.218  -0.042   0.086  0.144  0.159  0.058  0.506  0.401
## wsta     0.202 -0.165  -0.149  -0.032  0.134  0.051 -0.031 -0.019 -0.157
## wloc     0.348 -0.048   0.025   0.083  0.124  0.375  0.207  0.512  0.341
##           wtrd   wfir   wser   wmfg   wfed   wsta   wloc
## crmrte   0.410  0.330  0.280  0.354  0.486  0.202  0.348
## prbarr  -0.108 -0.184 -0.285 -0.168 -0.218 -0.165 -0.048
## prbconv -0.142  0.049  0.031  0.026 -0.042 -0.149  0.025
## prbpris  0.140  0.034 -0.001  0.009  0.086 -0.032  0.083
## avgsen   0.081  0.164 -0.030  0.116  0.144  0.134  0.124
## polpc    0.112  0.189  0.134  0.272  0.159  0.051  0.375
## taxpc    0.167  0.124  0.244  0.261  0.058 -0.031  0.207
## wcon     0.555  0.487  0.472  0.347  0.506 -0.019  0.512
## wtuc     0.361  0.333  0.365  0.462  0.401 -0.157  0.341
## wtrd     1.000  0.670  0.473  0.357  0.638 -0.001  0.591
## wfir     0.670  1.000  0.526  0.491  0.624  0.235  0.559
## wser     0.473  0.526  1.000  0.469  0.539  0.068  0.539
## wmfg     0.357  0.491  0.469  1.000  0.515  0.055  0.439
## wfed     0.638  0.624  0.539  0.515  1.000  0.186  0.520
## wsta    -0.001  0.235  0.068  0.055  0.186  1.000  0.156
## wloc     0.591  0.559  0.539  0.439  0.520  0.156  1.000
```

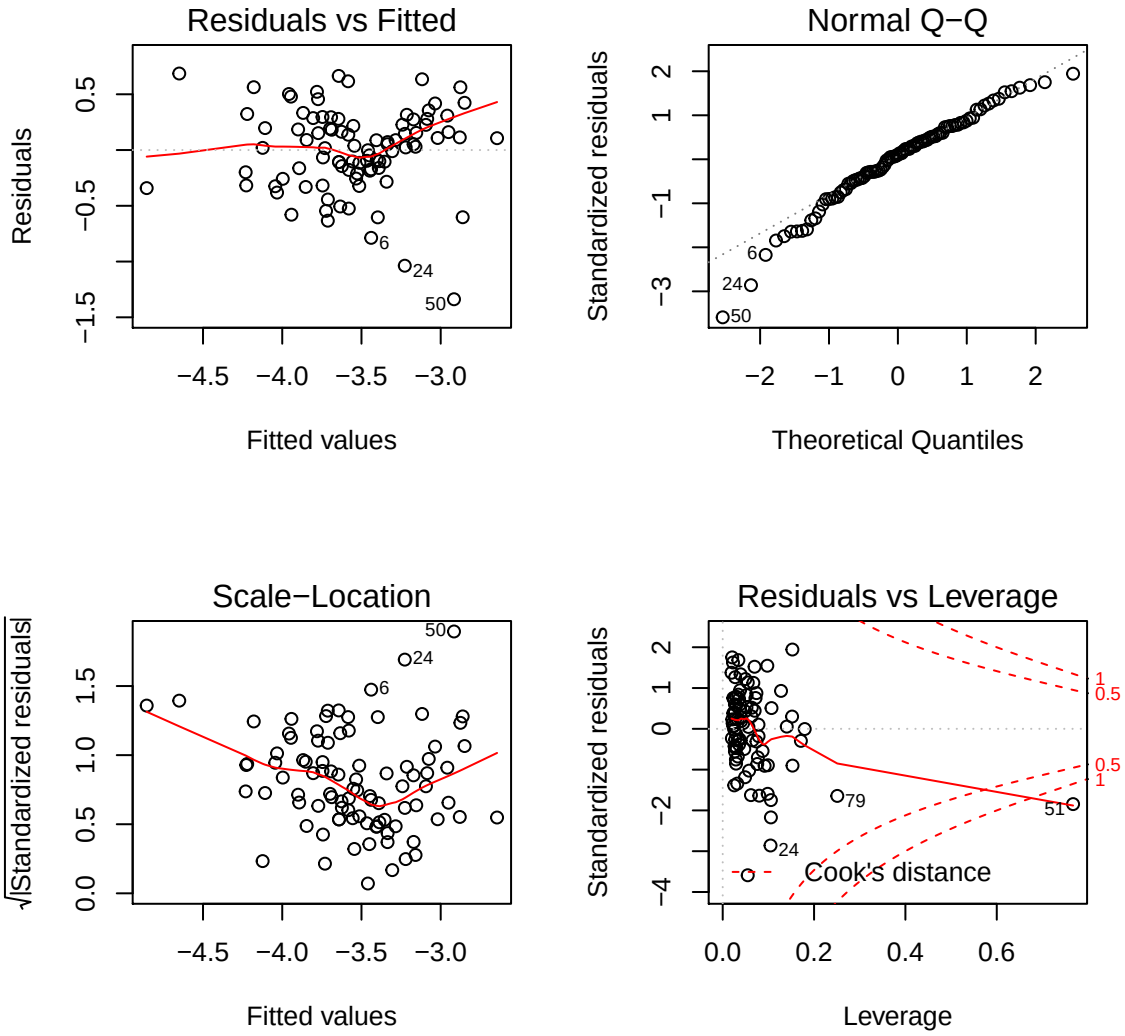
The correlation matrix provides additional insights regarding the linear relationship between the dependent variable and the predictors. Among the law enforcement variables, we see moderate negative correlations between the crime rate and the probabilities of arrest and conviction. Furthermore, between the probability of arrest, conviction, and prision sentence we see little linear relationship. However, we do note positive relationships between average sentence and police presence which could cause multi-collinearity in the OLS model. We also note strong correlations between the dependent variable and most of the labor market variables (weekly wages). However, including all of them might be problematic as they are very inter-correlated which will increase the multi-collinearity of the model.

Models

Log-Linear Models

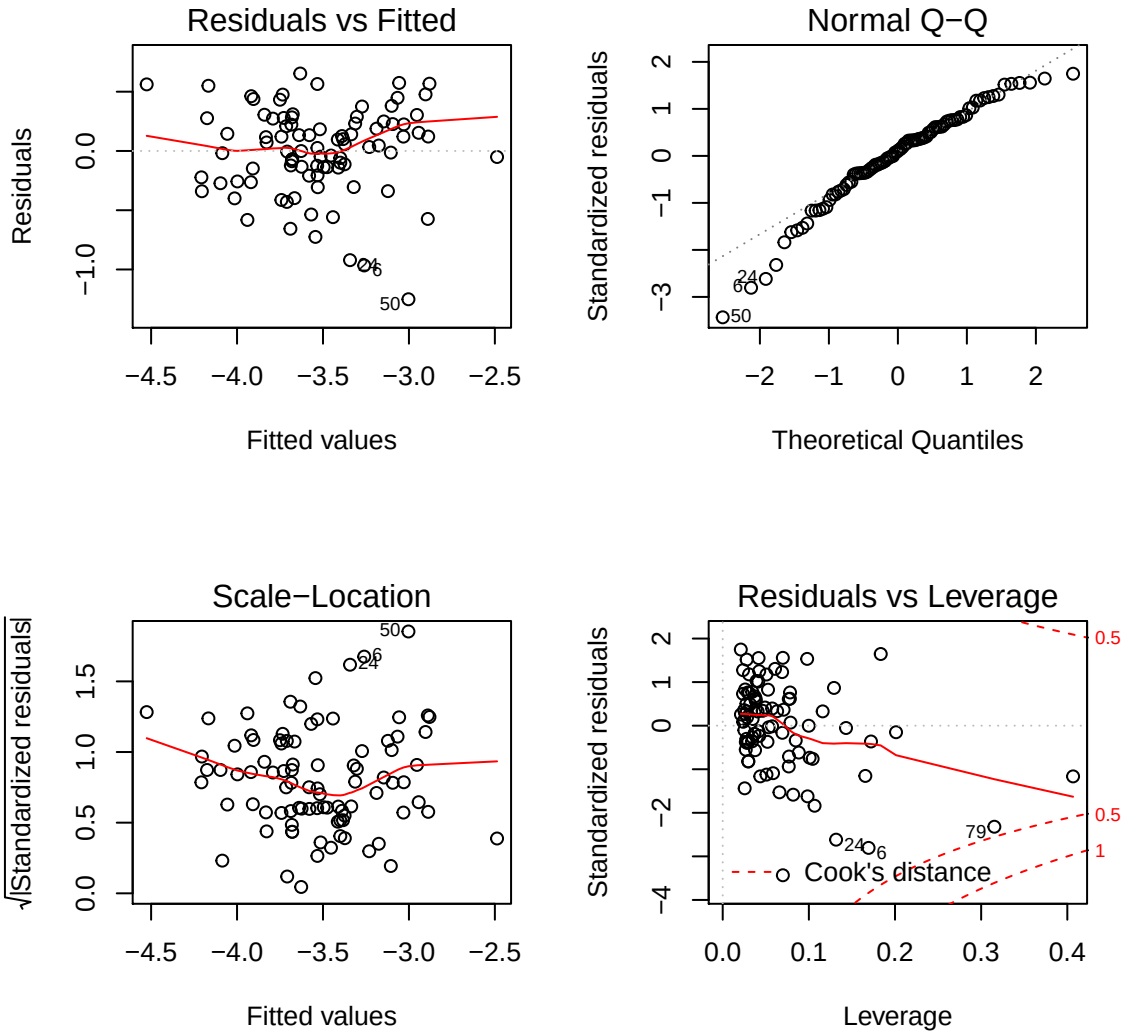
The model-building process will be conducted using the following logic. First, the most important and actionable variables will be introduced into the model. For instance, variables such as *pctymle* which can't be addressed through public policy won't be as prioratized as income or law-enforcement variables. Second, the model will be constructed iteratively, adding additional variables and having in consideration the potential omitted variable bias that the simpler models will likely have. The goal is to have a model that reflects at least the two most important dimensions that are determinants of crime according to literature: crime and law-enforcement. Afterwards, model validation will be conducted to assess whether the Classical Linear Model assumptions are fulfilled. If this happens, the model will be refined through t and F tests to obtain a parsimonous and interpretable model.

```
#Model 1: Focus on law enforcemnt
model_log_linear_1 = lm(log(crmrte) ~ prbarr + prbconv + prbpris +
                                      polpc + avgsen, data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_1)
```



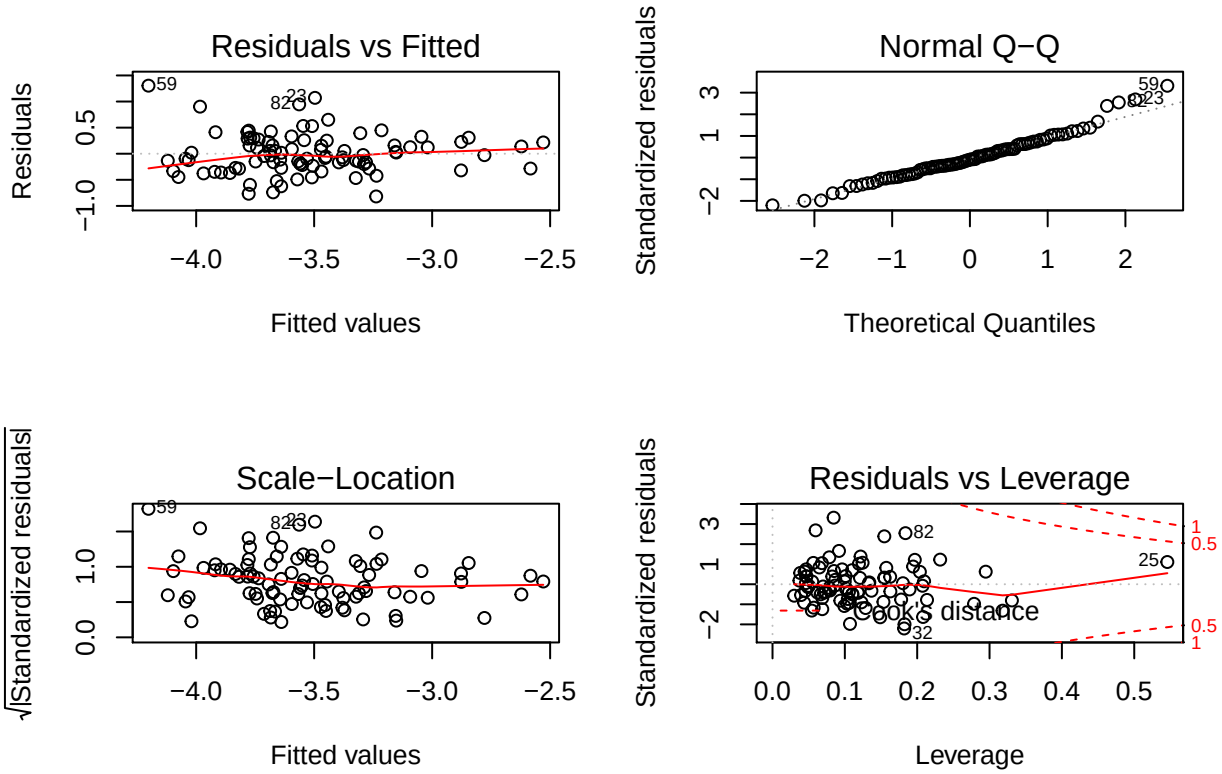
We note that observation 51 (county 115) has high leverage and influence with a Cook's distance larger than one. Given the issues we have noted before with this entry, and the fact that this county appears to have different crime dynamics than the rest, we propose eliminating it from the analysis and then, in a separate study, examine what is going on with its crime rate. This is mainly because we suspect that the differences in crime between county 115 and the rest is not simply a difference in level, which would mean that a single indicator variable for this county wouldn't be enough to explain them. Therfore, the incorporation of a series of interactions would be necessary to study this phenomenon. Unfortunately, this would mean that we would have to use several of our available degrees of freedom to simply be able to explain a single county (ovserspecifying our model). With this in mind, and given that our objective is to produce sensible public policy for the entire State, we consider that eliminating it is the most sensible thing to do.

```
#Model 2: Focus on law enforcemnt with indicator for county 115
crime_data = crime_data[-c(51),]
model_log_linear_2 = lm(log(crmrte) ~ prbarr + prbconv + prbpris +
                                      polpc + avgsen, data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_2)
```



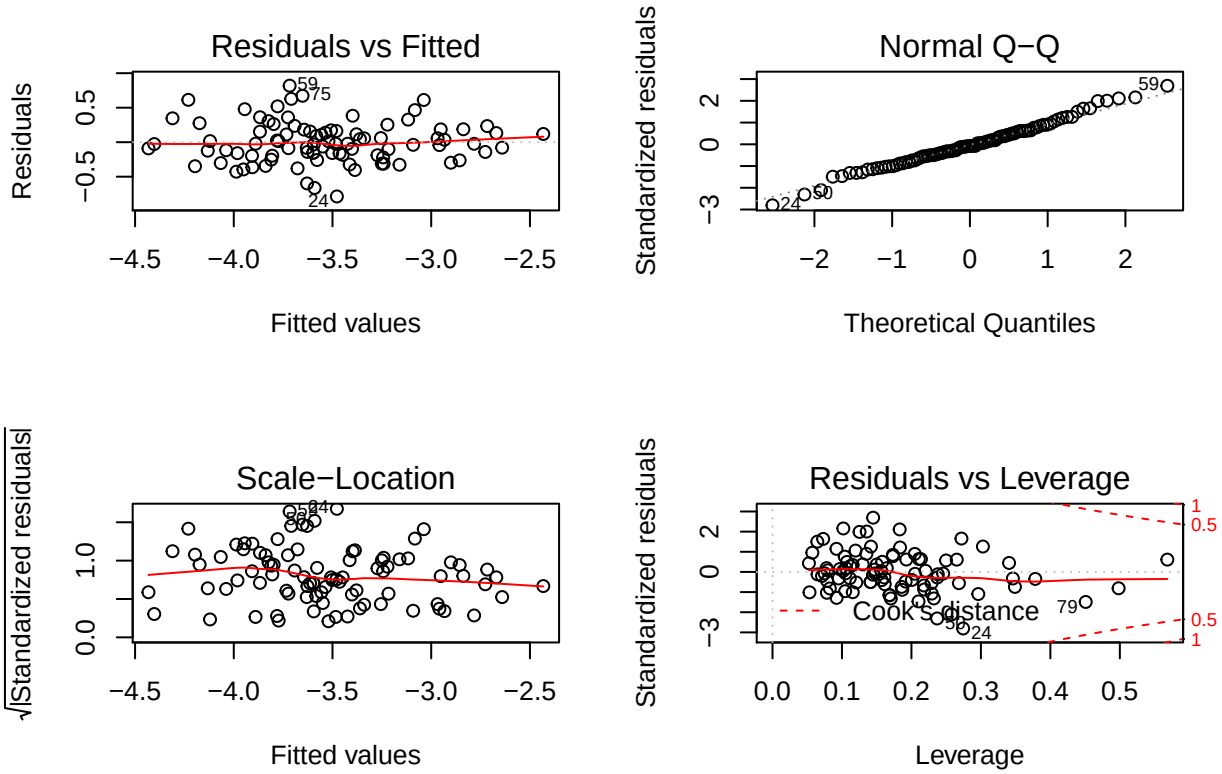
As expected due to the absence of economic variables, we fail to fulfill the zero-conditional mean assumption. Given that we expect a negative relationship in the population model between the crime rate and the economic variables (the higher the wealth of the county, the lower the crime rate) and an expected positive relationship between the economic variables and the judicial variables (richer counties have better access to law enforcemnt), we expect a negative omitted variable bias that will move the current beta coefficients towards zero.

```
#Model 3: Focus on economic variables
model_log_linear_3 = lm(log(crmrte) ~ taxpc + wcon + wtuc + wtrd + wfir +
                                      wser + wmfg + wfed + wsta + wloc,
                        data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_3)
```



Similarily to what was seen in "model_log_linear_2", we note a considerable omitted variable bias that causes issues with the zero-conditional mean assumption, given the lack of law-enforcement variables.

```
#Model 4: Full model of law-enforcement and economic variables
model_log_linear_4 = lm(log(crmrte) ~ prbarr + prbconv + prbpris +
                                      avgsen + polpc + taxpc + wcon +
                                      wtuc + wtrd + wfir + wser + wmfg +
                                      wfed + wsta + wloc,
                        data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_4)
```



The model with all law-enforcement and economic (labor) variables includes most of the actionable variables available to the researchers and therefore could be of significant importance in the creation of public policy.

Assumption Validation of Model 4 (model_log_linear_4)

MLR1: Linearity

Assumption 1 of the CLM is fulfilled given that there is no evidence or theoretical reason we are aware of that could lead us to doubt the linearity of the parameters of the population model for the crime rate. ### MLR2: Random Sampling and Lack of Autocorrelation Considering that the State of North Carolina had a total of 100 counties in 1987 (1987 Census of Governments: Preliminary report. Government units in 1987, APPENDIX A-161) and that we don't have any evidence that the sampling of the 91 counties included in this dataset was done in a non-random way, we consider this assumption fulfilled. However, it is very important to consider that there is a strong possibility of having spatial autocorrelation in the model, where the crime rate of a county depends on the crime rate of the neighboring counties. The existence of such spatial structure of dependency could make our estimates inconsistent and biased. While this issue is certainly worthy of keeping into consideration, we have no way of assessing the spatial autocorrelation between the counties given that we don't know the neighboring structure of the data to calculate statistics that could shed light into the issue such as Moran's-I. Therefore, we will proceed considering MLR2 fulfilled but keeping the spatial structure in mind while interpreting the models and developing conclusions.

```
dwtest(model_log_linear_4, order.by=crime_data$county)
```

```
##
##  Durbin-Watson test
##
## data:  model_log_linear_4
## DW = 2.175, p-value = 0.7968
## alternative hypothesis: true autocorrelation is greater than 0
```

Finally, while the Durbin-Watson test shows no indication of autocorrelation, it is important to take into consideration that it only tests whether there is autocorrelation at lag 1, but given that the data doesn't have a particular order (just county number which doesn't necessarily reflect the county's location in space) the results of the test are inclusive which means that spatial autocorrelation is still an issue to consider.

MLR3: No Perfect Colinearity

While we do expect some colinearity between our variables (especially the weekly wage set), we don't see evidence of perfect colinearity considering that the model was adjusted by R without any warnings about singlularity issues. ### MLR4: Zero-Conditional Mean The fitted values vs. residuals scatterplot above displays no critical issues that would compromise the zero-conditional mean assumption. Fuerthermore, we note that the spline line adjusted by R is practically flat and centered in zero which further confirms the validity of assumption MLR4. ### MLR5: Homoskedasticity Even though the Scale-Location plot shown above shows no serious violations to homoskedasticity, considering that the null hypothesis of homoskedasticity of the Breusch-Pagan test shown below is rejected with a p-value of 0.03451 at a significance level of $\alpha = 5\%$, we will proceed conservatively using White's Heteroscedasticity-consistent standard errors.

```
bptest(model_log_linear_4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_log_linear_4
## BP = 26.349, df = 15, p-value = 0.03451
```

MLR6: Normality

Model's 4 QQ-plot shows no serious violation of Normality of the standardized residuals. Furthermore, the Shapiro-Wilk test has an associated p-value to the null hypothesis of Normality equal to 0.5872 which isn't rejected at any reasonable significance level. Furthermore, the sample size of the data which is equal to 90 ($N > 30$) allows us to rely on a version of the Central Limit Theorem and assume Normality.

```
shapiro.test(model_log_linear_4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_log_linear_4$residuals
## W = 0.98804, p-value = 0.5872
```

Taking into consideration that model_log_linear_4 fulfills all the assumptions of the Classical Linear Model, we will proceed with t and F statistical tests to refine it through the elimination of statistically insignificant variables at a pre-determined $\alpha = 10\%$.

Omnibus test

```
linearHypothesis(model_log_linear_4, c("prbarr = 0", "prbconv = 0", "prbpris = 0",
                                       "avgsen = 0", "polpc = 0", "taxpc = 0",
                                       "wcon = 0", "wtuc = 0", "wtrd = 0",
                                       "wfir = 0", "wser = 0", "wmfg = 0",
                                       "wfed = 0", "wsta = 0", "wloc = 0")
                                       , vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## prbpris = 0
## avgsen = 0
## polpc = 0
## taxpc = 0
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wfed = 0
## wsta = 0
## wloc = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     89
## 2     74 15 13.333 1.975e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis that all $\beta_i = 0$ for $i = 1...15$ is rejected with a p-value equal to 1.975e-15 suggesting that all the variables included are jointly significant.

Omnibus test - Law Enforcement Variables

```
linearHypothesis(model_log_linear_4, c("prbarr = 0", "prbconv = 0", "prbpris = 0",
                                       "avgsen = 0", "polpc = 0")
                                       , vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## prbpris = 0
## avgsen = 0
## polpc = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc
##
## Note: Coefficient covariance matrix supplied.
```

```
##
##   Res.Df Df      F   Pr(>F)
## 1     79
## 2     74  5 4.6194 0.001008 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F test with null hypothesis that the coefficients for all law enforcement variables were all simmultaneusly equal to zero is rejected with an associated p-value equal to 0.001008 showing that these variables are jointly significant in explaining the crime rate.

Omnibus test - Economic and Labor Variables

```
linearHypothesis(model_log_linear_4, c("taxpc = 0",
                                       "wcon = 0", "wtuc = 0", "wtrd = 0",
                                       "wfir = 0", "wser = 0", "wmfg = 0",
                                       "wfed = 0", "wsta = 0", "wloc = 0")
                                       , vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## taxpc = 0
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wfed = 0
## wsta = 0
## wloc = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1     84
## 2     74 10 2.3482 0.01812 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The economic variables were proven to be jointly significantly with a p-value for the associated F test equal to 0.01812.

Including Demographic Variables

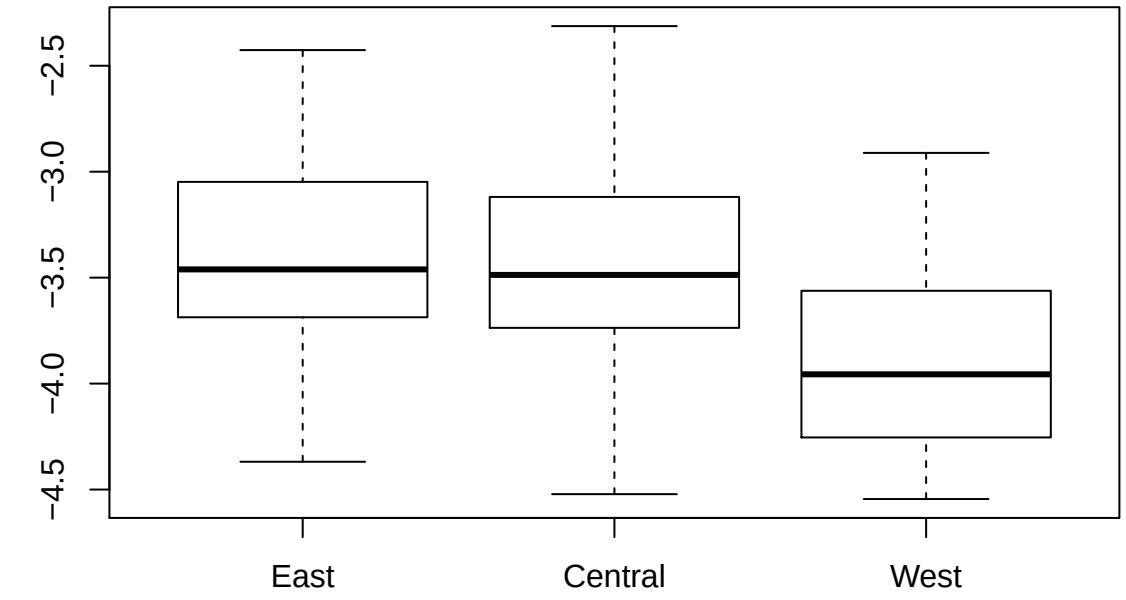
Demographic variables can provide relevant information that can help the policy-makers tailor the crime-reduction policies to the geographic area and population's characteristics. Furthermore, while it is difficult to take action directly on these variables (you can't simply make a rural county urban) these variables provide

important information that can help us gather additional knowledge about the crime in the State of North Carolina.

```
crime_data$cew = 0 # New variable that will display whether the county
#is located in the East (0), Center (1), or West(2) of NC.
crime_data$cew[crime_data$central==1] = 1
crime_data$cew[crime_data$west==1] = 2
table(crime_data$cew)
```

```
##
##  0  1  2
## 35 33 22
```

```
boxplot(log(crime_data$crmrte) ~ crime_data$cew, names = c("East", "Central", "West"))
title("Boxplot of the log of the Crime Rate on the regions of North Carolina")
```



We note that there doesn't seem to be a critical difference in the crime rate between counties in Eastern and Central North Carolina. However, we do see a considerably lower crime rate on Western counties of this State. This will be introduced into the model with a single dummy variable whose base level will be Western counties and will be equal to one on Eastern and Central counties.

```
crime_data$Icew = 0
crime_data$Icew[crime_data$cew==0 | crime_data$cew==1] = 1
crime_data$Icew = as.factor(crime_data$Icew)
```

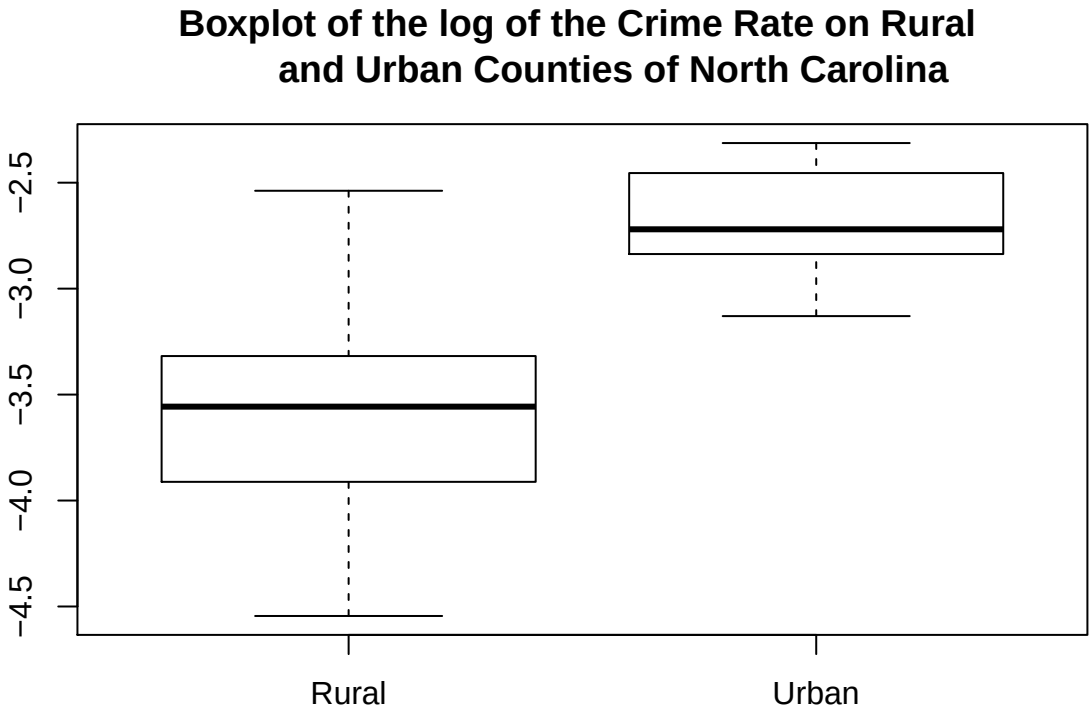
The urban variable could also be very relevant, considering that the dynamics of these types of counties are likely to differ significantly from their rural counterparts.

```
table(crime_data$urban)
```

```
##
##  0  1
## 82  8
```

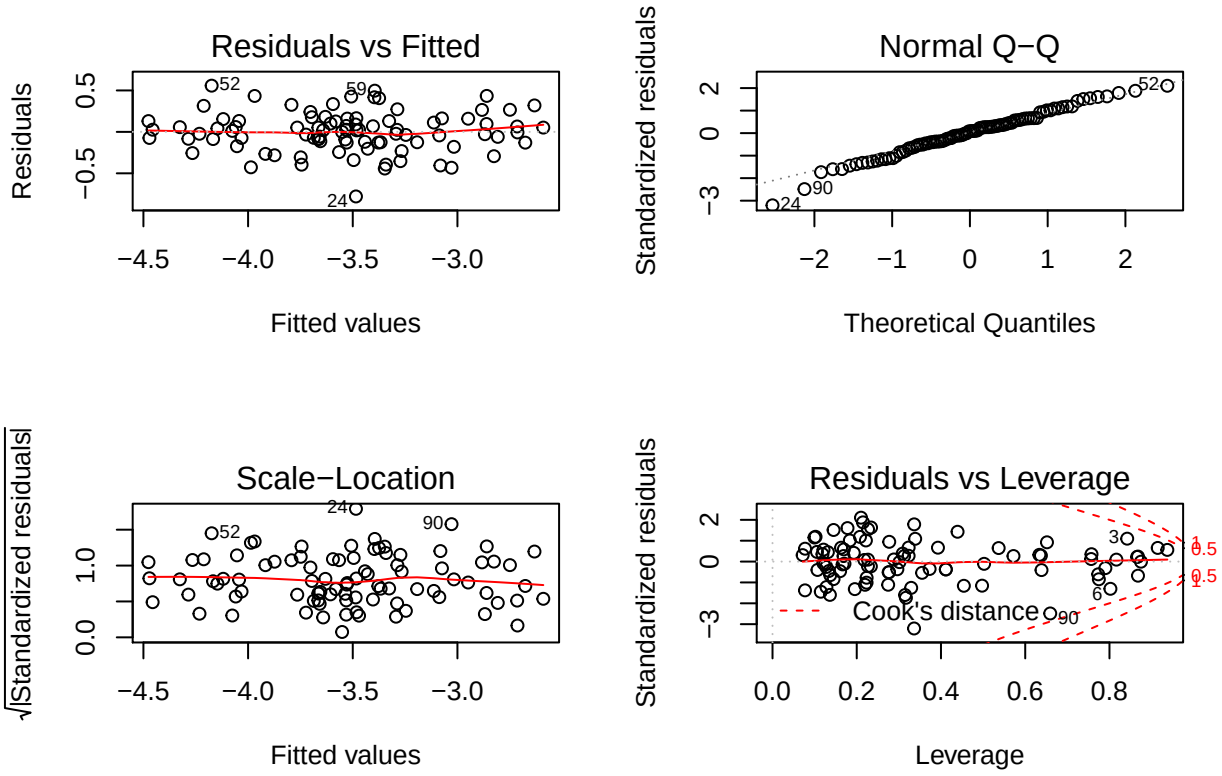
```
boxplot(log(crime_data$crmrte) ~ crime_data$urban,
        names = c("Rural", "Urban"))
```

```
title("Boxplot of the log of the Crime Rate on Rural
      and Urban Counties of North Carolina")
```



While we do note a considerable difference between the crime rates of Rural and Urban counties, with the former having typically lower crime rates, it is important to note that there are only 8 counties defined as urban in the data. With this in mind, it is highly probable that introducing this variable and their interactions could generate singularity issues due to the lack of specific combinations in a database of more than 20 variables and only 90 observations.

```
#Model 4_1: Law-enforcement, economic, and east/west/center variables
model_log_linear_4_1 = lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
            wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc + Icew +
            Icew*prbarr + Icew*prbconv + Icew*prbpris + Icew*avgsen + Icew*polpc +
            Icew*taxpc + Icew*wcon + Icew*wtuc + Icew*wtrd + Icew*wfir + Icew*wser +
            Icew*wmfg + Icew*wfed + Icew*wsta + Icew*wloc,
            data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_4_1)
```




```
coeftest(model_log_linear_4_1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value Pr(>|t|)
## (Intercept)   -5.5797e+00  2.3256e+00 -2.3993  0.01966 *
## prbarr        -1.4856e+00  1.3143e+00 -1.1303  0.26300
## prbconv       -7.8372e-01  4.3854e-01 -1.7871  0.07914 .
## prbpris       -7.5030e-01  1.4269e+00 -0.5258  0.60101
## avgsen         9.7860e-03  6.3956e-02  0.1530  0.87892
## polpc          8.5760e+01  4.0676e+02  0.2108  0.83375
## taxpc          7.6660e-03  1.9342e-02  0.3963  0.69331
## wcon           8.4627e-04  2.6792e-03  0.3159  0.75324
## wtuc          -1.1881e-04  2.0786e-03 -0.0572  0.95462
## wtrd          -1.3118e-03  8.3251e-03 -0.1576  0.87534
## wfir           1.9202e-03  5.0286e-03  0.3819  0.70396
## wser          -1.6403e-04  5.8867e-03 -0.0279  0.97787
## wmfg          -6.1893e-05  1.4003e-03 -0.0442  0.96490
## wfed           4.1913e-03  4.1481e-03  1.0104  0.31649
## wsta          -2.2729e-03  4.4828e-03 -0.5070  0.61406
## wloc           3.0167e-03  6.0112e-03  0.5018  0.61768
## Icew1          2.2352e+00  2.5155e+00  0.8886  0.37789
## prbarr:Icew1  -1.1185e-01  1.3936e+00 -0.0803  0.93631
## prbconv:Icew1 -3.3034e-01  5.9796e-01 -0.5524  0.58277
## prbpris:Icew1  7.9563e-01  1.6574e+00  0.4801  0.63299
## avgsen:Icew1  -3.1591e-02  6.7591e-02 -0.4674  0.64198
## polpc:Icew1    2.4466e+02  4.8337e+02  0.5062  0.61466
## taxpc:Icew1   -9.3215e-03  2.0352e-02 -0.4580  0.64865
```

```
## wcon:Icew1    -6.8949e-04  2.9764e-03 -0.2317  0.81762
## wtuc:Icew1     4.2821e-04  2.3796e-03  0.1800  0.85782
## wtrd:Icew1     2.2465e-03  8.6089e-03  0.2609  0.79506
## wfir:Icew1    -2.6632e-03  5.1789e-03 -0.5142  0.60904
## wser:Icew1    -1.1572e-03  6.0278e-03 -0.1920  0.84843
## wmfg:Icew1     8.4622e-04  1.7055e-03  0.4962  0.62165
## wfed:Icew1    -2.3986e-03  4.4187e-03 -0.5428  0.58933
## wsta:Icew1     2.4607e-03  4.6102e-03  0.5338  0.59555
## wloc:Icew1    -3.8910e-03  6.4968e-03 -0.5989  0.55156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model_log_linear_4_1, c("prbarr:Icew1 = 0", "prbconv:Icew1 = 0", "prbpris:Icew1 = 0",
                                         "avgsen:Icew1 = 0", "polpc:Icew1  = 0" , "taxpc:Icew1 = 0",
                                         "wcon:Icew1 = 0", "wtuc:Icew1 = 0",
                                         "wtrd:Icew1 = 0", "wfir:Icew1 = 0", "wser:Icew1 = 0",
                                         "wmfg:Icew1 = 0", "wfed:Icew1 = 0", "wsta:Icew1 = 0",
                                         "wloc:Icew1 = 0", "Icew1 = 0")
                                         , vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr:Icew1 = 0
## prbconv:Icew1 = 0
## prbpris:Icew1 = 0
## avgsen:Icew1 = 0
## polpc:Icew1 = 0
## taxpc:Icew1 = 0
## wcon:Icew1 = 0
## wtuc:Icew1 = 0
## wtrd:Icew1 = 0
## wfir:Icew1 = 0
## wser:Icew1 = 0
## wmfg:Icew1 = 0
## wfed:Icew1 = 0
## wsta:Icew1 = 0
## wloc:Icew1 = 0
## Icew1 = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     Icew + Icew * prbarr + Icew * prbconv + Icew * prbpris +
##     Icew * avgsen + Icew * polpc + Icew * taxpc + Icew * wcon +
##     Icew * wtuc + Icew * wtrd + Icew * wfir + Icew * wser + Icew *
##     wmfg + Icew * wfed + Icew * wsta + Icew * wloc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     74
## 2     58 16 1.4242 0.1629
```

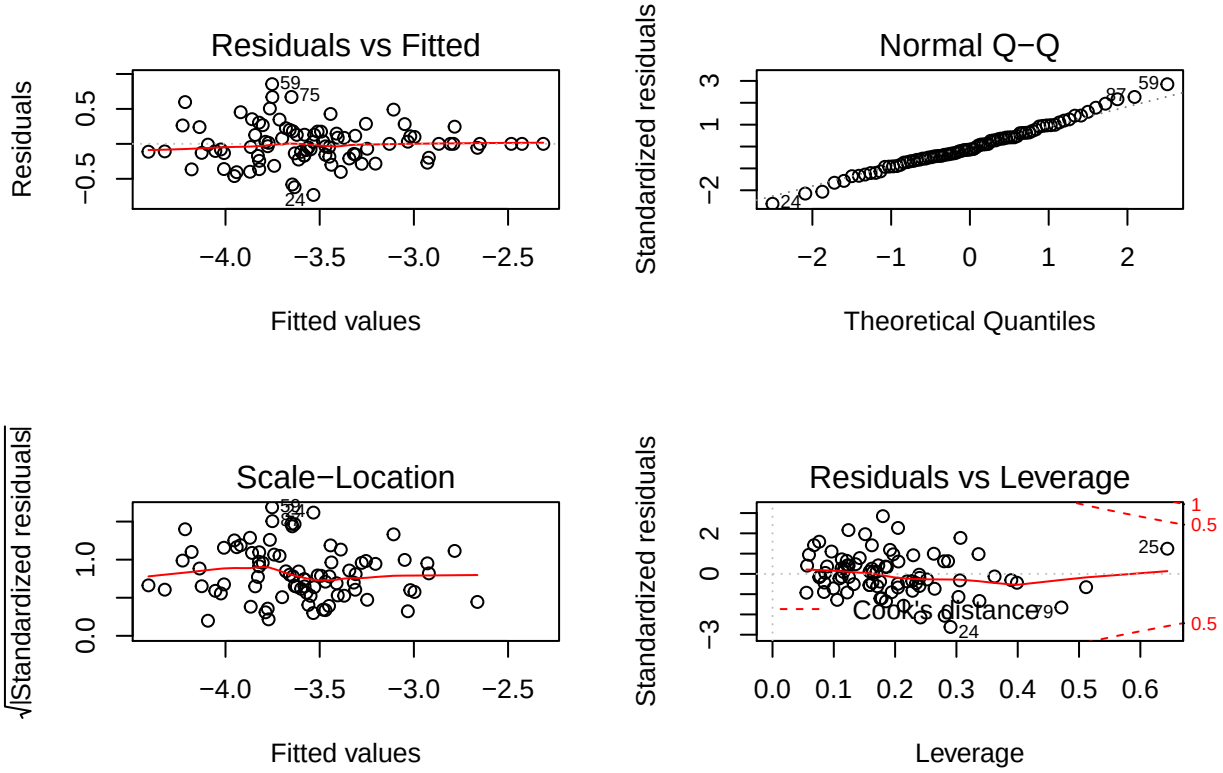
The p-value associated to the null hypothesis of beta-coefficients equal to zero for all location variables leads

us to not reject it and remove these variables from the model as they are not jointly significant.

```
#Model 4_2: Law-enforcement, economic, and urban/rural variables
model_log_linear_4_2 = lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
            wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc + urban +
            urban*prbarr + urban*prbconv + urban*prbpris + urban*avgsen + urban*polpc +
            urban*taxpc + urban*wcon + urban*wtuc + urban*wtrd + urban*wfir + urban*wser +
            urban*wmfg + urban*wfed + urban*wsta + urban*wloc,
            data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_4_2)
```

```
## Warning: not plotting observations with leverage one:
##   11, 23, 29, 31, 36, 52, 56, 82
```

```
## Warning: not plotting observations with leverage one:
##   11, 23, 29, 31, 36, 52, 56, 82
```




```
crime_data[crime_data$urban == 1,c(1,13)]
```

```
##    county urban
## 11     21     1
## 23     51     1
## 29     63     1
## 31     67     1
## 36     81     1
## 53    119     1
## 57    129     1
## 83    183     1
```

As suspected before, the low incidence of cases with urban indicator variable leads to singularities in the

model and won't be pursued further. However, the difference in crime rates between these two types of counties is still worth investigating further in subsequent studies.

Model Coefficients

Using a backwards variable selection, the log-linear model will be refined by iteratively dropping the variable with the largest p-value on the significance t-test ($H_0 : \beta_i = 0$).

```
coeftest(model_log_linear_4, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4.2815e+00  8.7283e-01 -4.9053 5.381e-06 ***
## prbarr      -1.9107e+00  5.5604e-01 -3.4363 0.0009707 ***
## prbconv     -9.4896e-01  2.4660e-01 -3.8482 0.0002504 ***
## prbpris     -6.6998e-02  6.0325e-01 -0.1111 0.9118687
## avgsen      -9.9141e-03  1.9033e-02 -0.5209 0.6039972
## polpc        1.4931e+02  1.2538e+02  1.1909 0.2375125
## taxpc        5.1736e-03  4.7027e-03  1.1001 0.2748402
## wcon         4.7709e-04  9.6304e-04  0.4954 0.6217837
## wtuc        -8.5298e-05  7.6343e-04 -0.1117 0.9113396
## wtrd         1.9579e-04  1.9472e-03  0.1005 0.9201822
## wfir        -6.8224e-04  1.0224e-03 -0.6673 0.5066473
## wser        -2.4152e-03  9.7800e-04 -2.4695 0.0158376 *
## wmfg         1.0451e-04  5.2962e-04  0.1973 0.8441044
## wfed         4.0554e-03  1.3219e-03  3.0680 0.0030105 **
## wsta        -5.6846e-04  9.9878e-04 -0.5692 0.5709766
## wloc         1.6992e-03  2.4585e-03  0.6912 0.4916284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, wtrd was dropped with an associated p-value equal to 0.9201822. Afterwards, prbpris is dropped due to a p-value equal to 0.9177290. Then wtuc is dropped since the p-value of the t-test $H_0 : \beta_{wtuc} = 0$ had a p-value of 0.8986431. The weekly wage for manufacturing jobs (wmfg) is also dropped due to its p-value of 0.8788609. Subsequently, wcon is removed from the model with a t-test p-value equal to 0.6056468. The variable wfir is then dropped with a p-value of 0.5325245. Later, the average sentence measured in days (avsen) is removed due to having the largest p-value of the remaining variables which equaled 0.4751385. Wloc is also removed since it had a t-test p-value of 0.360342. Next, wsta is dropped from the regression with a p-value of 0.3458571. Of the remaining variables, polpc is dropped with a p-value equal to 0.1648292 since it was larger than our pre-defined $\alpha = 10\%$. A F-test to assess the joint significance of the variables dropped in the backwards variable selection confirms our findings with an associated p-value of $H_0 : \beta_{prbpris} = \beta_{avgsen} = \beta_{polpc} = \beta_{wcon} = \beta_{wtuc} = \beta_{wtrd} = \beta_{wfir} = \beta_{wmfg} = \beta_{wsta} = \beta_{wloc} = 0$ equal to 0.9691.

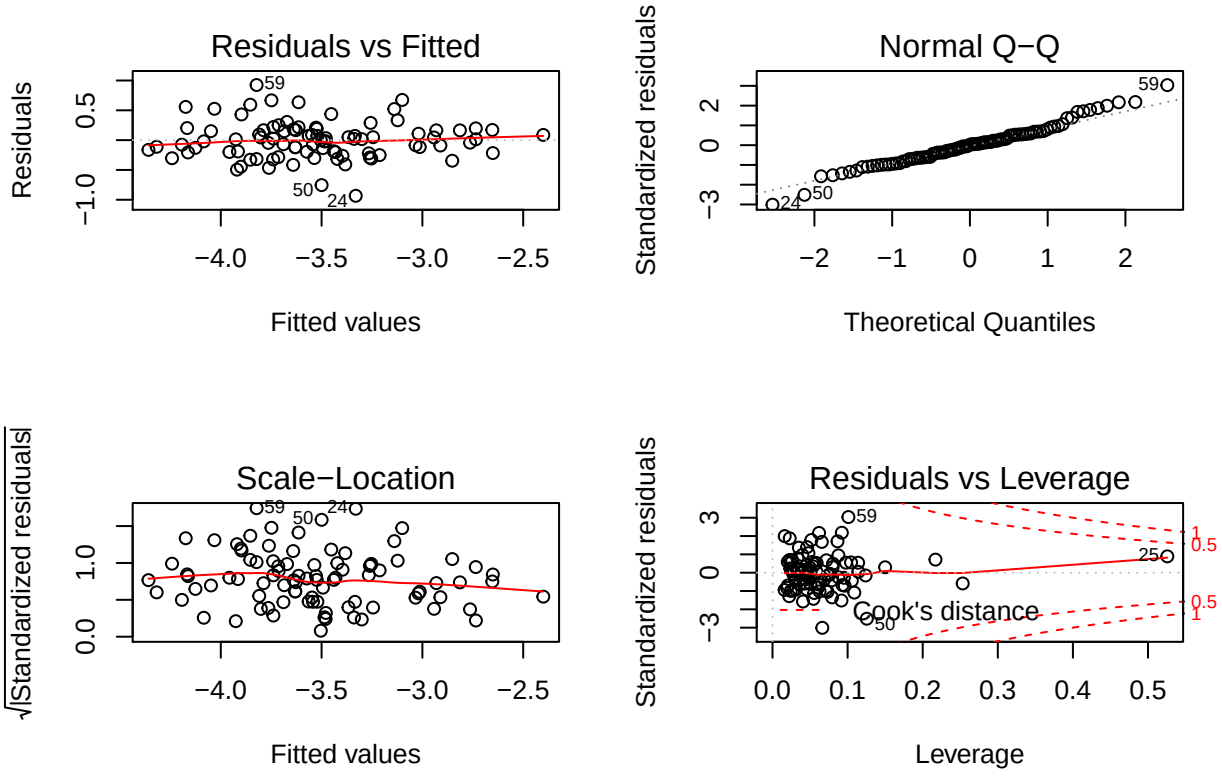
```
linearHypothesis(model_log_linear_4, c("prbpris = 0",
                                       "avgsen = 0", "polpc = 0",
                                       "wcon = 0", "wtuc = 0", "wtrd = 0",
                                       "wfir = 0", "wmfg = 0",
                                      "wsta = 0", "wloc = 0")
                                       , vcov = vcovHC)
```

```
## Linear hypothesis test
##
```

```
## Hypothesis:
## prbpris = 0
## avgsen = 0
## polpc = 0
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wmfg = 0
## wsta = 0
## wloc = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     84
## 2     74 10 0.3339 0.9691
```

Therefore, our final model is:

```
model_log_linear_4_final = lm(log(crmrte) ~ prbarr + prbconv + taxpc +
              wser + wfed,
            data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_4_final)
```



```r
coeftest(model_log_linear_4_final, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4.28176502  0.58964176 -7.2616 1.782e-10 ***
## prbarr      -1.89628726  0.44255022 -4.2849 4.843e-05 ***
## prbconv     -0.99537504  0.20613632 -4.8287 6.097e-06 ***
## taxpc        0.00941244  0.00319757  2.9436  0.004194 **
## wser        -0.00196457  0.00081242 -2.4182  0.017760 *
## wfed         0.00438807  0.00097881  4.4831 2.311e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

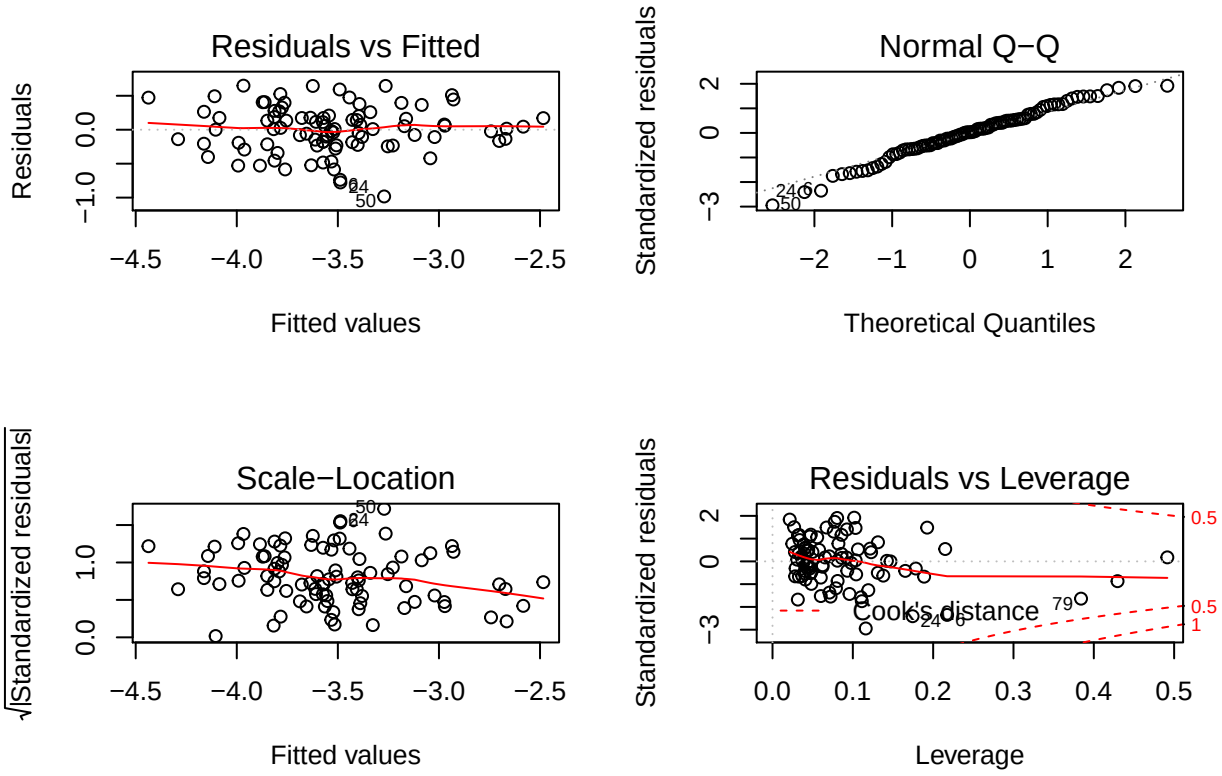
Starting from the previous discussion about the assumptions of the economic and law enforcement model, we consider assumptions MLR1, MLR2, and MLR3 met (taking into consideration the possible pressence of spatial autocorrelation) for this model. Furthermore, we don't observe any critical violations of the Zero-Conditional Mean assumption as the Residuals vs. Fitted values plot shows no relationship between the axes and has a flat adjusted spline line at zero. While heteroskedasticity is not suspected given the Scale-Location plot, we will use White's Heteroscedasticity-consistent standard errors. Finally, based on the QQ plot and asymptotic results provided by a sample size equal to 90 observations (that aren't highly skewed), we assume the normality. As the refined model meets all the assumptions of the Classical Linear Model, we determine that our regression has consistent, unbiased, and normally distributed parameters.

Alternate specification

During the model selection we noted that the weekly wage variables had a considerable amount of correlation between them. Therefore, an alternative model specification is obtained by summarizing the weekly labor variables through a summary statistic robust to deviations such as the median.

```r
for (i in 1:length(crime_data[,1])){
  crime_data$wage_median[i] = median(c(crime_data$wcon[i], crime_data$wtuc[i], crime_data$wtrd[i],
                                crime_data$wfir[i], crime_data$wser[i], crime_data$wmfg[i],
                                crime_data$wfed[i], crime_data$wsta[i], crime_data$wloc[i]))
}

#Model 5: Full model of law-enforcement and economic variables summarized by the median weekly wage
model_log_linear_5 = lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + taxpc +
            wage_median,
            data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_5)
```

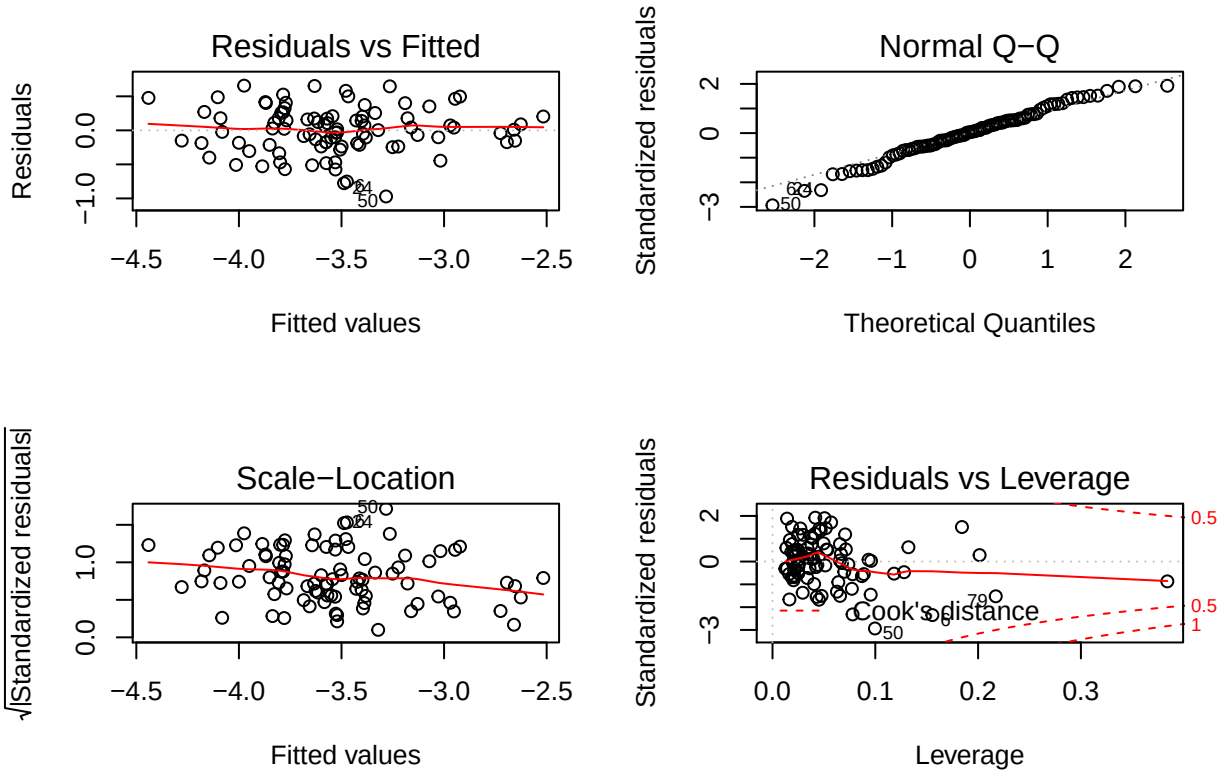



```
coeftest(model_log_linear_5, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4.1515e+00  7.2632e-01 -5.7158 1.710e-07 ***
## prbarr      -1.7142e+00  5.8096e-01 -2.9507  0.004132 **
## prbconv     -1.0013e+00  2.3017e-01 -4.3503 3.886e-05 ***
## prbpris      5.0914e-02  6.4705e-01  0.0787  0.937473
## avgsen      -2.1804e-03  1.7148e-02 -0.1271  0.899136
## polpc        1.9431e+02  1.2089e+02  1.6073  0.111830
## taxpc        9.3446e-04  3.6794e-03  0.2540  0.800151
## wage_median  4.0253e-03  1.4242e-03  2.8264  0.005912 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the same backwards variable selection, the final alternative model is obtained after removing prbpris (p-value=0.937473), avgsen (p-value=0.884356), and taxpc (p-value=0.797363). This leads us to the following final alternative model. One important caveat of this alternative method is that the median is calculated based on the salaries of the reported industries and is not necessarily the median wage in the county.

```
model_log_linear_5_final = lm(log(crmrte) ~ prbarr + prbconv + polpc +
            wage_median,
            data = crime_data)
par(mfrow=c(2,2))
plot(model_log_linear_5_final)
```




```
coeftest(model_log_linear_5_final, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -4.1347058   0.6140620 -6.7334 1.845e-09 ***
## prbarr       -1.7179948   0.5585538 -3.0758  0.002824 **
## prbconv      -1.0108910   0.2147549 -4.7072 9.663e-06 ***
## polpc       200.8021174  80.0009047  2.5100  0.013969 *
## wage_median   0.0040707   0.0013863  2.9364  0.004271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A quick review of the CLM assumptions shows that the model has Zero-Conditional Mean as shown by the spline line in the Residuals vs. Fitted values. Furthermore, while this model has some minor issues with heteroskedasticity as seen in the Scale-Location plot, heteroscedasticity-consistent standard errors will be used. Finally, we see that the normality assumption is met in the QQ plot (in addition to being granted by the asymptotical results given by the large sample size). Therefore, the alternative model also meets CLM assumptions and has consistent and unbiased parameters.

Regression Table

```
model_log_linear_full = lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc + density
                + taxpc + west + central + urban + pctmin80 + wcon + wtuc + wtrd + wfir + wser +
                  wmfg + wfed + wsta + wloc + mix + pctymle, data = crime_data)
model_log_linear_1 = lm(log(crmrte) ~ prbarr + prbconv + prbpris + polpc + avgsen, data = crime_data)
model_log_linear_3 = lm(log(crmrte) ~ taxpc + wcon + wtuc + wtrd + wfir +
```

```
                                    wser + wmfg + wfed + wsta + wloc,
                        data = crime_data)
#Models 1 and 3 are re-run to allocate for the changes in the data frame

(se.model1 = sqrt(diag(vcovHC(model_log_linear_1))))
```

(Intercept) prbarr prbconv prbpris polpc avgsen 0.5537213 0.6334766 0.2544122 0.7840402 114.3755338 0.0201737

```
(se.model3 = sqrt(diag(vcovHC(model_log_linear_3))))
```

(Intercept) taxpc wcon wtuc wtrd 0.6931966286 0.0051299948 0.0013239145 0.0007629119 0.0028644762 wfir wser wmfg wfed wsta 0.0014419875 0.0014479993 0.0007511258 0.0016130119 0.0011089075 wloc 0.0027262716

```
(se.model4 = sqrt(diag(vcovHC(model_log_linear_4_final))))
```

(Intercept) prbarr prbconv taxpc wser 0.5896417595 0.4425502154 0.2061363236 0.0031975722 0.0008124169 wfed 0.0009788083

```
(se.model5 = sqrt(diag(vcovHC(model_log_linear_5_final))))
```

(Intercept) prbarr prbconv polpc wage_median 0.614061979 0.558553824 0.214754914 80.000904680 0.001386289

```
(se.modelfull = sqrt(diag(vcovHC(model_log_linear_full))))
```

(Intercept) prbarr prbconv prbpris avgsen 8.186671e-01 3.646345e-01 2.674478e-01 4.643591e-01 1.470197e-02 polpc density taxpc west1 central1 1.512400e+02 5.535317e-02 7.624428e-03 1.426004e-01 1.073256e-01 urban1 pctmin80 wcon wtuc wtrd 2.240604e-01 3.513221e-03 9.097380e-04 7.214035e-04 1.880464e-03 wfir wser wmfg wfed wsta 1.077387e-03 1.283119e-03 4.446051e-04 1.098635e-03 9.774068e-04 wloc mix pctymle 2.109802e-03 5.868459e-01 1.293026e+00

```
stargazer(model_log_linear_1, model_log_linear_3,
          model_log_linear_4_final, model_log_linear_5_final,
          report = "vcs", type = "latex",
          omit.table.layout = "n", font.size="small",
          title = "Log-Linear Models Predicting Crime Rate", omit.stat = "f",
          se = list(se.model1, se.model3, se.model4, se.model5),
          star.cutoffs = c(0.05, 0.01, 0.001), float = F,
          add.lines=list(c("AIC", round(AIC(model_log_linear_1),1),
                           round(AIC(model_log_linear_3),1),
                           round(AIC(model_log_linear_4_final),1),
                           round(AIC(model_log_linear_5_final)),1)))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 15, 2018 - 13:15:05

	Dependent variable:			
	log(crmrte)			
	(1)	(2)	(3)	(4)
prbarr	−2.240		−1.896	−1.718
	(0.633)		(0.443)	(0.559)
prbconv	−1.057		−0.995	−1.011
	(0.254)		(0.206)	(0.215)
prbpris	0.207			
	(0.784)			
polpc	256.449			200.802
	(114.376)			(80.001)
avgsen	−0.001			
	(0.020)			
taxpc		0.013	0.009	
		(0.005)	(0.003)	
wcon		0.0003		
		(0.001)		
wtuc		0.00000		
		(0.001)		
wtrd		0.002		
		(0.003)		
wfir		−0.002		
		(0.001)		
wser		−0.002	−0.002	
		(0.001)	(0.001)	
wmfg		0.0005		
		(0.001)		
wfed		0.005	0.004	
		(0.002)	(0.001)	
wsta		0.001		
		(0.001)		
wloc		0.002		
		(0.003)		
wage_median				0.004
				(0.001)
Constant	−2.848	−6.689	−4.282	−4.135
	(0.554)	(0.693)	(0.590)	(0.614)
AIC	88	107.5	58.7	73
Observations	90	90	90	90
R^2	0.502	0.446	0.641	0.571
Adjusted R^2	0.472	0.376	0.619	0.551
Residual Std. Error	0.378 (df = 84)	0.411 (df = 79)	0.321 (df = 84)	0.349 (df = 85)

Model Interpretation

Our final model (model 4: model_log_linear_4) relfects that both economic metrics mainly measured through wages and law enforcement variables were both significant determinants of the crime rate in North Carolina in 1987. Most importantly from a public policy perspective, we see a negative relationship between the crime rate and the probability of arrest and conviction. Namely, an increase of 0.1 in the probability of arrest (prbarr) will lead to a decrease of 18.96% in the crime rate. In addition, an increase of 0.1 in the probability of conviction (prbconv) will reduce the crime rate in 9.95%. On the other hand, we see that both the tax revenue per capita (taxpc) and the weekly wage for federam employees have a positive relationship with the crime rate. More specifically, a 1 dollar increase in the tax revenue leads to an expected increase of 0.9% in the crime rate while a 1 dollar increase in the weekly wage for federal workers leads, in average, to a 0.4% increase in the crime rate. While this might be counter-intuitive at first, this could be caused by the fact that richer counties have more and better payed federal employees leading to larger inequalties in the population which could be increasing the crime rate (it is important to mention that the weekly wages for federal employees were usually the largest one of all the wages in the dataset). Also, the relationship between the weekly federal wage and the increased crime rate could be caused by a confounding effect where counties with larger crime problems receive more help from federal law-enforcement agencies. Moreover, we note that there is a negative relationship between the weekly wage of the service industry and the crime rate, where a 1 dollar increase in this wage leads to a reduction of 0.2% in the crime rate. Finally, it is important to mention that this model can explain roughly two-thirds of the variance in the log of the crime rate ($R^2 = 0.641$) and has the best fit in terms of AIC and $R^2_{adjusted}$ values with 58.7 and 0.619 respectively.

Omitted Variable Bias

While the model controls for a wide selection of variables and has no clear indication of violations of the zero-conditional mean assumption, there are still several additional variables that could be included in the model as they are very likely in the crime rate population model, they are:

- Average years of education

This socio-economic factor can introduce omitted variable bias in our co-efficients. There is no direct proxy for this variable that might stand in for it. We might expect certain types of jobs/industries that are associated with higher years of education to have more employment, and it is expected that higher years of education would drive wages higher. It is expected that the coefficient of average years of education will be negatively related to crime rate (higher the years of education, lower the crime rate). Also, wages are positively correlated with years of education under our assumption but we expect sign of coefficient of years of education to be negative. Hence it would result in negative bias.

- Unemployment rate in the county

This omitted variable is an important economic factor and it has possibly a high positive relationship with the dependent variable (crime rate). An increase in unemployment would result in an increase in crime rate. Also there is no direct proxy for this variable in the data. The wage variables do not effectively measure the rate of employment and hence can't cater to the effect of this variable. We can expect the tax per capita to be correlated with unemployment. Higher tax per capita would have a negative correlation with unemployment (as unemployment decreases, more taxes would be incurred). So with these considerations in mind, it can be said that the bias due to omitted variable of unemployment in tax per capita would be **negative** –> corr(unemp, taxpc) < 0 co-efficient of unemp with respect to crime rate is assumed to be positive, hence bias in co-efficient of taxpc is negative.

- Level of poverty (proportion of the population under the line of poverty)

This socio-economic variable is similar to unemployment but effectively comes from a different angle of economic conditions of an area. We expect co-efficient of level of poverty to be positively related to crime rate. There are several proxies for this variable. Number of arrests might be one that can give a hint about

the level of poverty (higher number of arrests mean higher level of poverty), however it is not a perfect proxy. We can expect correlation between the two. We expect a positive correlation between the these two (probability of arrest and level of poverty) and we also expect level of poverty to be positively related to crime rate. Hence this omitted variable will incur positive bias.

- Drug usage in the county

This social factor is expected to have positive relation with our dependent variable. Higher drug usage in the county would point towards higher crime rate. There is no direct proxy for this variable, however it is expected to be correlated to probability of arrests and probability of convictions. We can expect a positive bias due to this omitted variable as it is positively correlated with our (under consideration) independent and dependent variable.

- Percent of recurrent crime (percent of crime from recurrent criminals). Also known as recividism rate.

This is a very important factor which might have a proxy in probability of convictions. Higher convictions would mean that there are higher numbers of convicted criminals and higher tendency for repeat criminals. While this may not be a perfect proxy, it is a slightly overlapping variable and we might expect positive correlation between the two. And recividism may be positively related to crime rate (higher recurrent criminals will drive crime rate high). Thus we expect the bias due to this omitted variable to be positive.

- High school completion rate in the county
- Past crime rate
- Economic growth
- Average debt in the county
- GDP per capita in the county
- Inequality

Conclusions

The above discussion including EDA, model building and omitted variables show that it is a highly complicated task predicting crime rate in a certain area, and there are a number of socio-economic and law enforcement factors involved that influence the predictive power of this variable.

However a general trend is that the most effective and actionable variables at reducing the crime rate are a high probability of arrest and conviction. Therfore in order to prevent future crime, it is recommended to channel sufficient resources to improve the criminal system to make sure that once a crime is committed the perpetrator is arrested at high rates and that once that they are arrested they are actually convicted. This means that work needs to be done to make police enforcement agencies better at capturing criminals and the prosecution better at convicting guilty criminals. An additional economic variable that can help control the crime rate are the wages, in particular for the service industry. Therefore, a better criminal system together with economic factors for the population such as better living wages for residents are likely to lead to reduced crime rates.