Sample R Analysis

James Kajdasz, Yueh-Wen Wang, Aidan Feay

8/4/2018

Introduction

This demo of an R analysis proposes that a politician on the political campaign trail is running on the promise of putting more police on the streets to lower crime. We have been hired by a rival politician to examine the validity of that strategy. Our goal in this lab is determine how hiring more police would impact crime rate. Enlarging the police force is an action politicians can do relatively quickly, making it an attractive campaign platform and important action to evaluate properly. It's anticipated that this report can help inform future policy initiatives and election campaign platforms with regards to crime.

Our analysis concludes with an identification of counties that are considered over-policed or under-policed based on conditions of population density, percent of young males in the community, relative wealth of the community, and crime rate.

The Initial Data Loading and Cleaning

Loading the Data

```r
library('car')
```

```
## Loading required package: carData
```

```r
library('stargazer')
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library('lmtest')
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library('sandwich')
# setwd('~/Documents/Berkeley/W203/Labs/w203_lab3') # Aidan's WD
# setwd('D:/Teresa/Berkeley/Semester2/W203 Statistics')
# setwd('D:/programs/dropbox/education/berkeley/w203stats/lab3')  # This is Jim's working directory
getwd()
```

```
## [1] "D:/Programs/Dropbox/Education/Berkeley/W203Stats/CrimeData"
```

```r
crime <- read.csv('crime_v2.csv')
```

Studying and Cleaning

To examine our research question, we consulted data from Cornwell and Trumball, who drew together many sources from FBI Uniform Crime Reports, North Carolina's Department of Corrections files, census data, and the North Carolina Employment Security Commission. The data is not current. The original report by Cornwell and Trumball was published in 1994, and data for that report is even older. We began our examination by inspecting the data variables and sample size. Each data observation represents a county in North Carolina, with values on 25 different variables ($n = 97$).

```r
summary(crime)
```

```
##      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        
```

```r
#head(crime)
nrow(crime)
```

```
## [1] 97
```

The variable ‘probability of arrest’ (prbarr) is a numeric probability but is coded as a factor with 92 levels. We changed the data type to numeric.

```r
crime$prbconv <-as.numeric(levels(crime$prbconv)[crime$prbconv])
```

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

There are 100 counties in North Carolina but only 97 represented in our dataset. Six rows have null values for every field. We ommited these empty rows from our analysis, bringing our total sample size to 91.

```r
#navals <- crime[is.na(crime$county),]
#navals
cleancrime <- crime[complete.cases(crime),]
```

Examining each variable, we note ambiguity in how region is coded. Counties are categorized within a region of the state using the variables ‘west’ and ‘central’ (Western NC: (1,0), Central NC: (0,1)). There is a third category (coded 0,0) but it is not apparent what region this refers to. Thirty-three counties are coded (0,0). It’s likely this code represents Eastern NC or Coastal NC, but this is an assumption and cannot be confirmed with the existing data. In addition, county 71 is coded (1,1). The meaning of this coding is not certain.

We note some confusion about the variable Density, which is reported as ‘people per square mile’ in the codebook. The highest population density reported is 8.82 people per square mile. This means an urban area like Mecklenburg (which contains the city of Charlotte and has a land area 546 square miles) would have a population of only 4,815. It seems more likely that this scale is in hundreds or thousands of people per square mile. For now we note the issue but don’t make any changes to the data.

Next we examined individual values to identify possible outlier data. Some counties contained values that are of suspect validity.

County 185 reports an average weekly service industry wage of $2,177. This number is abnormnally high. The next highest service industry wage is $391. Mean service industry wage is $275 (SD = $206). The service industry wage in county 185 is 9.23 SD above the mean. While suspect, we did not remove outliers at this stage of the anlaysis.

```r
#summary(cleancrime$wser)
#suspect <- subset(cleancrime, cleancrime$wser >= 2177)
#suspect
```

```r
#cleancrime$wser
#hist(cleancrime$wser)
```

County 55 reports tax income per capita of 119.76. This is unusually high. It is 6.25 SD above the mean (M = 38.06, SD = 13.08). The reported value seems even more suspicious when examining wage data. The majority of wages in county 55 fall below the mean for the state, with only government workers earning near the average. For now we note the unusual value but do not change the data.

```r
#summary(cleancrime$taxpc)
#suspect3 <- subset(cleancrime, cleancrime$taxpc >= 119.7)
#suspect3
#hist(cleancrime$taxpc)
#cleancrime[25,'taxpc']
```

County 133 reports an unusually high percentage of young males (24.9%). This value is 7.06 standard deviations above the mean (M = 8.4%, SD = 2.3). The next highest reported value in a county is 15.1%. For now we note the suspect value, but do not make any changes.

```r
#summary(cleancrime$pctymle)
#suspect4 <- subset(cleancrime, cleancrime$pctymle >= .2487)
#suspect4
#hist(cleancrime$pctymle)
#cleancrime[59, 'pctymle']
```

Counties 3, 19, 99, 115, 127, 137, 149, 185, 195, and 197 all have a Probability of Conviction (prbconv) that is greater than 1, which is not possible for a probability. Such a large number of aberrations in a single variable make us cautious of using this variable for any analysis.

```r
#summary(cleancrime$prbconv)
#suspect5 <- subset(cleancrime, cleancrime$prbconv > 1.00)
#suspect5
```

County 115 reports a Probability of Arrest that is greater than 1 (1.09), which is not possible for a probability.

```r
#summary(cleancrime$prbarr)
#suspect6 <- subset(cleancrime, cleancrime$prbarr >= 1.00)
#suspect6
#hist(cleancrime$prbarr)
#cleancrime[51, 'prbarr']
```

County 115 is a particularly interesting data point. Of all the counties, it has the highest police per capita (.009), the highest probability of arrest (1.09), the highest average sentence (20.7), and the lowest crime rate (.005). It's possible the county might exert undo leverage in any model created. For now we note the unusual county but do not change the data.

```r
subset(cleancrime, cleancrime$county == 115)
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     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
```

The Model Building Process

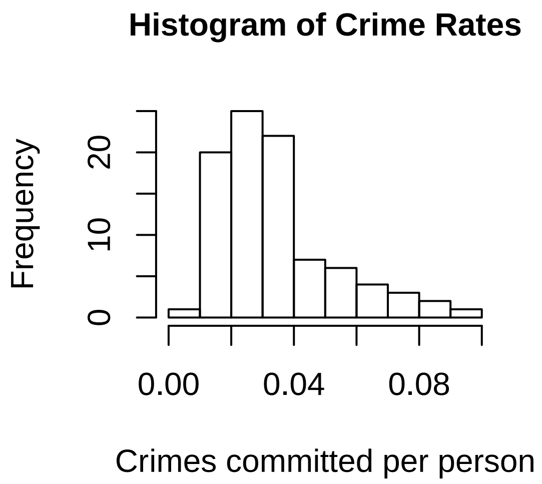
Next we identify our primary variables of interest, potential covariates, and evalulate the need for transformations of the data.

Crime Rate

Our research question is to evaluate the effect on crime rates, and so we identify the variable 'crimes commited per person' (crmrte) as our criterion variable of interest. Other variables such as offense type (mix) and average sentence (avgsen) might also be interesting outcome variables if we wanted to consider the general severity of crimes committed, but our research question does not make such a specification so the broader metric of crime rate is deemed to be most useful. We begin our examination of our selected criterion variable: crimes commited per person.

Crimes committed per person (variable crmrte) ranges from 0.005 to .099 ($M = .033, SD = 0.02$). There are no unusual values noted. The distribution is positively skewed. Tests of normality indicate a non-normal distribution ($W = 0.89, p < .001$). Our analysis might benefit by taking the log of crime rate so we can examine percent changes in crime, a metric that is more readily understood than changes in crimes committed per person. However, in this case it is not advisable. Crime rate ranges from zero to one, with many values near zero. Taking the log of such a variable can create extreme values that do not reflect the original scaling (Wooldridge, p.172). For this reason, we choose to not transform the crime rate variable.

```
hist(cleancrime$crmrte, main='Histogram of Crime Rates', cex.main = 1,
  xlab='Crimes committed per person')
```



```
# mean(cleancrime$crmrte)
# sd(cleancrime$crmrte)
```

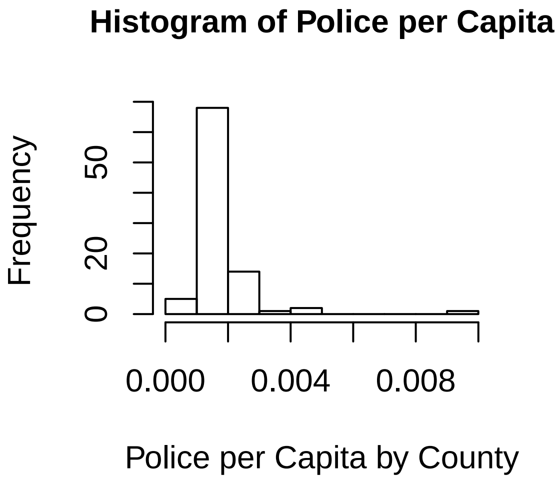
```
#qqnorm(cleancrime$crmrte)
shapiro.test(cleancrime$crmrte)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cleancrime$crmrte
## W = 0.88954, p-value = 1.269e-06
```

Police Per Capita

The other variable of primary interest as dictated by our research question is police per capita (polpc). The variable ranges from 0.00075 to 0.00905 ($M = 0.0017, SD = .00098$). There is one outlier: county 155 reports a police per capita presence 0.009, which is 7.45 SD above the mean. The distribution is positively skewed and significantly strays from normality ($W = 0.88954, p < .001$). We choose to not conduct a log transformation for the same reasons we did not transform the variable 'crimes comitted per person' (i.e. values between 0 and 1 with many values near 0).

```
hist(cleancrime$polpc, main='Histogram of Police per Capita',
     cex.main = 1, xlab='Police per Capita by County')
```



```
#qqnorm(cleancrime$polpc)
#summary(cleancrime$polpc)
#mean(cleancrime$polpc)
#sd(cleancrime$polpc)
#shapiro.test(cleancrime$crmrte)
```

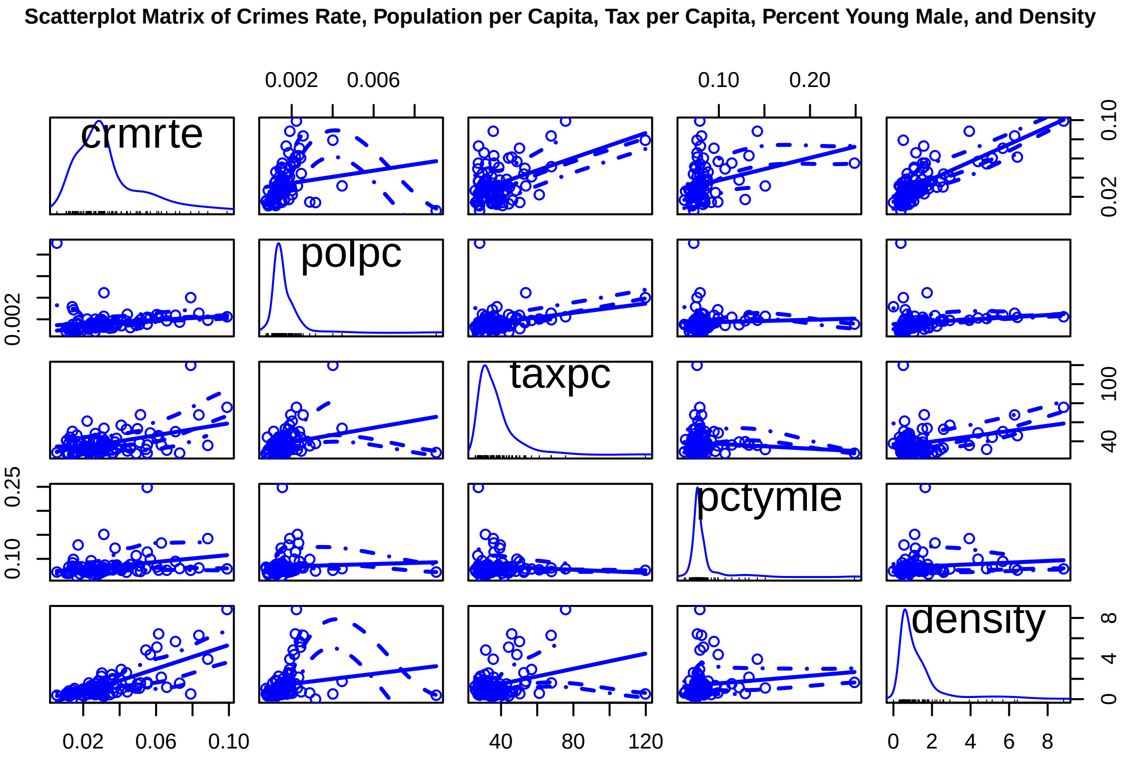
Beyond our primary variables of interest, We want to identify additional variables affecting crime rate. Identifying additional variables helps us understand crime as a phenomenon better. In addition, by accounting for the affects of other variables, we can better isolate the contribution of police presence by itself.

We look to theory to suggest which variables have a connection to crime. We examine population density (density), as dense urban areas are generally associated with more crime than rural areas. For the same reason, we also examine the dummy coded 'urban' variable that identifies counties that are largely metropolitan. We examine the percent of young males in a particular county (pctymle), as young males represent a disproportionate share of those committing crimes. Crime is often associated with poorer socieconomic areas. We therefore examine the measure of taxes collected per capita (taxpc), which serves as a proxy variable for relative wealth of a county.

Additional variables might also arguably be important. Weekly wage for various industries is also an indicator of economic wealth and opportunity of a region. The various wages do not give an indication of proportion of a particular industry in a particular county however. Unable to know how important a particular industry was to a particular county, we chose to omit examination. Poor socioeconomic areas often have higher percentages of minortiy residents. It might be tempting to then examine the percentage of minority in a county as a variable of interest (pctmin80). This data was collected in 1980, and has a considerable temporal gap before other measures were collected circa 1987. For this reason, we chose to exclude this variable from analysis.

As a first step to check our intuitions, we create scatter plots to compare crime rate with other non-categorical variables.

```r
spm(~crmrte + polpc + taxpc + pctymle + density, data=cleancrime, main="Scatterplot Matrix of Crimes Ra
```



We note in the first column weak to moderate positive correlations between crime rate and our other variables of interest. The positive correlation between police per capita and crime rate is noteworthy, as it would seem to go against our initial hypothesis that higher police presence will reduce crime. We examine each of our predictor variables in greater detail.

Density

Population density (density) ranges from 0.00002 to 8.83 people (unit uncertain) per square mile, ($M = 1.43, SD = 1.51$). Though the variable is described as people per square mile, a populous county like Mecklenburg covers 546 square miles and has a much higher population than 4,821 (max density * area). In 1994, the year of our study, Mecklenburg County had a population of 576,276: over 100 times greater than the density variable indicates assuming that our field refers to singular digits of people. It would be reasonable to suggest that the variable is actually hundreds of people per square mile (or even thousands). The distribution is positively skewed and not normal ($W = 0.67, p < .001$). With many values near zero, it is inadvisable to consider a log transformation. Density has a large positive correlation with crime rate ($r = 0.73$). The variable should be useful in our analysis.

```r
#hist(cleancrime$density, main='Histogram of Population Density', xlab='People per Sq Mile')
#qqnorm(cleancrime$density)
#summary(cleancrime$density)
#mean(cleancrime$density)
#sd(cleancrime$density)
```

```
#shapiro.test(cleancrime$density)
#cor(cleancrime$crmrte, cleancrime$density)
```

Urban

Each county is coded by the 'Urban' variable to denote it as either a Standard Metropolitan Statistical Area or not. It takes values of either 0 or 1. It has a mean value of 0.088, meaning that most counties are not urban. Crime is significantly higher ($t(7.9) = 6.05, p < 0.001$) in counties coded as urban ($m = .07$) versus non-urban ($m = .03$). This variable should be an important predictor of crime, although it reflects much the same information as the previous variable 'density'. Furthermore, as 'density' is a continuous variable and 'urban' is a dichotomous one, there is more information conveyed in the 'density' variable than the 'urban' variable.

```
#summary(cleancrime$urban)
#mean(cleancrime$density)
t.test(crmrte ~ urban, data=cleancrime)
```

```
##
##  Welch Two Sample t-test
##
## data:  crmrte by urban
## t = -6.0523, df = 7.8626, p-value = 0.0003269
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05621223 -0.02512652
## sample estimates:
## mean in group 0 mean in group 1
##      0.02982490      0.07049427
```

Percent Young Male

According to the Bureau of Justice Statistics, young people (especially men) have the highest rates of violent crime. The percentage of young men in a given county (pctymle) ranges from 0.06 to 0.25, or 6% to 25%, ($M = 0.078, SD = 0.023$). There is one outlier. County 133 reports 24.9% young males (7.06 SD above the mean). The distribution is positively skewed and is not normally distributed ($W = 0.53, p < .001$). There is a moderate positive correlation between percent young males and crime rate ($r = .29$). This variable will be useful in our analysis.

```
#hist(cleancrime$pctymle, main='Histogram of % Young Male by County', xlab='% Young Male')
#qqnorm(cleancrime$pctymle)
#summary(cleancrime$pctymle)
#mean(cleancrime$pctymle)
#sd(cleancrime$pctymle)
#shapiro.test(cleancrime$pctymle)
#cor(cleancrime$crmrte, cleancrime$pctymle)
```

Tax Per Capita

Crime is often associated with poorer socioeconomic areas. Higher tax income per capita stands in as a proxy variable for the relative wealth of a county. We expect this variable to be inversely related to crime rate. Per capita tax income ranges from 252.69 to 119.76 ($M = 38.06, SD = 13.08$). The distribution is positively skewed and non-normal ($W = 0.70, p < .001$). There is one outlier: county 55, reporting a tax income of

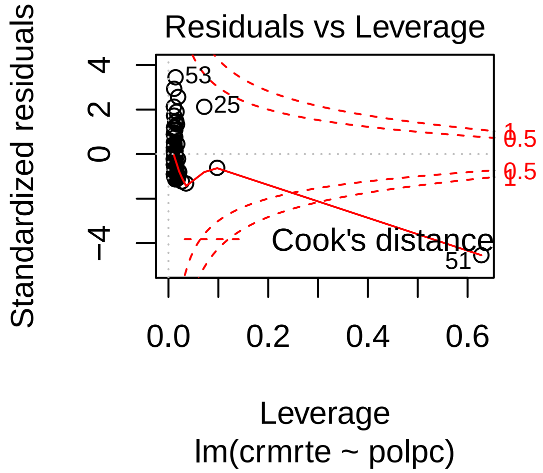
119.76. This is 6.25 SD above the mean. There is a moderate to large positive correlation between tax income and crime rate ($r = 0.45$), which goes against our initial intuition. One possible explanation is that high tax per capita areas are also urban/high density areas. This intuition is partially supported by a moderate positive correlation between population density and tax per capita income ($r = 0.32$).

```
#hist(cleancrime$taxpc, main='Histogram of Tax Rates', xlab='Taxes per Capita')
#qqnorm(cleancrime$taxpc)
#shapiro.test(cleancrime$taxpc)
#summary(cleancrime$taxpc)
#mean(cleancrime$taxpc)
#sd(cleancrime$taxpc)
#cor(cleancrime$crmrte, cleancrime$taxpc)
#cor(cleancrime$density, cleancrime$taxpc)
```

Regression Models: Base Model

In our first model, we include our primary variables of interest: crime rate (crmrte) regressed on police per capita (polpc). Our initial model is not impressive ($\beta_{polpc} = 3.24, SE = 11.74, p = .78$). Diagnostic graphs indicate heteroscedasticity and a violation of 0 conditional mean for error. There are significant violations of our CLM assumptions. Diagnostic graphs indicate an outlier with a Cook's distance in excess of 1. County 155, a data point highlighted as suspect in the initial EDA, contributes much of the residual of our initial model. We believe the suspect values of county 115 and the high leverage and residual effect on attempts to model are enough justification to remove the county from further analysis.

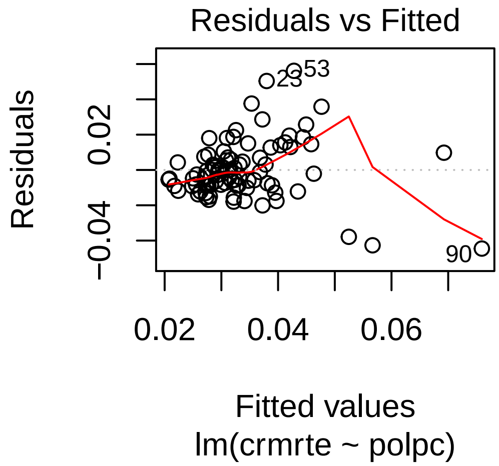
```
basemodel <- lm(crmrte ~ polpc, data=cleancrime)
#summary(basemodel)
plot(basemodel, which=5)
```



```
#cleancrime[51, 'county']
```

Removing county 115 improves model fit significantly ($\beta_{polpc} = 14.87, SE = 6.21, p = .02$). However, assumptions of homoscedasticity and zero conditional mean of residuals are still violated. A second outlier, county 195, has Cook's distance in excess of one, high leverage and high residual in our model. County 195 was highlighted in our EDA as having a suspect prbconv value, but other values are unremarkable. The large leverage and residual that county 195 adds to the model gives us pause. The clear violation of zero conditional mean in our model is enough for us to judge removing county 195 from further analysis is in the best interest of understanding.

```
crimeminusleverage <- subset(cleancrime, cleancrime$county != 115)
basemodel2 <- lm(crmrte ~ polpc, data=crimeminusleverage)
plot(basemodel2, which=1)
```



```
#coeftest(basemodel2, vcov=vcovHC)
#cleancrime[90,'county']
#subset(cleancrime, cleancrime$county == 195)
```

Removing county 195 helps significantly in meeting our assumption of 0 conditional mean of residuals. While evidence of heteroscedasticity is still present, we can use robust standard errors to compensate. Our final baseline model estimate is statistically significant, although in the opposite direction that we first hypothesized. $(\beta_{polpc} = 20.02, SE = 5.14, p < .001, R^2 = 0.32)$.

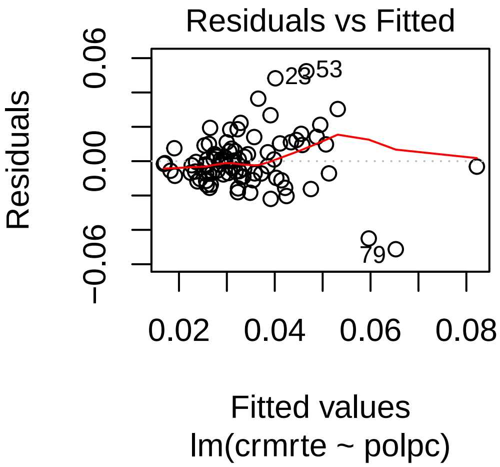
```
crimeminusleverage2 <- subset(crimeminusleverage, crimeminusleverage$county != 195)
basemodel3 <- lm(crmrte ~ polpc, data=crimeminusleverage2)
summary(basemodel3)
```

```
##
## Call:
## lm(formula = crmrte ~ polpc, data = crimeminusleverage2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.051270 -0.007397 -0.001616  0.007284  0.052365
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001940   0.005228   0.371    0.711
## polpc       20.015231   3.122807   6.409 7.27e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01557 on 87 degrees of freedom
## Multiple R-squared:  0.3207, Adjusted R-squared:  0.3129
## F-statistic: 41.08 on 1 and 87 DF,  p-value: 7.273e-09
```

```
coeftest(basemodel3, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.0019402  0.0071287  0.2722 0.7861361
## polpc       20.0152311  5.1431415  3.8916 0.0001945 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(basemodel3, which=1)
```



CLM1: Linear Parameters Since our linear model has its coefficient derived from linear multiplication on a single x, our condition is met.

CLM2: Random Sampling Each of our values is a distinct county with no relationship between one another aside from the state government, so our second condition is met.

CLM3: Sample Variation Because our sample variance is greater than 0, even though it's small, we meet this assumption. It might be worth performing a log transform to address this.

```
var(cleancrime$polpc)
```

```
## [1] 9.740909e-07
```

CLM4: Zero Conditional Mean By examining the Residuals vs Fitted Values plot above and the fitted curve line roughly fits along the zero value, so the condition is met.

CLM5: Homoscedasticity There is evidence of heteroscedasticity in the graph of residuals, confirmed by hypothesis testing ($BP(1) = 18.66, p < .001$). It will be necessary to use robust standard errors for testing regression coefficients.

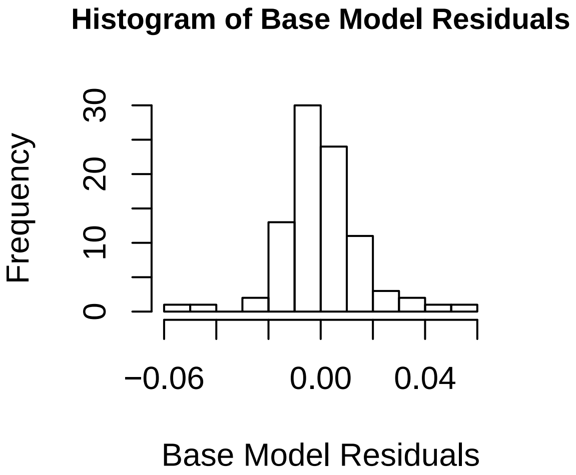
```
bptest(basemodel3)
```

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

```
## data:  basemodel3
## BP = 18.662, df = 1, p-value = 1.561e-05
```

CLM6: Normality A histogram of residuals resembles a normal distribution, although hypothesis testing indicates it is not normal ($W = 0.93, p < .001$). This assumption may not be met.

```
hist(basemodel3$residuals, main='Histogram of Base Model Residuals',
     xlab='Base Model Residuals', cex.main = 0.9)
```



```
shapiro.test(basemodel3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  basemodel3$residuals
## W = 0.93159, p-value = 0.0001574
```

Interpretation of Model 1

An increase of 1 policeman per 1,000 people is associated with a rise in crime of .020 crimes per person. This is opposite of our initial hypothesis. Higher police presence is associated with higher crime. Note: To facilitate understanding of the regression, we have divided the regression weights by 1,000 in our interpretation.

Regression Model: Second Model

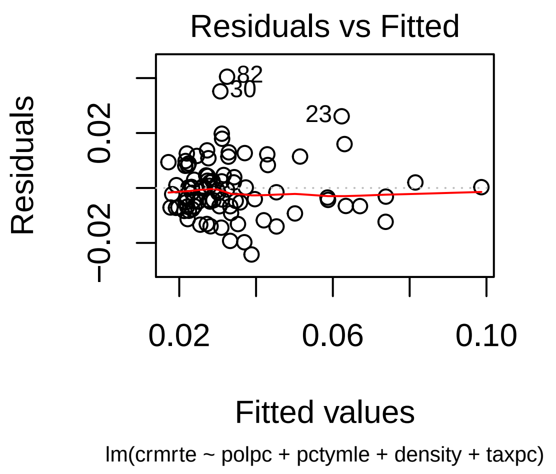
In our second model, we add additional covariates that we believe related to crime. This includes tax per capita (taxpc), percent of young male (pctymle), population density (density), and whether the county is considered an urban area (urban). We note that the regression weight for urban is statistically insignificant ($\beta_{urban} = 0.0008, SE = 0.007, p = .92$). Dropping the variable 'urban' raises $AdjR^2$ from 0.6445 to 0.6487. It seems the variable 'urban' is not contributing much to the model and so we drop it.

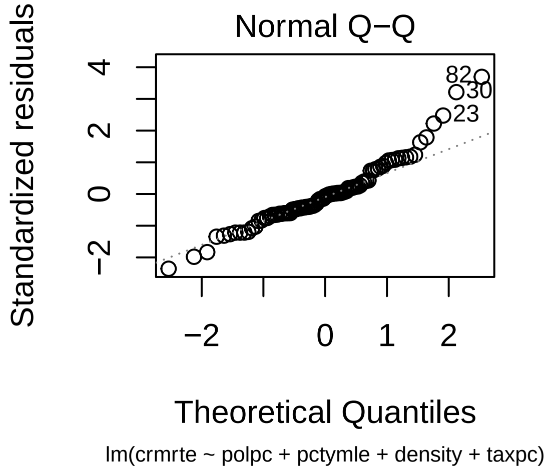
The revised model 2 regresses crime rate on police, percent of young male, population density and tax income per capita (Adj $R^2 = 0.6487$).

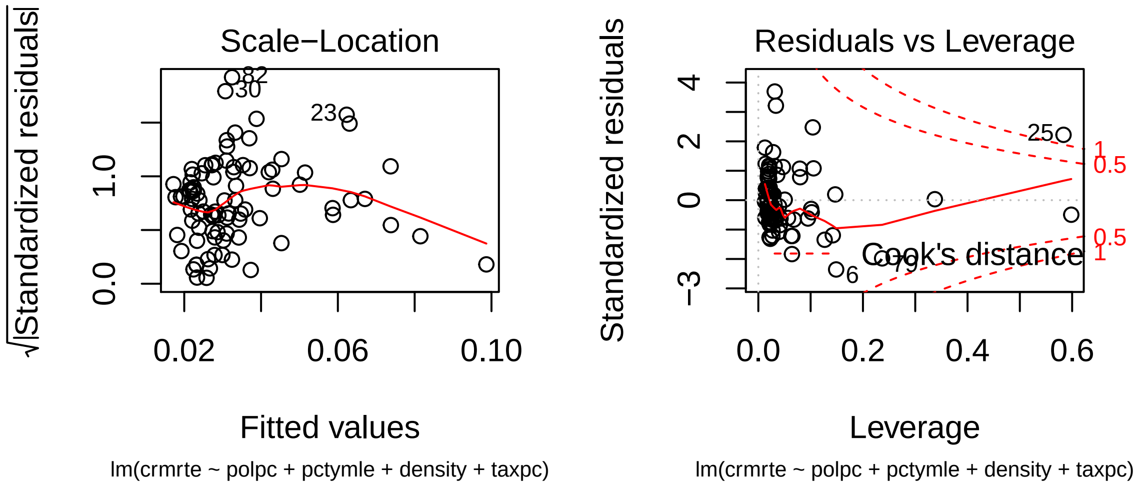
```
#model2_1 <- lm(crmrte ~ polpc + pctymle + density + taxpc +urban, data=crimeminusleverage2)
#summary(model2_1)
#plot(model2_1)

model2 <- lm(crmrte ~ polpc + pctymle + density + taxpc, data=crimeminusleverage2)
coeftest(model2, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -0.01033582  0.01059775 -0.9753  0.33222
## polpc        7.11835146  5.29183358  1.3452  0.18219
## pctymle      0.15953946  0.06267232  2.5456  0.01274 *
## density      0.00703611  0.00130072  5.4094 5.85e-07 ***
## taxpc        0.00024293  0.00024190  1.0043  0.31813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#summary(model2)
plot(model2, cex.sub=0.7)
```







CLM1: Linear Parameters This condition is met. Any population distribution can be represented as a linear model plus some unconstrained error. Our linear model describes 65% of the variance in the dependent variable ($AdjR^2 = .65$)

CLM2: Random Sampling This condition is met. Given that the population of interest is counties in North Carolina, we have sampled essentially all of the population. Our models was based on 89 of 100 counties. It can therefore be said that our sample is representative of counties in North Carolina.

CLM3: No perfect colinearity This condition is met. Our highest correlation between predictors is between population desnity and tax per capita ($r = .59$)

```
cor(crimeminusleverage2$polpc, crimeminusleverage2$pctymle)  # r=.19
```

```
## [1] 0.1862292
```

```
cor(crimeminusleverage2$polpc, crimeminusleverage2$density) # r=.40
```

```
## [1] 0.3997005
```

```
cor(crimeminusleverage2$polpc, crimeminusleverage2$taxpc) # r=.59
```

```
## [1] 0.5856566
```

```
cor(crimeminusleverage2$pctymle, crimeminusleverage2$density) # r=.11
```

```
## [1] 0.1133805
```

```
cor(crimeminusleverage2$pctymle, crimeminusleverage2$taxpc) # r=-0.09
```

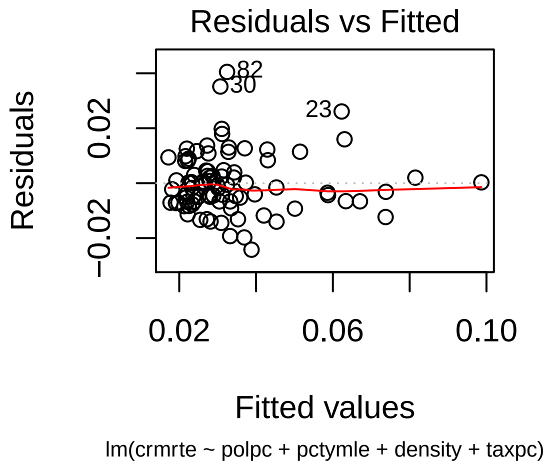
```
## [1] -0.09188169
```

```
cor(crimeminusleverage2$density, crimeminusleverage2$taxpc) # r=.32
```

```
## [1] 0.3181109
```

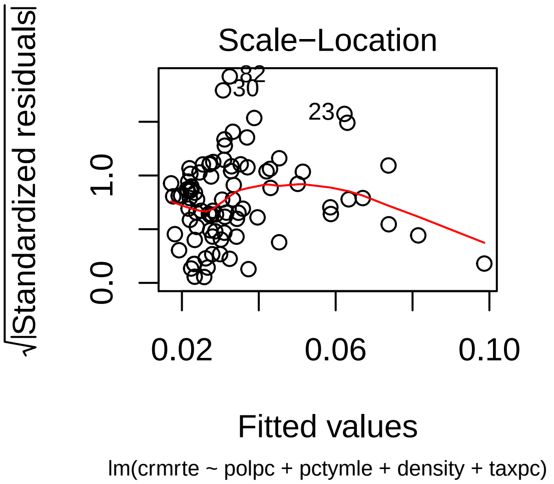
CLM4: Zero Conditional Mean This condition is met. Examining the graph of residuals versus fitted values, the red spline line appears to tracks near zero across all values of x.

```
plot(model2, which=1, cex.sub=0.7)
```



CLM5: Homoscedasticity This condition is not met. Examining The Residuals vs. Fitted plot above, and the Scale-Location Plot below, it's a little hard to tell whether the variance follows an even band given that the number of data points drops off significantly for certain values of X. The results of the Breusch-Pagan Test rejects the null hypothesis, indicating heteroscedasticity $(BP = 10.762(4), p = 0.029)$.

```
plot(model2, which=3, cex.sub=0.7)
```

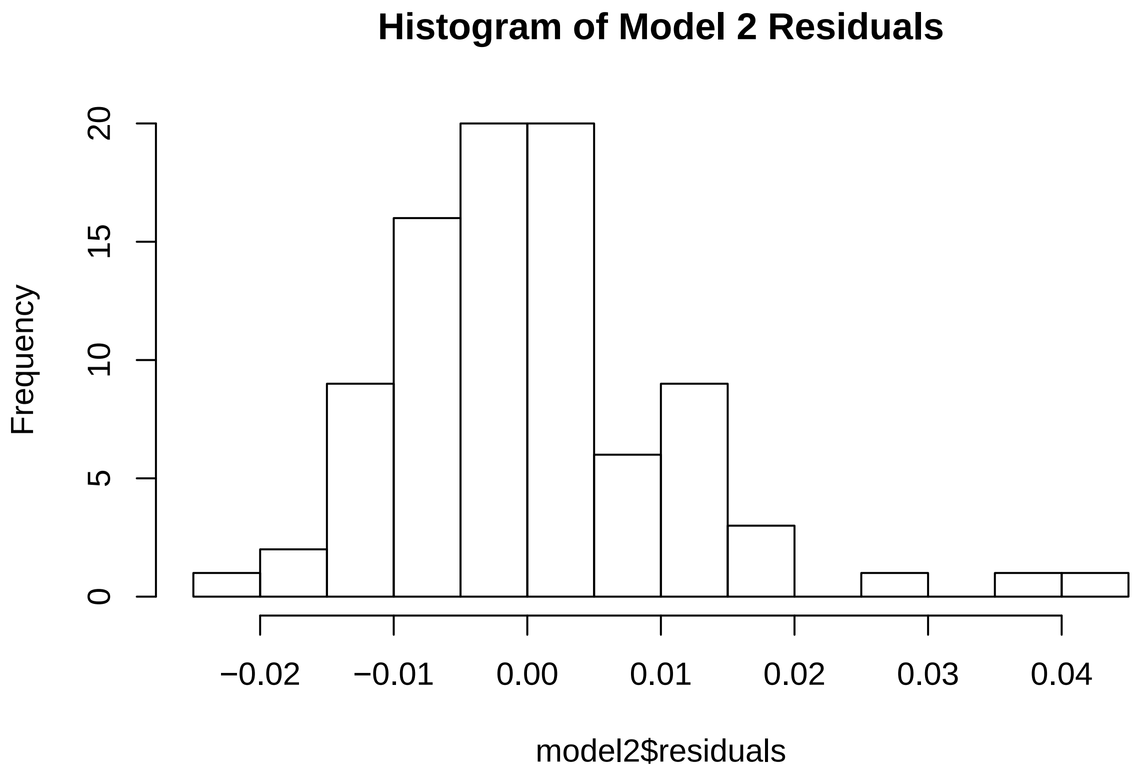


```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 10.762, df = 4, p-value = 0.02937
```

CLM6: Normality This condition is marginally not met. The distribution of residuals has a mean of 0 $(t(88) < .001, p = 1.00)$. A histogram of residuals visually approximates a normal distribution with a positive skew. The results of the Shapiro-Wilk test indicate a non-normal distribution $(W = 0.94, p < .001)$.

```r
hist(model2$residuals, breaks=20, main='Histogram of Model 2 Residuals')
```



```r
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.94439, p-value = 0.0008269
```

```r
t.test(model2$residuals, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  model2$residuals
## t = -6.0577e-17, df = 88, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.002291728  0.002291728
## sample estimates:
##     mean of x
## -6.985643e-20
```

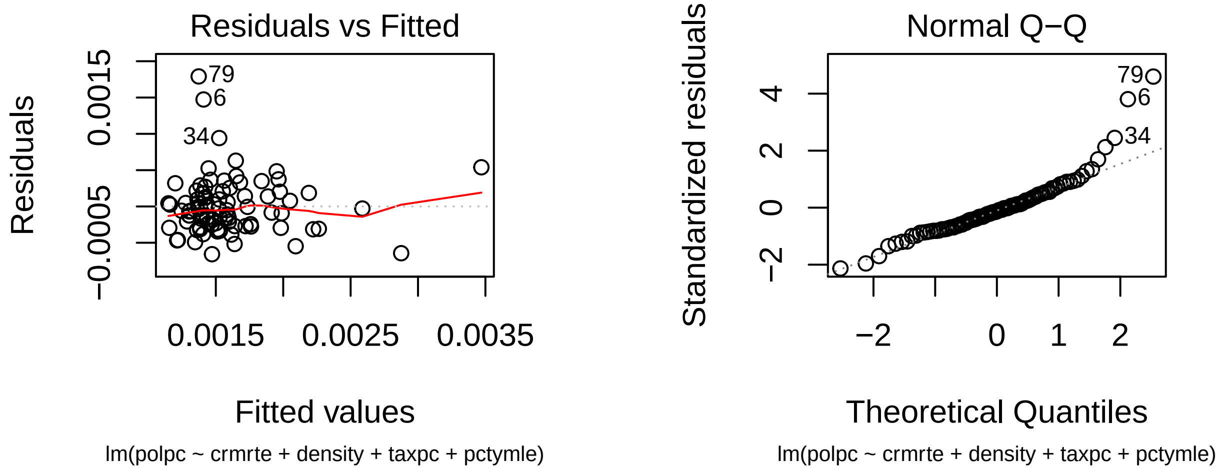
Interpretation of Model 2

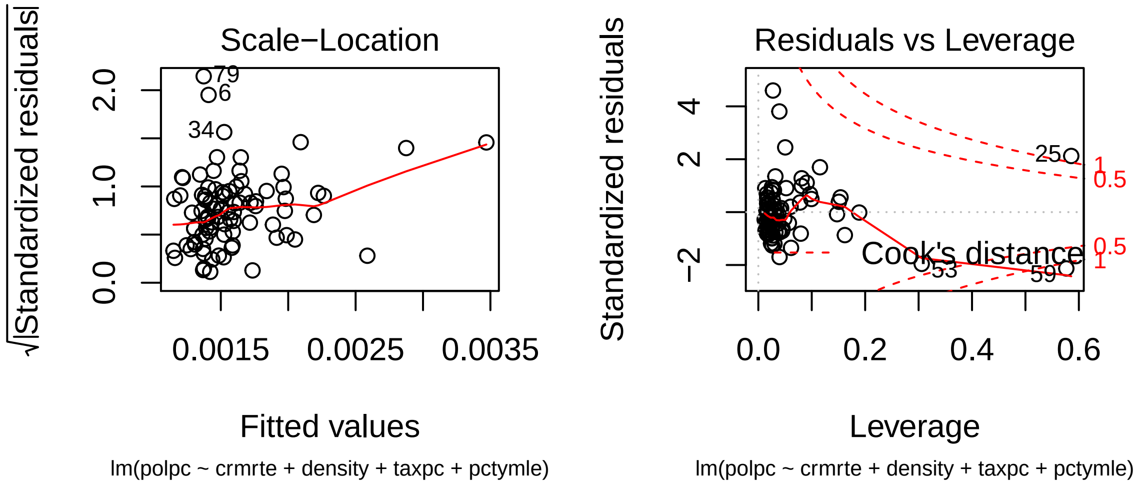
We now have a model that can predict the crime rate, given factors of police per capita rate, population density, wealth (with proxy variable tax per capita), and percentage of young males. An increase of 10 people per square mile is associated with almost a 0.07 unit increase in crimes per person. An increase of 1 policeman per 100 people is associated with almost a 0.07 unit increase in crimes per person. An increase in tax per capita by 100 would result in a 0.02 increase in crimes per person. Likewise an increase in the percentage of young men by 1% would result in a 0.2 increase in crimes per person. Given the positive correlation between crime rate and police presence, it is difficult to conclude that higher police presence resulted in lower crime. We continue to investigate this question in model 3, and consider a broader change to our research question.

Regression Model: Third Model

Our research question was to examine the data to see if having higher police presence resulted in lower crime. We did not see this in the data. In fact, higher police presence was associated with higher rates of crime, even after controlling for possible covariates such as population density, young males present, etc. We do not believe it makes theoretical sense to infer a causal relationship here. Having more police doesn't cause more crime. Rather, we believe the direction of causality is reversed: high crime results in more police being sent to a problem area. Given this change in causal direction, we believe it's appropriate and an interesting analysis to designate the variable 'police per capita' as the criterion variable and 'crime rate' as a predictor variable. The resulting model will give a predicted level of police presence for every county based on factors of crime rate, density, etc… This can then be compared to actual police presence, and counties can be classified as 'over policed' or 'under policed' based on significant characteristics. We proceed with this analysis for our third model.

```
model3 <- lm(polpc ~ crmrte + density + taxpc + pctymle, data=crimeminusleverage2)
plot(model3, cex.sub=0.7)
```





```r
#summary(model3)
coeftest(model3, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.2487e-04 4.7371e-04  0.6858  0.49473
## crmrte      8.9492e-03 4.6310e-03  1.9325  0.05667 .
## density     3.4035e-06 6.5668e-05  0.0518  0.95879
## taxpc       1.8392e-05 7.8237e-06  2.3508  0.02108 *
## pctymle     3.0687e-03 4.8961e-03  0.6268  0.53252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Detailed CLM Assumptions Analysis for Model 3

CLM.1 (Linear in Parameters):

This condition is met. Any population distribution can be represented as a linear model plus some unconstrained error. Our linear model describes 45% of the variance in the dependent variable ($AdjR^2 = .45$)

CLM.2 (Random Sampling):

This condition is met. Given that the population of interest is counties in North Carolina, we have sampled essentially all of the population. Our models was based on 89 of 100 counties. It can therefore be said that our sample is representative of counties in North Carolina.

CLM.3 (No Perfect collinearity):

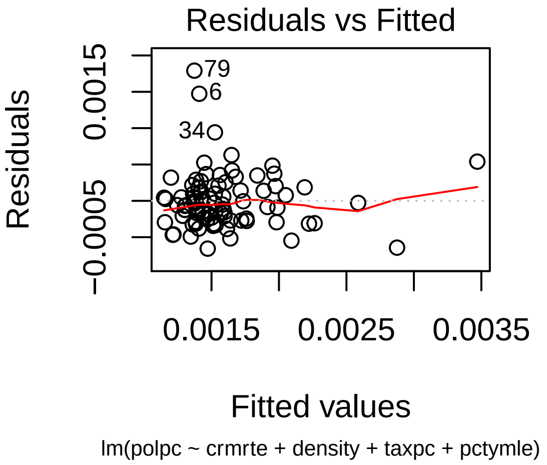
This condition is met. Our highest correlation between predictors is between population density and crime rate ($r = .73$)

```r
cor(crimeminusleverage2$crmrte, crimeminusleverage2$density)  # r=.73
cor(crimeminusleverage2$crmrte, crimeminusleverage2$taxpc) # r=.45
cor(crimeminusleverage2$crmrte, crimeminusleverage2$pctymle) # r=.28
cor(crimeminusleverage2$taxpc, crimeminusleverage2$pctymle) # r=-.09
cor(crimeminusleverage2$taxpc, crimeminusleverage2$density) # r=.31
cor(crimeminusleverage2$pctymle, crimeminusleverage2$density) # r=.11
```

CLM.4 (Zero Conditional Mean: $E(u|x_1, x_2, ..., x_k) = 0$)

This condition is met. Examining the graph of residuals versus fitted values, the red spline line appears to tracks near zero across all values of x.

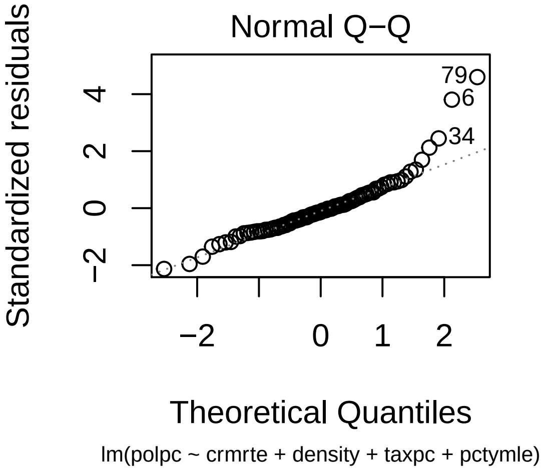
```r
plot(model3, which=1, cex.sub=0.7)
```



CLM.5 (Homoscedasticity: $Var(u|x_1, ..., x_k) = \sigma^2$)

This condition is met. Examining The Residuals vs. Fitted plot above, and the Scale-Location Plot below, it's a little hard to tell whether the variance follows an even band given that the number of data points drops off significantly for certain values of X. The results of the Breusch-Pagan Test fail to reject the null hypothesis. We do not reject the hypothesis of homoscedasticity (BP = 2.96(4), p = 0.56). We can arguably use non-robust standard errors of our regression weights to test for significance, but we choose to be more conservative and use robust standard errors. This does not change statistical significance of any of the regression weights at the $p = .05$ level.

```r
plot(model3, which =2, cex.sub=0.7)
```



```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 2.9602, df = 4, p-value = 0.5645
```

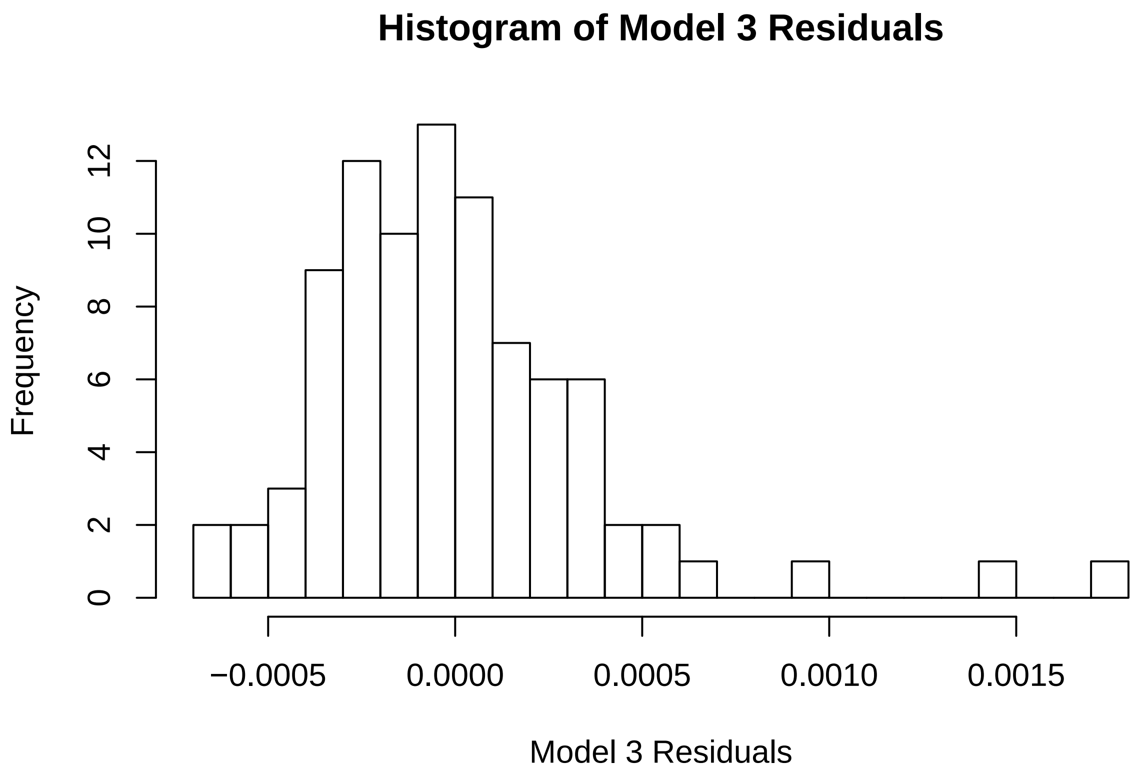
```
summary(model3)
```

```
##
## Call:
## lm(formula = polpc ~ crmrte + density + taxpc + pctymle, data = crimeminusleverage2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.575e-04 -2.538e-04 -5.155e-05  1.765e-04  1.791e-03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.249e-04  2.113e-04   1.538   0.1279
## crmrte      8.949e-03  3.743e-03   2.391   0.0191 *
## density     3.403e-06  4.076e-05   0.083   0.9337
## taxpc       1.839e-05  3.738e-06   4.920 4.25e-06 ***
## pctymle     3.069e-03  1.955e-03   1.570   0.1202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0003948 on 84 degrees of freedom
## Multiple R-squared:  0.4734, Adjusted R-squared:  0.4483
## F-statistic: 18.88 on 4 and 84 DF,  p-value: 4.199e-11
```

CLM.6 (Normality: $u \sim N(0, \sigma^2)$)

This condition is marginal. The distribution of residuals has a mean of 0 ($t(88) < .001, p = 1.00$). A histogram of residuals visually approximates a normal distribution with a positive skew. The results of the Shapiro-Wilk test indicate a non-normal distribution (W=0.87, p <.001).

```
hist(model3$residuals, breaks=20, main='Histogram of Model 3 Residuals', xlab='Model 3 Residuals')
```



```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.87168, p-value = 3.033e-07
```

```
t.test(model3$residuals, mu = 0)
```

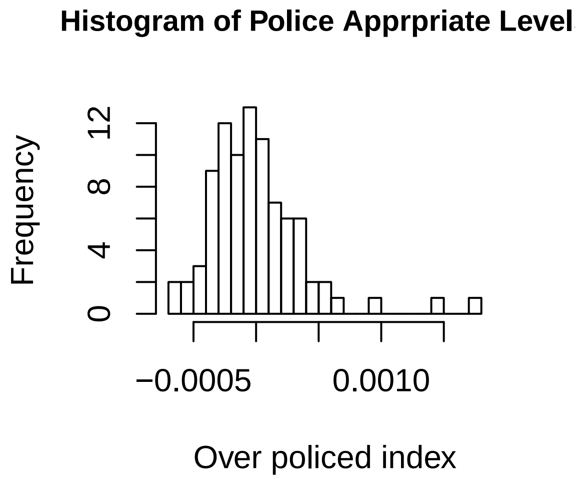
```
##
##  One Sample t-test
##
## data:  model3$residuals
## t = 4.1316e-16, df = 88, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -8.125798e-05  8.125798e-05
## sample estimates:
##   mean of x
## 1.68935e-20
```

Interpretation of Model 3

We took the unconventional step of switching the places of our predictor and criterion variables during our analysis. This comes after our analysis indicated we were mistaken in our initial understanding of the direction of causality. It's not that high police per capita lowers crime, it's that higher crime causes police per capita to increase. Adjusting mid-course to our new insight, we now have a model that suggests the appropriate police presence, given factors of crime, population density, wealth (with proxy variable tax per capita), and percentage of young males. We can use this information to identify counties that are over-policed or under-policed based on the conditions.

In the analysis below, we predict the appropriate police per capita rate for every county based on our model. We then subtract the predicted police per capita from the actual police per capita to create an index of policing level. A positive value represents a county that is 'over policed' based on the conditions. A negative number indicates a county that is 'under policed' based on the conditions. Interestingly, a small positive correlation is observed where counties that are 'over-policed' have a higher probability of arrest than counties that are 'under-policed' ($r = 0.14$).

```r
crimeminusleverage2$pred_polpc <- predict(model3)
crimeminusleverage2$delta_polpc <- crimeminusleverage2$polpc - crimeminusleverage2$pred_polpc
hist(crimeminusleverage2$delta_polpc, breaks = 20, main='Histogram of Police Apprpriate Levels',
     xlab='Over policed index', cex.main=0.9)
```



```r
summary(crimeminusleverage2$delta_polpc)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -6.575e-04 -2.538e-04 -5.155e-05  0.000e+00  1.765e-04  1.791e-03
```

```r
cor(crimeminusleverage2$prbarr, crimeminusleverage2$delta_polpc)
```

```
## [1] 0.1377155
```

```r
cor(crimeminusleverage2$mix, crimeminusleverage2$delta_polpc)
```

```
## [1] 0.2253812
```

The Regression Table

Here is our initial regression table for the first two models, both of which regress on crime rate. Our second model explains significantly more variation in crime rate and has a lower AIC, indicating a higher quality model.

```
vcovmatrix1 <- vcovHC(basemodel3, "HC1")
robust1 <- coeftest(basemodel3, vcov=vcovmatrix1)

vcovmatrix2 <- vcovHC(model2, "HC1")
robust2 <- coeftest(model2, vcov=vcovmatrix2)


stargazer(basemodel3, model2, title="Models to Predict Crime Rate", type="latex", add.lines=list(c("AIC
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Aug 15, 2018 - 12:43:32 PM

Table 1: Models to Predict Crime Rate

	Dependent variable:	
	crmrte	
	(1)	(2)
polpc	20.015***	7.118**
	(3.123)	(2.978)
pctymle		0.160***
		(0.053)
density		0.007***
		(0.001)
taxpc		0.0002**
		(0.0001)
Constant	0.002	−0.010*
	(0.005)	(0.006)
AIC	-462.567789741388	-541.15533123602
Observations	89	89
R^2	0.321	0.665
Adjusted R^2	0.313	0.649
Residual Std. Error	0.016 (df = 87)	0.011 (df = 84)
F Statistic	41.080*** (df = 1; 87)	41.616*** (df = 4; 84)
Note:		*p<0.1; **p<0.05; ***p<0.01

However, it is our last model in which we adjusted our outcome which will provide the most actionable results for the political campaign. To ensure that we can draw useful conclusions from our final model, we need to apply more robust standard errors.

```
vcovmatrix3 <- vcovHC(model3, "HC1")
coeftest(model3, vcov=vcovmatrix3)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.2487e-04 2.4026e-04  1.3522  0.17996
## crmrte      8.9492e-03 3.8634e-03  2.3164  0.02297 *
## density     3.4035e-06 4.2768e-05  0.0796  0.93676
## taxpc       1.8392e-05 4.1293e-06  4.4539 2.58e-05 ***
## pctymle     3.0687e-03 2.5672e-03  1.1953  0.23533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By computing a variance-covariance matrix for our final model and then conducting a t-test, we can see that a more robust assessment of our model is more conservative. Tax Per Capita is still highly statistically significant, but not quite as much as it was without robust standard errors. For even greater detail, we can conduct Akaike's Information Criterion and add that to our final table.

```
vcovmatrix <- vcovHC(model3, "HC1")
robust <- coeftest(model3, vcov=vcovmatrix)
stargazer(model3, robust, title="Model to Predict Police Per Capita with AIC", type="latex", add.lines=
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Aug 15, 2018 - 12:43:32 PM

In practical terms, density has no effect on police per capita. An increase in crimes per person of 1 is associated with almost a 1 unit increase in police per capita. An increase in tax per capita by $100 would result in a 0.2 increase in police per capita. Likewise an increase in the percentage of young men by 1% would result in a 0.3 increase in police per capita.

Relative to our first two models, this third AIC is much lower and therefore denotes a much better model specification. It seems our inclination to change outcomes has paid off in our ability to extrapolate practical conclusions.

By using model 3, we can see if the county is over policed or under policed and decide whether we should have more police in the county. From the plot Residuals/Fitted, we can see that most of the difference is between $(-0.0005\ 0.0005)$, which is quite significant compared with the mean 0.0017022. Therefore, we do believe that the model is practically significant.

The Omitted Variables Discussion

Economist Gary Becker, who has written on the economics of crime, hypothesizes that criminals decide to engage in criminal behavior after making a rational evaluation of several factors. These factors include the economic opportunity of engaging in legal behavior, the economic opportunity engaging in illegal behavior, the probability of getting caught and the severity of punishment if caught. We can use this theory to identify potential omitted variables. Regarding legal economic opportunity, our variable of tax per capita (a proxy for wealthy areas) may have some overlap. However, we did not look at average wage data which may have also informed this construct. We did not evaluate any variables concerning economic opportunity for illegal behavior. There don't seem to be any in the database. This is likely very difficult data to get. With regards to probability of getting caught, we did not use variables on probability of arrest in our analysis. Lastly, for severity of punishment, there was information on probability of conviction, prison sentence, and length of sentence that we did not use. Our research question focused specifically on police presence (and not the economics of crime), and so these potential variables were not immediately obvious ones for us to consider. However, they still could have been useful covariates that we could have investigated.

Table 2: Model to Predict Police Per Capita with AIC

	Dependent variable:	
	polpc	
	OLS	*coefficient test*
	(1)	(2)
crmrte	0.009**	0.009**
	(0.004)	(0.004)
density	0.00000	0.00000
	(0.00004)	(0.00004)
taxpc	0.00002***	0.00002***
	(0.00000)	(0.00000)
pctymle	0.003	0.003
	(0.002)	(0.003)
Constant	0.0003	0.0003
	(0.0002)	(0.0002)
AIC	-1135.57427333374	
Observations	89	
R^2	0.473	
Adjusted R^2	0.448	
Residual Std. Error	0.0004 (df = 84)	
F Statistic	18.876*** (df = 4; 84)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Conclusion

Our goal was to analyze the claim that adding additional police will reduce crime. In the course of our analysis, we found that higher police rates were associated with higher crime. We reject the notion that more police cause more crime, and instead conclude the direction of causality is reversed: higher crime brings more police.

Adjusting our analysis appropriately, we then decided to focus our attention on making police presence the criterion variable of interest, predicted by crime rate, percentage of male youth, relative wealth (tax income), and density. Using this model, we were then able to identify counties that are 'over-policed' or 'under-policed' based on what conditions suggest their police presence should be. We believe this classification has some useful applications. For instance, it is a way to evaluate the appropriate allocation of police resources. There are research applications as well. For instance, if one were to examine the correlation between police per capita and probability of arrest ($r = -0.09$), they may conclude increasing police force does not increase the probability of arrest. However, if we take our index of 'over/under policing', we note that over-policed counties have a higher probability of arrest than under-policed counties ($r = 0.14$).

```r
cor(crimeminusleverage2$prbarr, crimeminusleverage2$polpc)
```

```
## [1] -0.09578462
```

```r
cor(crimeminusleverage2$prbarr, crimeminusleverage2$delta_polpc)
```

```
## [1] 0.1377155
```

We changed our focus and shifted our criterion relatively late in our study. If we were to continue going forward, we would start this process over again from the beginning, with the premise of modeling police presence. We need to reevaluate our model choices with this new focus in mind, as assumptions we made for model 1 and model 2 had a different initial purpose.

Research is an iterative process and, going forward, we would continue to look at ways of improving our model. Some things we would consider going forward: regression weights near zero are difficult to interpret. We should consider a transformation where we multiply our variable by a large constant (100 or 1000) to make regression weights easier to interpret. We had difficulty with some of our residuals approximating a normal distribution. We might further investigate possible transformations which might be helpful. We have additional covariates that can be considered. Our research started by looking for covariates of crime, and our new focus would be finding covariates for allocation of police.

```r
library(knitr)
allocation <- data.frame('County'= crimeminusleverage2$county, 'Allocation of Resources'=crimeminusleve
kable(allocation, title="Adjusted Allocation of Police Resources", type="latex")
```

County	Allocation.of.Resources
1	0.3671413
3	-0.4671322
5	-0.0695658
7	-0.0515550
9	-0.2945351
11	1.4728843
13	-0.0357928
15	-0.3255316
17	-0.1821804
19	0.1377504
21	-0.0849726
23	0.0970368
25	0.5249399
27	-0.2704918

County	Allocation.of.Resources
33	-0.4624214
35	-0.0939534
37	-0.0816369
39	0.0428377
41	-0.0272672
45	-0.2537867
47	-0.1533609
49	-0.2756555
51	-0.3142606
53	-0.5170161
55	0.5396317
57	-0.0067387
59	-0.3340306
61	0.0483422
63	0.1857459
65	0.1417669
67	0.2022262
69	-0.3128790
71	0.4184370
77	0.9418442
79	-0.3812070
81	0.4842108
83	-0.2112484
85	-0.1581055
87	0.2703584
89	-0.1264890
91	0.0971563
93	0.0049694
97	0.1109738
99	0.2550697
101	-0.1690176
105	-0.3081604
107	-0.2945925
109	-0.2061274
111	-0.0589526
113	0.0306090
117	-0.3177530
119	-0.6438961
123	0.2092239
125	0.3305630
127	0.1765431
129	-0.0286186
131	-0.2304744
133	-0.5476846
135	0.3734401
137	-0.6574929
139	0.2892773
141	-0.2302867
143	-0.4919000
145	-0.2698705
147	0.3489846
149	0.3193969

County	Allocation.of.Resources
151	0.0475802
153	-0.1448781
155	0.0229219
157	0.3551356
159	0.1831066
161	-0.0615052
163	-0.0087444
165	-0.2445830
167	0.1275984
169	-0.2883203
171	0.2169897
173	1.7908265
175	0.0581506
179	-0.1103114
181	-0.0058148
183	0.0772585
185	-0.1672567
187	-0.3439380
189	0.6301725
191	-0.3874093
193	-0.1237594
193	-0.1237594
197	0.0258201