Lab 4: Reducing Crime

w203: Statistics for Data Science

Colby Carter & Jennifer Philippou

August 22, 2017

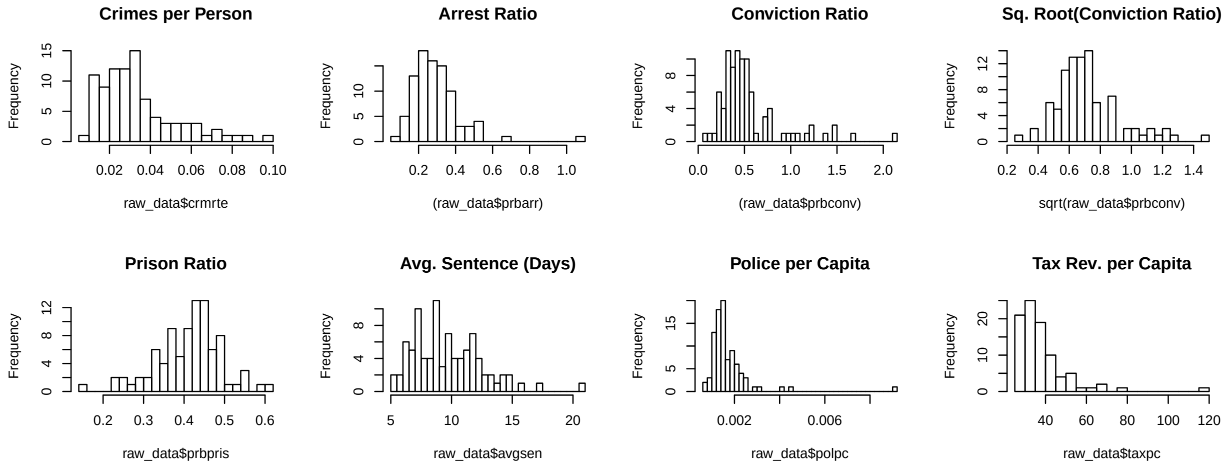
Introduction

Safety and security are fundamental human needs and society relies on the government to provide a peaceful environment. The political leaders, law enforcement agencies, and the legal system work together to bring about a halcyon community. This analysis leverages data to help local officials better understand the factors associated with crime. A stronger understanding of the current environment and relationships within the data facilitates the creation of strategic policy that will mitigate future crime and enhance the community.

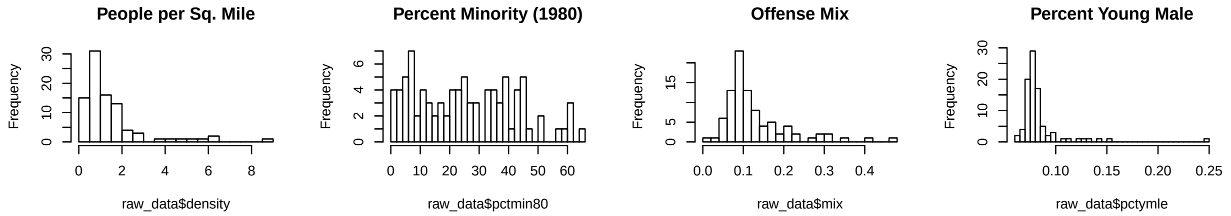
Exploratory Data Analysis

The dataset examined in this analysis include twenty-six different variables for ninety different counties. As the describe function shows, the data also has no missing values and solely represents 1987 metrics for a given point in time. The file seems to be missing a dummy variable for region so a column has been added to adjust for that. Additionally there appears to be an erroneous entry for service wage for county no. 185 as it is well above all of the other wages across the board and it has been removed. Initally the probabilites for conviction, arrest and prison that exceeded one concerned us, however further research into the dataset reveals that the metrics are actually ratios and can be included in the analysis: "PA (which is measured by the ratio of arrests to offences), probability of conviction given arrest PC (which is measured by the ratio of convictions to arrests), probability of a prison sentence given a conviction PP (measured by the proportion of total convictions resulting in prison sentences)".

```r
par(mfrow = c(2,4))
hist(raw_data$crmrte, breaks = 30, main = "Crimes per Person")
# hist(sqrt(raw_data$crmrte), breaks = 30, main = "Sq-Root of Crimes/Person")
hist((raw_data$prbarr), breaks = 30, main = "Arrest Ratio")
hist((raw_data$prbconv), breaks = 30, main = "Conviction Ratio")
hist(sqrt(raw_data$prbconv), breaks = 30, main = "Sq. Root(Conviction Ratio)")
hist(raw_data$prbpris, breaks = 30, main = "Prison Ratio")
hist(raw_data$avgsen, breaks = 30, main = "Avg. Sentence (Days)")
hist(raw_data$polpc, breaks = 30, main = "Police per Capita")
hist(raw_data$taxpc, breaks = 30, main = "Tax Rev. per Capita")
```



```
hist(raw_data$density, breaks = 30, main = "People per Sq. Mile")
hist(raw_data$pctmin80, breaks = 30, main = "Percent Minority (1980)")
hist(raw_data$mix, breaks = 30, main = "Offense Mix")
hist(raw_data$pctymle, breaks = 30, main = "Percent Young Male")
```



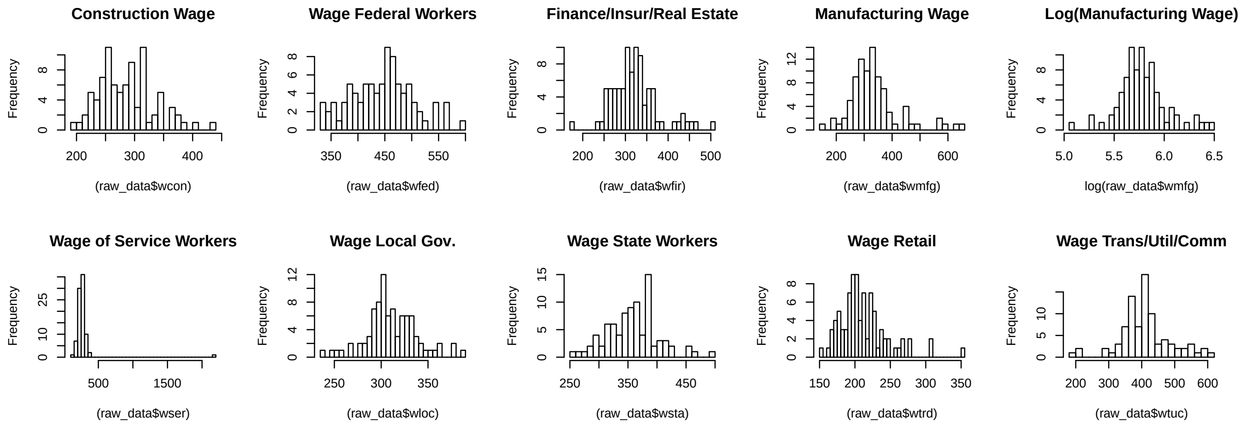
There are a handful of instances of very large outliers that tend to be associated with very small population densities (i.e., low population and/or high land area), which include: police per capita (0.9 compared to 0.5), tax revenue per capita (119, nearly double the next highest value), and percent young male (10 pct points higher at 25% of total population–an unusually high rate possibly explained by the presence of something like a military base). With likely explanations outside of our data, we exclude these observations from our proposed models.

```
#Remove record with top coded variable (like done in the batting averages example 13.12)
raw_data$removalFlag = ifelse(raw_data$county == "55",1, #taxpayer
                              ifelse(raw_data$county == "133",1, #pctmale
                                     ifelse(raw_data$county == "115",1, 0))) #polpc
raw_data = subset(raw_data, raw_data$removalFlag !=1)
```

For the crimes per person, probability of arrest, probability of conviction, average sentance days, offene mix, density, and tax per capita, we see varying right skewness and attempt to mitigate violations of the assumptions for linear regression by transforming these variables with the square root function (e.g., square root of the Conditional Conviction Ratio, in histogram above).

For the average weekly wage levels by industry, most distributions are fairly normal with some right skew in a couple sample distributions (e.g., see Construction and Manufacturing). We considered transformations, including taking the natural log or square root to mitigate this skew (e.g., Log(Manufacturing Wage) below), but given the Central Limit Theorem and our sample size above 30, we did not consider this necessary for linear regression. However, there is one case of an extreme outline in the average wage of service workers nearly ten times higher than median county wage value, which would appear to be a data entry error likely off by an order of magnitude; we see reason to ignore this wage value error but not the observation itself and its other variables.

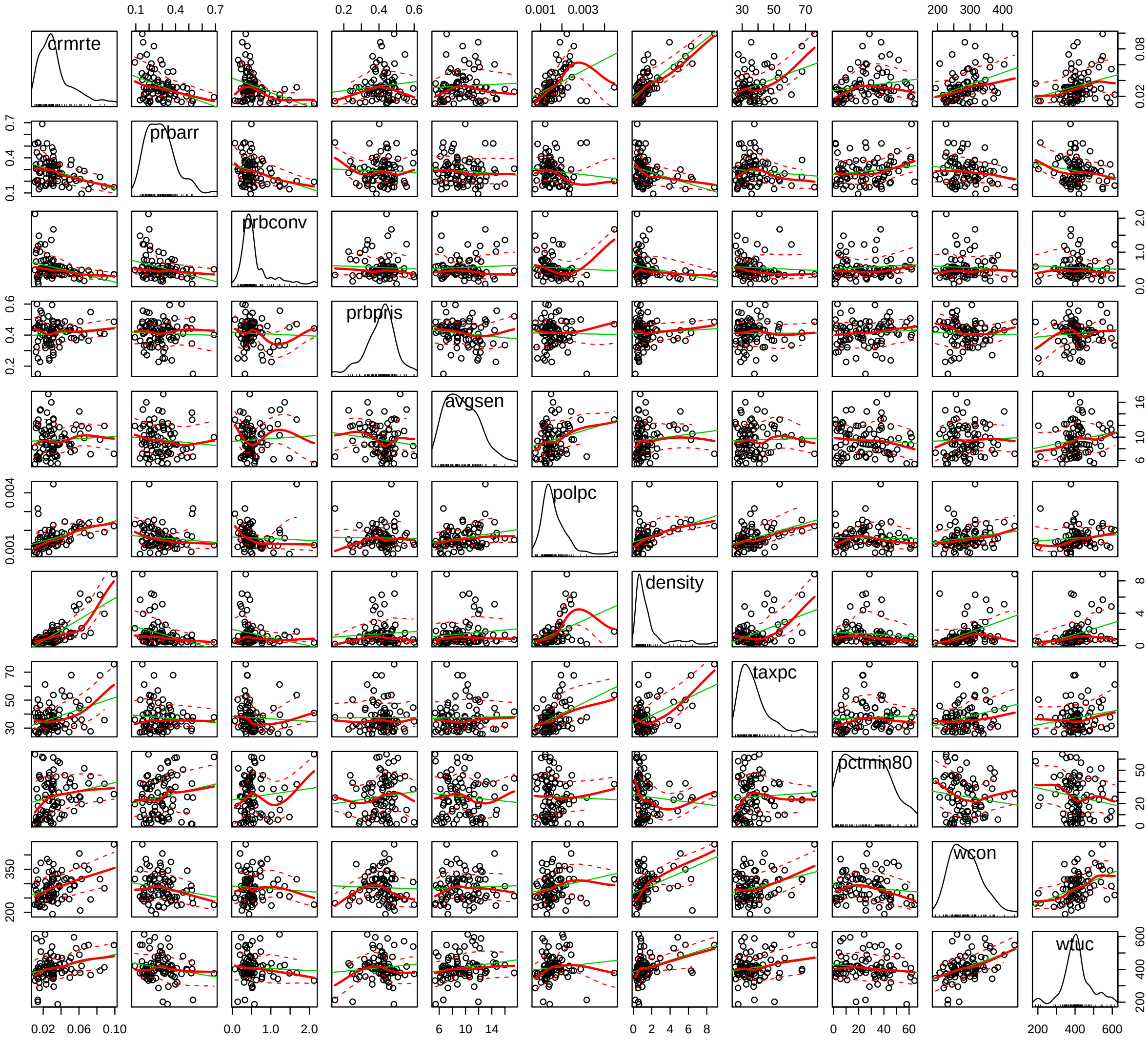
```r
#Dropping service wage outlier due to likely data entry error or extremely low desity/population influe
par(mfrow = c(2,5))
hist((raw_data$wcon), breaks = 30, main = "Construction Wage")
#hist(sqrt(raw_data$wcon), breaks = 30, main = "Sq-Root of Contstruction Wage")
hist((raw_data$wfed), breaks = 30, main = "Wage Federal Workers")
hist((raw_data$wfir), breaks = 30, main = "Finance/Insur/Real Estate")
hist((raw_data$wmfg), breaks = 30, main = "Manufacturing Wage")
hist(log(raw_data$wmfg), breaks = 30, main = "Log(Manufacturing Wage)")
hist((raw_data$wser), breaks = 30, main = "Wage of Service Workers")
hist((raw_data$wloc), breaks = 30, main = "Wage Local Gov.")
hist((raw_data$wsta), breaks = 30, main = "Wage State Workers")
hist((raw_data$wtrd), breaks = 30, main = "Wage Retail")
hist((raw_data$wtuc), breaks = 30, main = "Wage Trans/Util/Comm")
```



```r
raw_data$wser[84] = NA
```

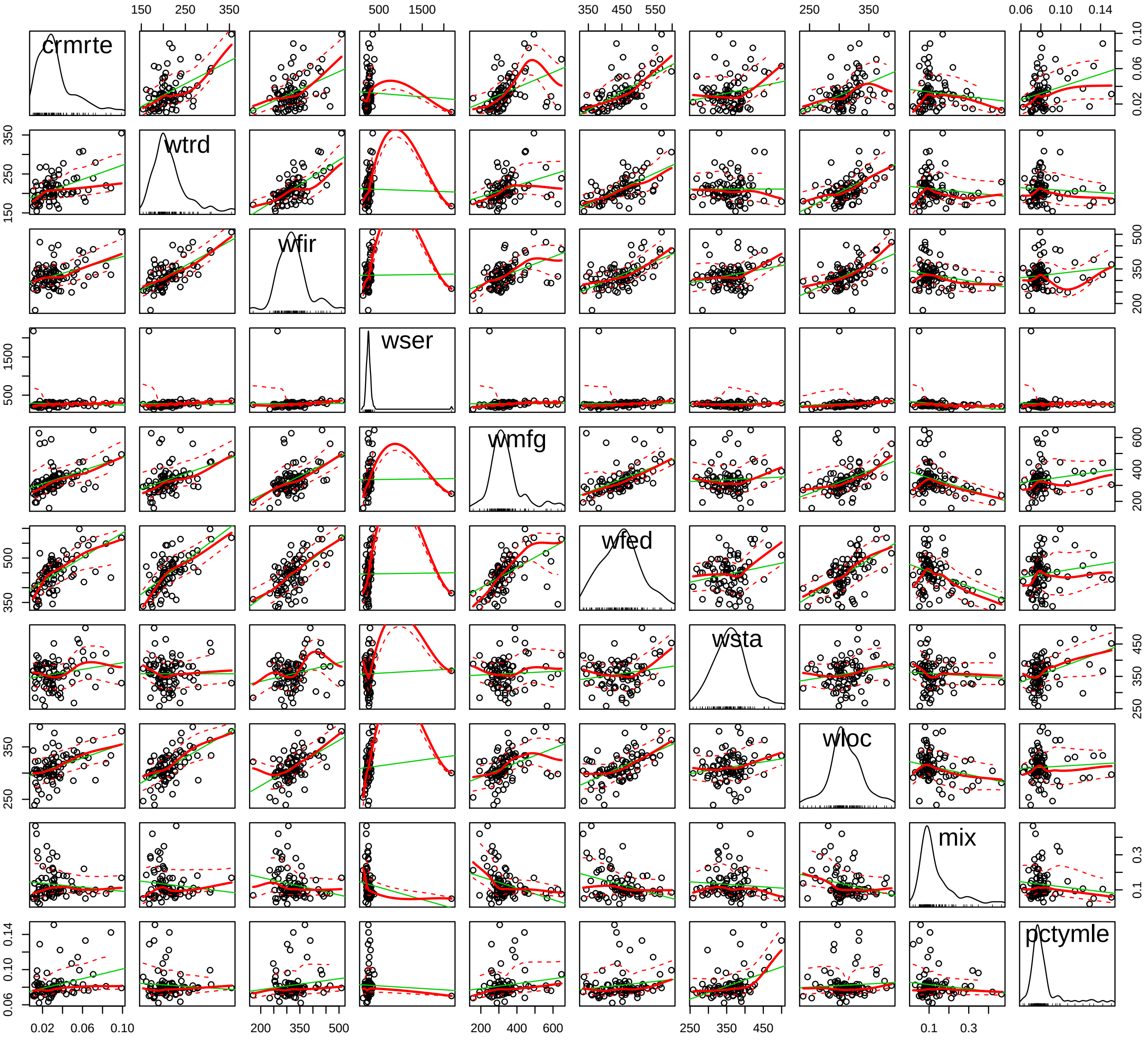
Scatter plot matrix Part I

```r
scatterplotMatrix(raw_data[c(4:11,15:17)])
```



Scatterplot matrix Part II

```
scatterplotMatrix(raw_data[c(4, 18:26)])
```



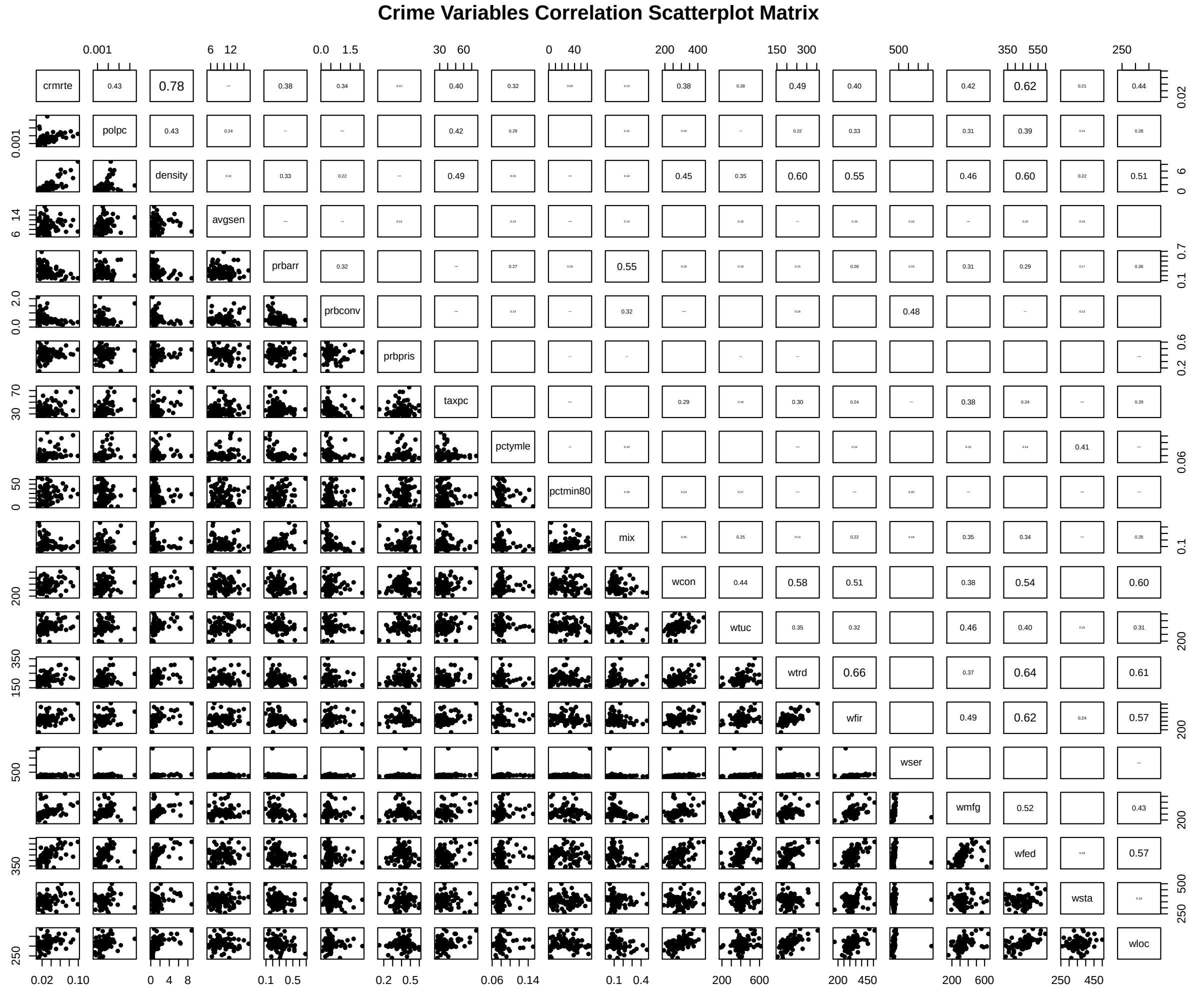
Correlation between all the features: There is a high positive correlation between the crime rate and density (0.73), and a moderate relationship between the crime rate and urban variabless (0.61). There is also a moderately strong relationship between the probability of arrest and the mix (0.57). The west region and percentage minority has the lowest negative correlation (0.63), but given the binary value for west the spearman isn't super meaningful because of the number of ties in the ranking. Density is positively correlated with urban (0.8) and the federal wage(0.58). Amongst the wage values we see many moderate correlations. For example, the wage for construction has a 0.56 correlation with the wage for local gov. workers. Another example is the mild correlation between federal workers and the crime rate (0.59), density(0.58), wage for retail (0.62), wage for the service industry (0.58), wage for Finanace and insurance (0.59), and the wage for local workers (0.54).

```
correlationMatrix = round(cor(raw_data[c(4:26)], use="pairwise.complete.obs"),2)
```

Correlation visual between selected variables (excluding binary variables)

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{   usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y,use="pairwise.complete.obs"))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)}

pairs(~crmrte +polpc + density + avgsen + prbarr +prbconv +prbpris + taxpc + pctymle + pctmin80 + mix +
```


Crime Variables Correlation Scatterplot Matrix

Identifying Key Determinants of Crime

Given our cross-section of the above county-level data, including crime statistics and likelihoods of punishment, we first seek to explain the key causal drivers of crime by hypothesizing one population model. We first conjecture which key variables may have an effect that *also* have the potential to be addressed from a policy-making position. This is the first step in an iterative process to test the robustness of these model effects and the likely biases inherent in these limited data fields and from missing, or omitted, variables that we would like to have.

Before hypthothesizing our population model, we consider the types of variables in the data and which types could be translated into policy terms. First, we have fields related to police presence and the likelihood and severity of punishment for any given crime. These are our most likely levers for which we can estimate the causal relationships and compare the relative effects on crime rates, and politically can be adjusted by increasing police staffing or proposing legislation to improve crime prevention effectiveness and punishment deterents through statute.

Next, we are given locational differences such as population density, which may offer targeted policy prescriptions or resource allocation, as well as additional effects when interacted with the aforementioned police and punishment severity levers. For example, we conjecture that there may be a difference in effect of adding police presence or increasing the severity of punishment for particular crimes in urban versus non-urban areas. We must recognize, unfortunately, that the size of the urban segment is only $n = 8$, which is likely produce too large of standard errors to make confident conclusions, but the direction of coefficients could still lead to more targeted future analysis on urban or non-urban policy.

Lastly, we have a number of economic variables including various wage levels, tax levies and demographics. While these may have explanatory power on our sample of counties, these variables do not have direct links to policy and will be incorporated later when attempting to control for non-crime factors and test for model robustness.

In expectation, we hypothesize that crime prevention proxies such as police per capita, average sentance duration, and likelihood of capture and punishment will have negative relationships with crime rates, while these effect sizes may differ in highly dense, or urban areas. We will then estimate the following population model with their raw, untransformed variables:

$$crmrte = \beta_0 + \beta_1 polpc + \beta_2 avgsen + \beta_3 prbarr + \beta_4 prbconv + \beta_5 prbpris + \beta_6 density + \beta_7 urban + \beta_8 (polpc * urban) + \beta_9 (avgsen * urba$$

```
model1 = lm(crmrte ~polpc + avgsen + prbarr + prbconv + prbpris + density + urban + (polpc * urban) +
              (avgsen * urban) + (prbarr * urban) + (prbconv * urban) + (prbpris * urban), data=raw_data
# summary(model1)
```

From this regression output, we see we have fairly strong explanatory power of our sample, with approximately 75% of the variation in crime rate explained by the model, with several statistically significant relationships. We now test the six key assumptions of classical linear modeling:

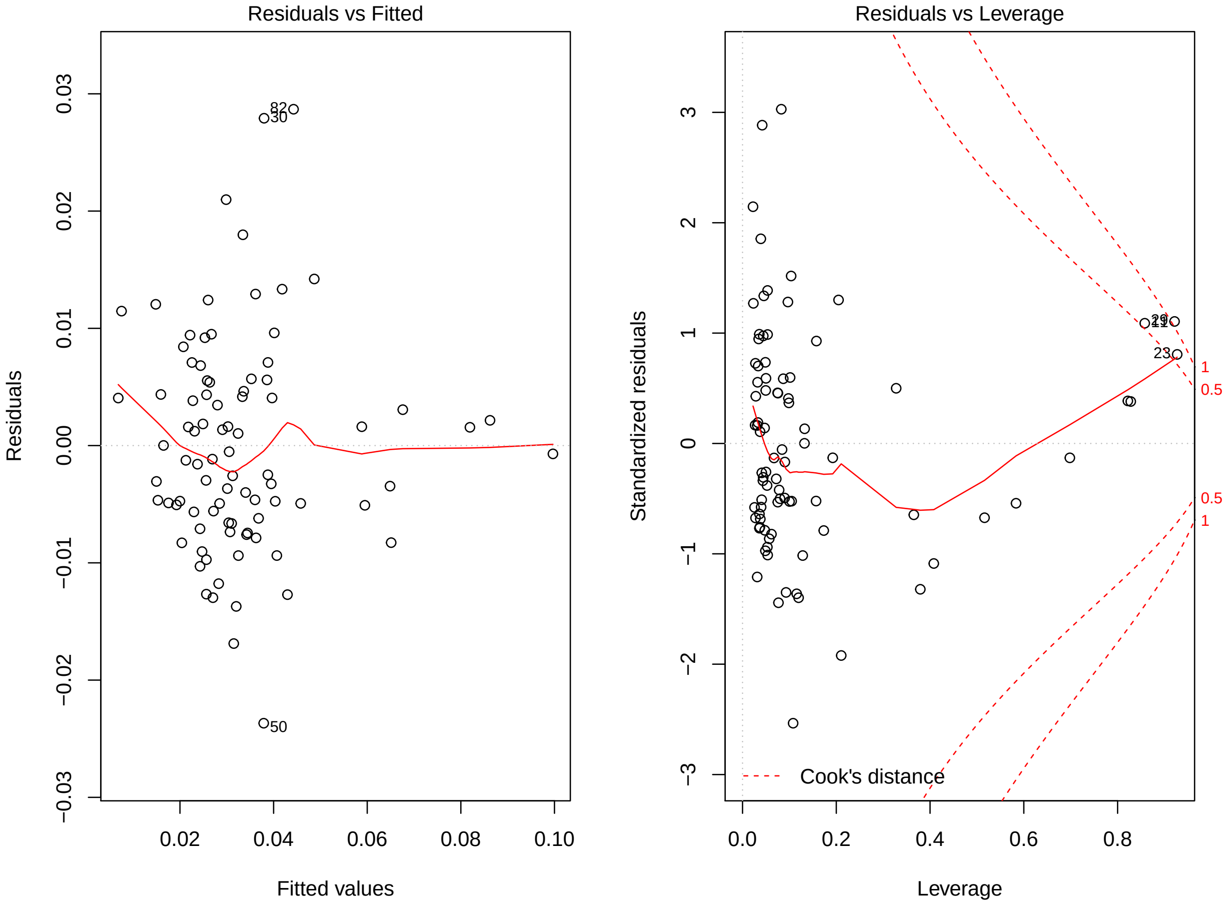
1.) Linearity in parameters 2.) Random sample 3.) No perfect collinearity 4.) Zero conditional mean 5.) Homoskedasticity 6.) Normality of errors

For assumptions #1-#3, we can treat each as satisfied given our population model definition: the coefficients we are estimated are strictly linear and the variables come from a cross-sectional dataset of county level data from 1987; while the sample is not randomly selected, it contains all the available cross-sections and is populated fully. Lastly, we know there is no perfect multicollinearity, as no fields can be derived from the values in another field; this is confirmed by our correlation matrix above where no variables are perfectly correlated, and the trivial fact that the model ran with no variables forced out.

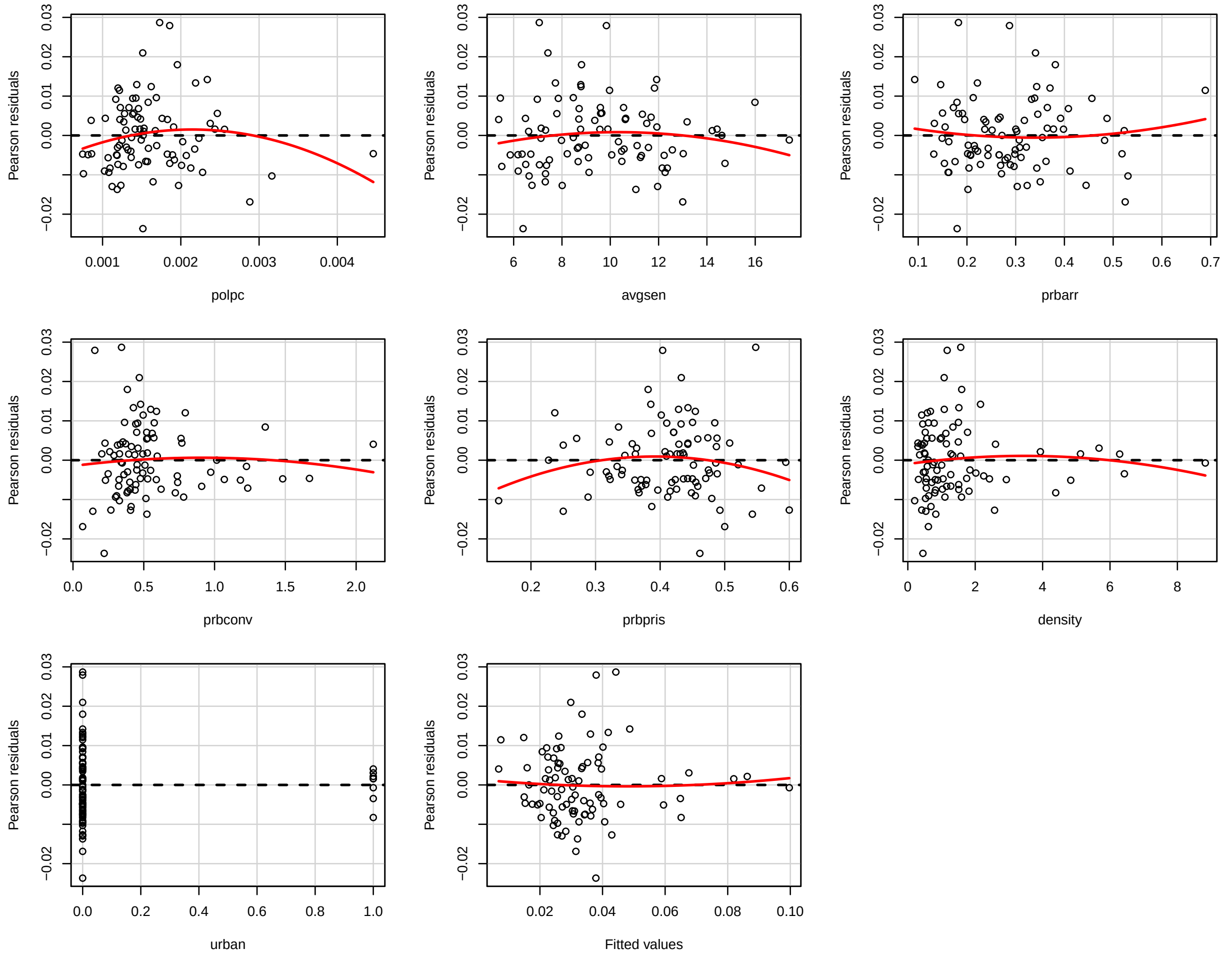
We then must confirm that the expected values of our errors are all zero for any given observation and variable, as well as the homoskedasticity and normality of the errors. Looking at the first plot of residuals versus observed crime rate values, we see the fitted line of expected residuals (red) closely following zero on our residual axis, with no obvious trend as the observed crime rate values increase. Similarly, the Pearson residuals for each independent variable have expectation roughly tracking the zero line, thus not producing evidence that we do not have zero-conditional mean of the errors. However, we do see a couple data points with high leverage with Cook's distances approaching 1 (Chart: Residuals vs Leverage), likely due to the right-skew of both our depend variable and several covariates; we test a model with skewed variables transformed to more normal sample distributions shortly.

Looking at these same plots, we also do not see reason to believe there is heteroskedasticity in these data, where there would be distinct differences in variation of the errors dependendent about the value of the underlying variables; while we do see clustering of the fitted values due to skew of the underlying variables, this does not appear to lead to large differences in variance across the errors, with any trends being pulled by values at the extremes. And lastly, we can look to the Q-Q plot of the standardized residuals against the line that would represent a normal distribution and see that the residuals do appear to be close to normal, with a few points deviating from the line at the tales. [INTERPRET ACF AND PACF CHARTS]

```r
test_resids <- data.frame(model1$fitted.values,model1$residuals, raw_data$urban)
par(mfrow = c(1,2))
plot(model1, which = 1)
plot(model1, which = 5)
```

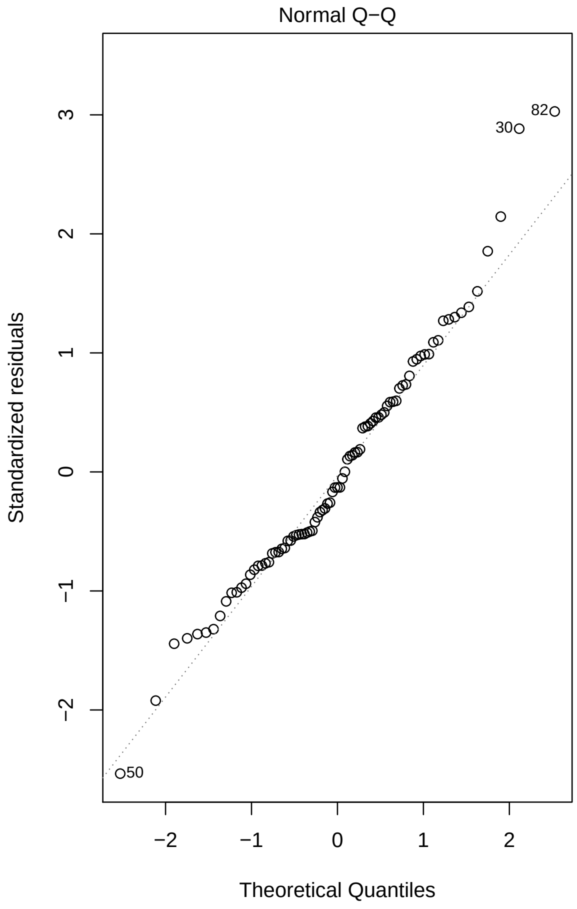


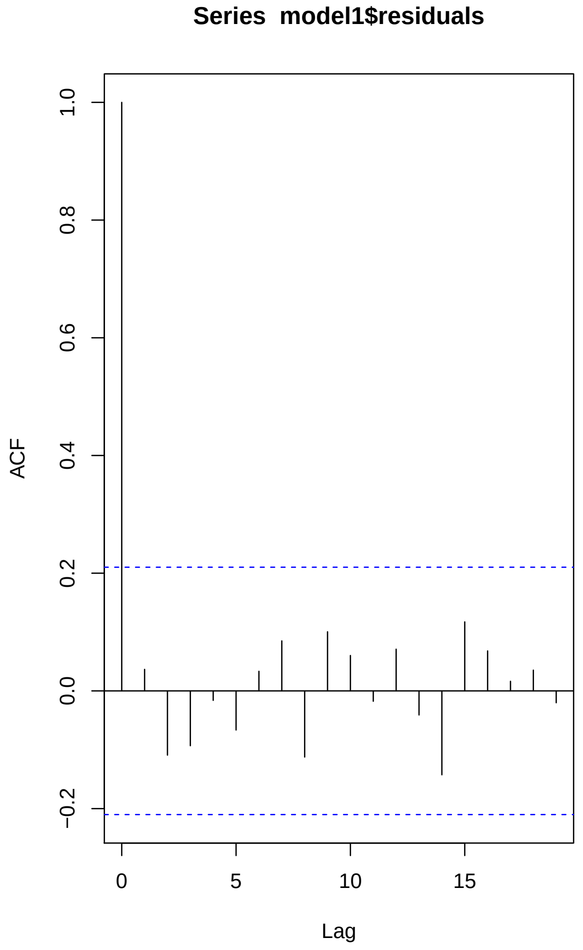
```r
residualPlots(model1)
```



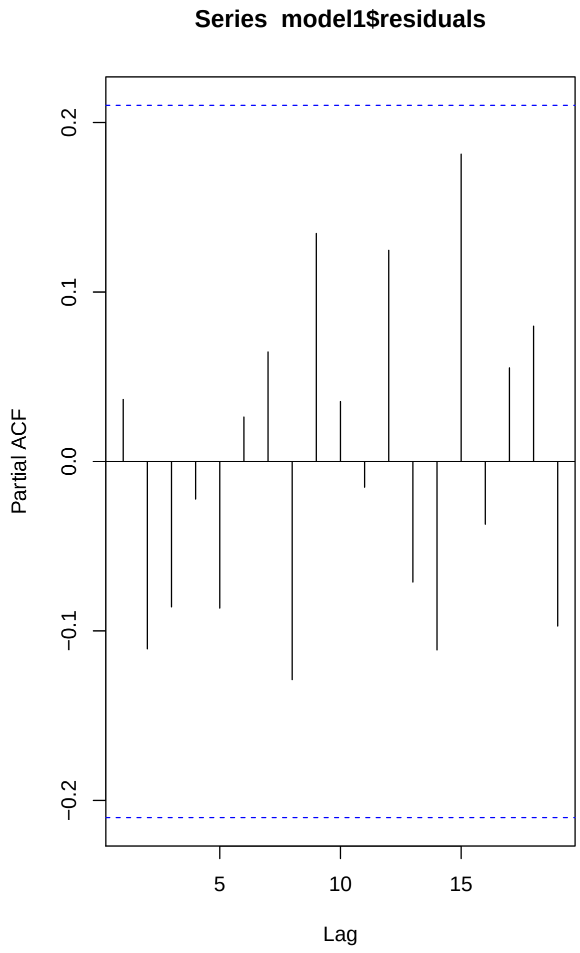
```
##            Test stat Pr(>|t|)
## polpc         -1.948    0.055
## avgsen        -0.979    0.331
## prbarr         0.684    0.496
## prbconv       -0.511    0.611
## prbpris       -1.470    0.146
## density       -1.279    0.205
## urban          0.765    0.446
## Tukey test     1.032    0.302
```

```
plot(model1, which = 2)
acf(model1$residuals)
```





```r
pacf(model1$residuals)
```



While it would appear that interacting the *urban* indicator with our key crime covariates would absorb much of the effects of those variables, we should test the bias that omitting them would introduce as well as whether the explanatory power is signifcantly improved by adding them. By running the model with omitted interaction terms, we see below that the coefficients to our raw numeric variables are for the most part unchanged, while several of our interaction terms have strongly significant effects. Further, by running an F-test on the two models, we do not see evidence that the restricted model has just as much explanatory power as the full model ($p = .08$):

Call: lm(formula = crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density + urban, data = raw_data)

Residuals: Min 1Q Median 3Q Max -0.024036 -0.006474 -0.002591 0.005560 0.028504

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0367082 0.0096197 3.816 0.000268 ***polpc 5.3460641 2.2305449 2.397 0.018904***
avgsen -0.0004364 0.0004502 -0.969 0.335314
prbarr -0.0460002 0.0117684 -3.909 0.000195 *prbconv -0.0152580 0.0036697 -4.158 8.09e-05* **prbpris 0.0115132 0.0139378 0.826 0.411268**
density 0.0056645 0.0013865 4.085 0.000105 * urban 0.0056689 0.0067357 0.842 0.402540
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 0.01021 on 79 degrees of freedom Multiple R-squared: 0.7101, Adjusted R-squared: 0.6845 F-statistic: 27.65 on 7 and 79 DF, p-value: < 2.2e-16

Comparison of Key Crime Determinants Only

```
                                Dependent variable:
                    ---------------------------------------------
                                          crmrte
                             (1)                    (2)
```

polpc 5.461 5.346 (2.177)* (2.231)*
avgsen -0.0004 -0.0004 (0.0004) (0.0005)
prbarr -0.044 -0.046 (0.012)*** (0.012)***
prbconv -0.015 -0.015 (0.004)*** (0.004)***
prbpris 0.014 0.012 (0.014) (0.014)
density 0.006 0.006 (0.002)** (0.001)***
urban 0.590 0.006 (0.207)** (0.007)
polpc:urban 22.181 (20.135)
avgsen:urban -0.015 (0.006)*
prbarr:urban -0.073 (0.099)
prbconv:urban -0.403 (0.143)**
prbpris:urban -0.766 (0.281)**
Constant 0.034 0.037 (0.010)*** (0.010)***

Observations 87 87
R2 0.745 0.710
Adjusted R2 0.704 0.684
Residual Std. Error 0.010 (df = 74) 0.010 (df = 79)
F Statistic 18.040*** (df = 12; 74) 27.650*** (df = 7; 79) ==
Note: *p<0.05;* ***p<0.01;*** p<0.001 Analysis of Variance Table

Model 1: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density + urban + (polpc * urban) + (avgsen * urban) + (prbarr * urban) + (prbconv * urban) + (prbpris * urban) Model 2: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density + urban Res.Df RSS Df Sum of Sq F Pr(>F)
1 74 0.0072373
2 79 0.0082347 -5 -0.00099736 2.0396 0.08283 . — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ‘’ 1

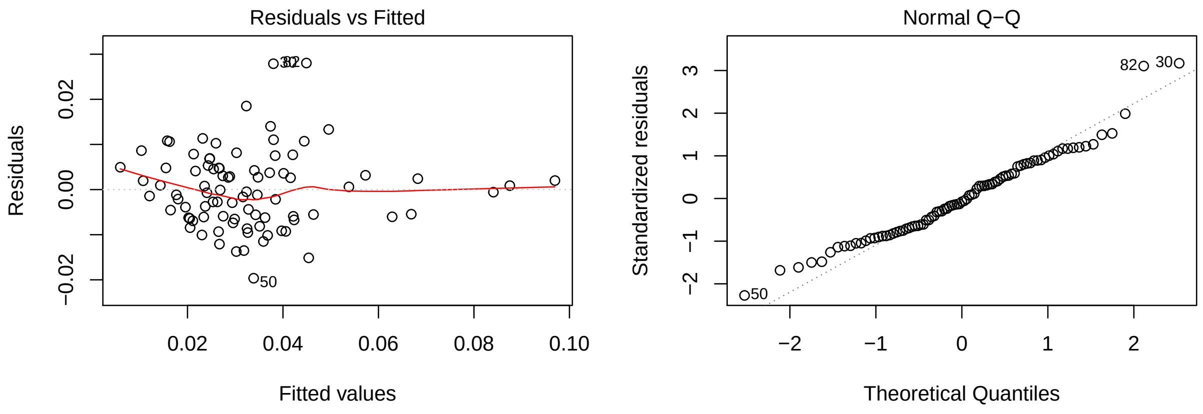
Since we also noted the right skew of several of the independent variables, we consider whether the model is signifcantly improved when transforming these variables with the square root function. In this case, while the ANOVA test does suggest there is modest evidence to reject the null that the full model has the same explanatory power as the model with transformed variables, the improvement does not appear to be strong enough to warrant the interpretation of square-roots of key variables. Further, looking at the Q-Q plot of standardized residuals, we still see similar approximate normality to the original model, still with some deviation from the normal curve at the extremes:

```r
model1_trans = lm(crmrte ~sqrt(polpc) + sqrt(avgsen) + prbarr + sqrt(prbconv) + prbpris + sqrt(density)
                    urban + ((polpc) * urban) + ((avgsen) * urban) + (prbarr * urban) +
                    ((prbconv) * urban) + (prbpris * urban), data=raw_data)
summary(model1_trans)
```

```
##
## Call:
## lm(formula = crmrte ~ sqrt(polpc) + sqrt(avgsen) + prbarr + sqrt(prbconv) +
##     prbpris + sqrt(density) + urban + ((polpc) * urban) + ((avgsen) *
##     urban) + (prbarr * urban) + ((prbconv) * urban) + (prbpris *
##     urban), data = raw_data)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0196475 -0.0061519 -0.0005745  0.0047507  0.0280411
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.115078   0.064632  -1.781  0.07927 .
## sqrt(polpc)     2.972121   1.329752   2.235  0.02856 *
## sqrt(avgsen)    0.041917   0.035147   1.193  0.23700
## prbarr         -0.035492   0.011542  -3.075  0.00299 **
## sqrt(prbconv)   0.022101   0.029573   0.747  0.45733
## prbpris         0.015379   0.013185   1.166  0.24735
## sqrt(density)   0.012659   0.003985   3.176  0.00221 **
## urban           0.562689   0.197268   2.852  0.00568 **
## polpc         -27.411824  15.121081  -1.813  0.07409 .
## avgsen         -0.007162   0.005614  -1.276  0.20617
## prbconv        -0.022658   0.017688  -1.281  0.20437
## urban:polpc    39.555495  18.047658   2.192  0.03168 *
## urban:avgsen   -0.015177   0.005574  -2.723  0.00814 **
## prbarr:urban   -0.126253   0.092637  -1.363  0.17723
## urban:prbconv  -0.428430   0.136854  -3.131  0.00253 **
## prbpris:urban  -0.715574   0.266961  -2.680  0.00913 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009455 on 71 degrees of freedom
## Multiple R-squared:  0.7766, Adjusted R-squared:  0.7294
## F-statistic: 16.45 on 15 and 71 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(1,2))
plot(model1_trans, which = c(1,2))
```



```r
anova(model1, model1_trans)
```

```
## Analysis of Variance Table
##
## Model 1: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density +
##     urban + (polpc * urban) + (avgsen * urban) + (prbarr * urban) +
##     (prbconv * urban) + (prbpris * urban)
## Model 2: crmrte ~ sqrt(polpc) + sqrt(avgsen) + prbarr + sqrt(prbconv) +
##     prbpris + sqrt(density) + urban + ((polpc) * urban) + ((avgsen) *
##     urban) + (prbarr * urban) + ((prbconv) * urban) + (prbpris *
##     urban)
##   Res.Df       RSS Df  Sum of Sq      F  Pr(>F)
## 1     74 0.0072373
## 2     71 0.0063468  3 0.00089049 3.3206 0.02457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ?waldtest
# waldtest(model1, model1_trans, vcov = vcovHC)
```

Summary of Key Crime Determinants Model

Based on these three initial OLS regression models, before controlling for non-crime related control variables, we see several noteworth effects that could influence the direction of public policy. At a high level, we see crime increase as population density increases, and correspondingly, stronger enforcement–or higher likelhood of conviction and prison–in urban counties is associated with the largest decreases in crime rate. On the other hand, we see a moderately significant *positive* relationship between police per capita, which is likely reflective of a dual-causality problem: counties facing high crime rates may be responding by *then* increasing their policing presence. While we do proceed cautiously given the low sample size of our *urban* counties, the most effective allocation of resources to combat high crime rates would be improve the capabilities of law enforcement to bring convictions and put violent criminals in jail. There appears to be less of an effect from either increasing police staff volume, but with some effect from increasing sentence duration. Given these relationships, however, we need to test robustness by controlling for other factors, including demographic and economic variables.

Increasing Explanatory Power with Unbiased Controls

After determining the explantory features that are most malleable to public policy, we look to identify the variables that are more difficult to influence, but nevertheless bare important relationships within the dataset. The distinguishing characteristics of theses explantory features is their ability to increase the amount of variability explained while not introducing bias and violating any of the six aforementioned assumptions.

Model 1 focuses so specifically on crime and policy that it currently does not leverage much background information on each of the counties, and that is the objective of model 2. Starting with a demographic variable, introducing the tax revenue per capita to the model will enable an understanding of the relative resources available within the counties. Traditionally wealthier communities tend to have more resources and less crime and we expect to see that trend in North Carolina. Another demographic feature is the percentage of young males, the classic perpetrator of crime; naturally we anticipate a positive coefficient. Finally we have the bedrock of demographic information, the percentage of the population that's a minority, and we hypothesize that as diversity increases crime will too. A higher level characteristic of the counties is what region they belong to; however without more subject matter expertise its difficult to anticpate a trend. Outside of the demographic variables, we have added the crime mix variable which delineates the more violent crimes (face to face) with the less interactive counterpart (other – think theft). The crime mix is difficult to control with public policy because its a facet of human nature, and didn't make it into model 1 because of that, but it could provide significant insight into crime rates. Typically cities have a reputation for more violent crimes and we expect to an interaction between the urban indicator and the mix variable. ###more on interactions…?

$$crmrte = \beta_0+\beta_1 polpc+\beta_2 avgsen+\beta_3 prbarr+\beta_4 prbconv+\beta_5 prbpris+\beta_6 density+\beta_7 urban+\beta_8(polpc*urban)+\beta_9(avgsen*urb$$

```
model2 = lm(crmrte ~polpc + avgsen + prbarr + prbconv + prbpris + density + urban + (polpc * urban) +
              (avgsen * urban) + (prbarr * urban) + (prbconv * urban) + (prbpris * urban) + taxpc + wes
model2
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris +
##     density + urban + (polpc * urban) + (avgsen * urban) + (prbarr *
##     urban) + (prbconv * urban) + (prbpris * urban) + taxpc +
##     west + central + pctymle + pctmin80 + mix + missingLabel,
##     data = raw_data)
##
## Coefficients:
##   (Intercept)          polpc         avgsen         prbarr        prbconv
##     3.211e-02      5.897e+00     -3.009e-04     -4.882e-02     -1.881e-02
##       prbpris        density          urban          taxpc           west
##     6.539e-03      7.239e-03      5.156e-01     -6.121e-05     -1.083e-03
##       central        pctymle       pctmin80            mix   missingLabel
##    -2.081e-03      2.176e-02      3.714e-04     -1.943e-02             NA
##   polpc:urban   avgsen:urban   prbarr:urban  prbconv:urban  prbpris:urban
##     3.575e+00     -1.342e-02      2.988e-02     -3.188e-01     -6.600e-01
```

```
#summary(model2)
```

Adding interactions for model 2:

```
model2_rest <- lm(crmrte ~polpc + avgsen + prbarr + prbconv + prbpris + density + urban + (polpc * urba
              (avgsen * urban) + (prbarr * urban) + (prbconv * urban) + (prbpris * urban) + taxpc + wes
```

Comparing Model versions: The control variables added to model 2 clearly enhanced the model. The output below shows an increase in adjusted R-squared from 0.70 to 0.82 and the anova test confirms the modesl is statistically significant so we reject the null hypthesis that the models are the same. Finally, model 2 also introduces the minority rate and it is highly statistically significant with a tiny p-value (although the power is very low). The results from the 2nd iteration of model 2 (includes the interactions) compared with the first version of model 2 are strikingly different. The model adjusted R-square moves less than 1/100th of a point and the anova test confirms the models are not significantly different.

```
#summary(model2_rest)
stargazer(model1, model2, model2_rest, type = "text", report = "vcs*",
          header = FALSE,
          title = "Comparison of Key Crime Determinants Only",
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Comparison of Key Crime Determinants Only
## ===========================================================================================
##                                                Dependent variable:
##                      -----------------------------------------------------------------------
##                                                      crmrte
##                               (1)                     (2)                     (3)
## -------------------------------------------------------------------------------------------
## polpc                        5.461                   5.897                   6.120
##                            (2.177)*                (2.090)**               (2.684)*
##
## avgsen                      -0.0004                 -0.0003                 -0.0002
##                            (0.0004)                (0.0004)                (0.0004)
```

```
##
## prbarr                      -0.044                   -0.049                  -0.049
##                           (0.012)***               (0.011)***              (0.011)***
##
## prbconv                     -0.015                   -0.019                  -0.018
##                           (0.004)***               (0.003)***              (0.004)***
##
## prbpris                      0.014                    0.007                  -0.0004
##                             (0.014)                  (0.011)                 (0.012)
##
## density                      0.006                    0.007                   0.007
##                            (0.002)**                (0.002)***              (0.002)***
##
## urban                        0.590                    0.516                   0.360
##                            (0.207)**                 (0.173)**               (0.218)
##
## taxpc                                                -0.0001                 -0.0001
##                                                     (0.0001)                (0.0001)
##
## west                                                 -0.001                   0.012
##                                                      (0.004)                 (0.014)
##
## central                                              -0.002                  -0.012
##                                                      (0.003)                 (0.009)
##
## pctymle                                               0.022                   0.093
##                                                      (0.067)                 (0.071)
##
## pctmin80                                             0.0004                  0.0004
##                                                    (0.0001)***             (0.0001)***
##
## mix                                                  -0.019                  -0.036
##                                                      (0.014)                 (0.021)
##
## missingLabel
##
##
## polpc:urban                 22.181                    3.575                  10.582
##                            (20.135)                 (16.267)                (23.808)
##
## avgsen:urban                -0.015                   -0.013                  -0.013
##                            (0.006)*                 (0.005)**               (0.008)
##
## prbarr:urban                -0.073                    0.030                   0.008
##                             (0.099)                  (0.084)                 (0.121)
##
## prbconv:urban               -0.403                   -0.319                  -0.232
##                            (0.143)**                (0.119)**               (0.133)
##
## prbpris:urban               -0.766                   -0.660                  -0.242
##                            (0.281)**                (0.234)**               (0.337)
##
## polpc:west                                                                   -8.294
##                                                                              (5.170)
```

```
##
## polpc:central                                                                4.090
##                                                                             (4.610)
##
## central:mix                                                                  0.035
##                                                                             (0.031)
##
## west:mix                                                                     0.042
##                                                                             (0.026)
##
## urban:mix                                                                   -0.266
##                                                                             (0.154)
##
## urban:taxpc                                                                 -0.001
##                                                                             (0.001)
##
## taxpc:west                                                                  -0.0001
##                                                                            (0.0004)
##
## Constant                     0.034                   0.032                   0.031
##                            (0.010)***               (0.011)**               (0.011)**
##
## -------------------------------------------------------------------------------------------
## Observations                  87                      87                      87
## R2                           0.745                   0.861                   0.880
## Adjusted R2                  0.704                   0.824                   0.831
## Residual Std. Error     0.010 (df = 74)         0.008 (df = 68)         0.007 (df = 61)
## F Statistic         18.040*** (df = 12; 74) 23.332*** (df = 18; 68) 17.963*** (df = 25; 61)
## ===========================================================================================
## Note:                                                         *p<0.05; **p<0.01; ***p<0.001
```

```r
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density +
##     urban + (polpc * urban) + (avgsen * urban) + (prbarr * urban) +
##     (prbconv * urban) + (prbpris * urban)
## Model 2: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density +
##     urban + (polpc * urban) + (avgsen * urban) + (prbarr * urban) +
##     (prbconv * urban) + (prbpris * urban) + taxpc + west + central +
##     pctymle + pctmin80 + mix + missingLabel
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     74 0.0072373
## 2     68 0.0039589  6 0.0032784 9.3854 1.711e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

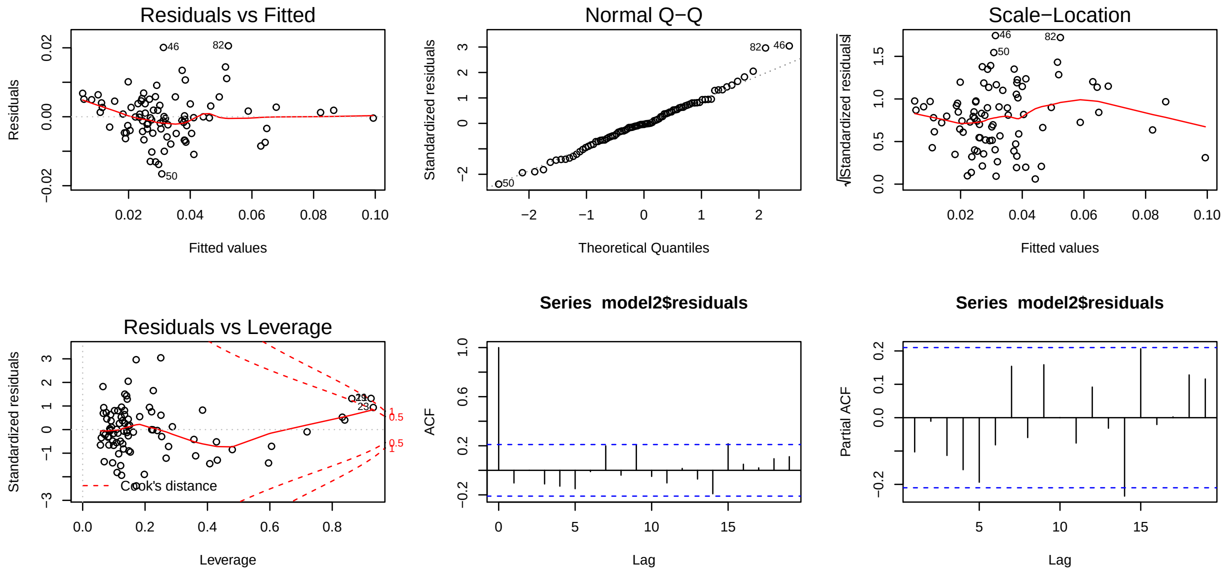
```r
anova(model2, model2_rest )
```

```
## Analysis of Variance Table
##
## Model 1: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density +
##     urban + (polpc * urban) + (avgsen * urban) + (prbarr * urban) +
##     (prbconv * urban) + (prbpris * urban) + taxpc + west + central +
##     pctymle + pctmin80 + mix + missingLabel
## Model 2: crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris + density +
##     urban + (polpc * urban) + (avgsen * urban) + (prbarr * urban) +
##     (prbconv * urban) + (prbpris * urban) + taxpc + west + central +
##     pctymle + pctmin80 + mix + (west * polpc) + (central * polpc) +
##     (central * mix) + (west * mix) + (urban * mix) + (urban *
##     taxpc) + (taxpc * west)
##   Res.Df       RSS Df  Sum of Sq     F Pr(>F)
## 1     68 0.0039589
## 2     61 0.0033974  7 0.00056143 1.44 0.2061
```

A quick note on the assumptions of the versions on model2: After running the diagnostic plots, we have additional evidence to reject the version of model2 that has the interactions. Introducing interactions causes the model to have heterosckedacity (shown by a clear upward trend in the scale-location plot), reduces the normallity in the Q-plot, and has 8 points with leverage above 1. Given the number of variables in the model it must be overfitting. The more parsimonious version of model 2 does not engender violated assumptions and showed results similar to model1 where the assumptions are not violated.

The results of the model comparison and the assumptions analysis both concur that the simpler version of model2 should be the go forward model.

```r
#light version of model2
par(mfrow = c(2,3))
plot(model2, which = c(1,2,3,5))
acf(model2$residuals)
pacf(model2$residuals)
```

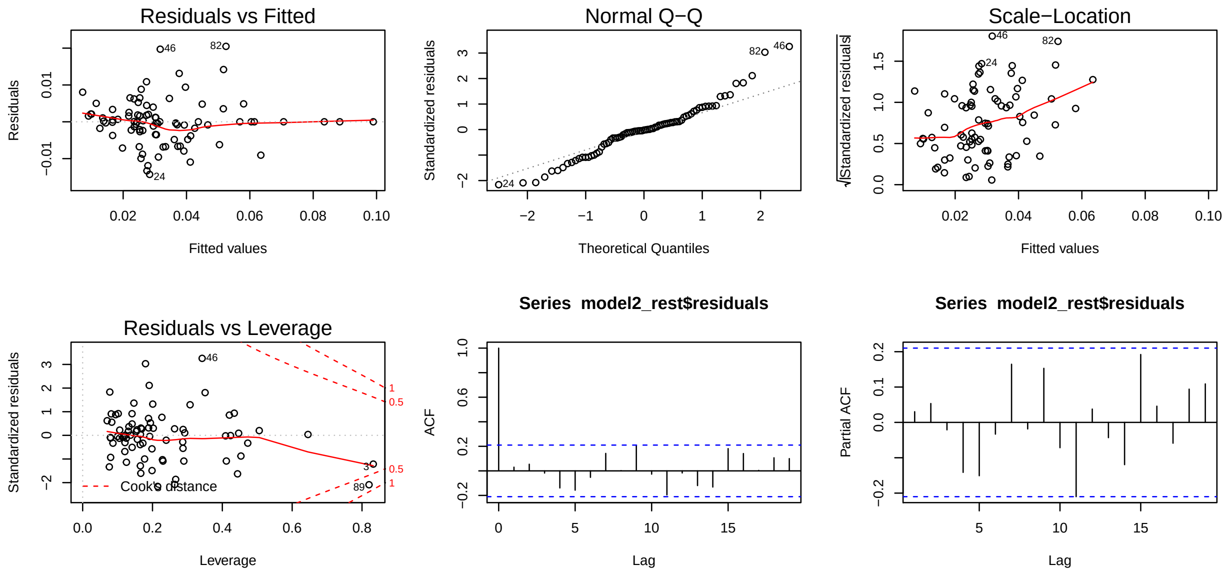


```r
#model 2 (interactions)
par(mfrow = c(2,3))
plot(model2_rest, which = c(1,2,3,5))
```

```
## Warning: not plotting observations with leverage one:
##   11, 23, 28, 30, 35, 51, 55, 80

## Warning: not plotting observations with leverage one:
##   11, 23, 28, 30, 35, 51, 55, 80
```

```
#residualPlots(model2)
acf(model2_rest$residuals)
pacf(model2_rest$residuals)
```



```
#takes forever to run because of the number of variables
#residualPlots(model2)
```

Testing Robustness with All Covariates Included

Having focused on the key determinants of crime from our available data while controlling for influential regional, demographic and economic factors, we now turn to including the remainder of our available variables in order to both maximize explanatory power of our model while testing whether the robustness of our original crime determinants hold, or whether there are strong biases generated by these other variables or omitted variables.

```
model3 <- lm(crmrte ~polpc + avgsen + prbarr + prbconv + prbpris + density + urban + (polpc * urban) +
              (prbconv * urban) + (prbpris * urban) + taxpc + west+ central + pctymle + pctmin80 + mix +
              wcon +wtuc + wtrd + wfir+ wser +wmfg +wfed +wsta +wloc, data=raw_data)
# summary(model3)
stargazer(model3, model1, type = "text", report = "vcs*",
          header = FALSE,
          title = "Comparison of Full Model to Key Crime Determinants Only",
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Comparison of Full Model to Key Crime Determinants Only
## ==================================================================
##                                     Dependent variable:
##                         ------------------------------------------
##                                           crmrte
##                                  (1)                     (2)
## ------------------------------------------------------------------
## polpc                           4.578                   5.461
##                            (2.213)*                (2.177)*
##
## avgsen                      -0.0004                 -0.0004
##                            (0.0004)                (0.0004)
##
## prbarr                      -0.050                  -0.044
##                           (0.011)***              (0.012)***
##
## prbconv                     -0.016                  -0.015
##                           (0.004)***              (0.004)***
##
## prbpris                      0.002                   0.014
##                             (0.011)                 (0.014)
##
## density                      0.006                   0.006
##                            (0.002)**               (0.002)**
##
## urban                        0.458                   0.590
##                            (0.175)*                (0.207)**
##
## taxpc                       -0.0001
##                            (0.0001)
##
## west                         0.001
##                             (0.004)
##
## central                     -0.001
##                             (0.003)
##
## pctymle                      0.085
##                             (0.075)
##
## pctmin80                    0.0004
##                           (0.0001)***
##
## mix                         -0.006
##                             (0.015)
##
## wcon                        0.00002
##                            (0.00003)
##
## wtuc                        0.00000
##                            (0.00001)
##
## wtrd                        0.00001
##                            (0.00004)
##
## wfir                       -0.00004
##                            (0.00003)
##
## wser                       -0.00000
##                            (0.00001)
##
## wmfg                       -0.00000
##                            (0.00001)
##
## wfed                        0.0001
##                           (0.00003)*
##
## wsta                       -0.00003
##                            (0.00003)
##
## wloc                        0.00000
##                            (0.00005)
##
## polpc:urban                 10.752                  22.181
##                            (17.190)                (20.135)
##
## avgsen:urban                -0.013                  -0.015
##                            (0.005)*                (0.006)*
##
## prbarr:urban                 0.001                  -0.073
##                             (0.097)                 (0.099)
##
## prbconv:urban               -0.306                  -0.403
##                            (0.123)*                (0.143)**
##
## prbpris:urban               -0.564                  -0.766
##                            (0.237)*                (0.281)**
##
## Constant                     0.015                   0.034
##                             (0.018)               (0.010)***
##
## -------------------------------------------------------------------
## Observations                  86                      87
## R2                           0.885                   0.745
## Adjusted R2                  0.832                   0.704
## Residual Std. Error     0.007 (df = 58)         0.010 (df = 74)
## F Statistic         16.559*** (df = 27; 58) 18.040*** (df = 12; 74)
## ===================================================================
## Note:                                 *p<0.05; **p<0.01; ***p<0.001
```

notes on robustness and limitations

Thoughts on potentially missing variables/endogeneity: county specifics: -Education levels -Age distribution (older communities?) -Crime reporting rate (if the data isn't captured it LOOKS safer) -Recidivism rate vs rehabilitation programs quality -divorce rates/single parent homes

Crime specifics -Only a few industries are represented for wage (overall measure would be helpful) -Type of crime (property vs violent) -violent: murder, rape, robbery, assult -property: burgulary, theft -misdemenor: vandalism, disorderly conduct, traffic violations -other: drugs/alcohol -X & Y coordinates of crimes (some streets are safer than others) -Crime time (certian crimes have higher rate during the day vs night) -Seasonality of crimes (summer vs winter)

Conclusion