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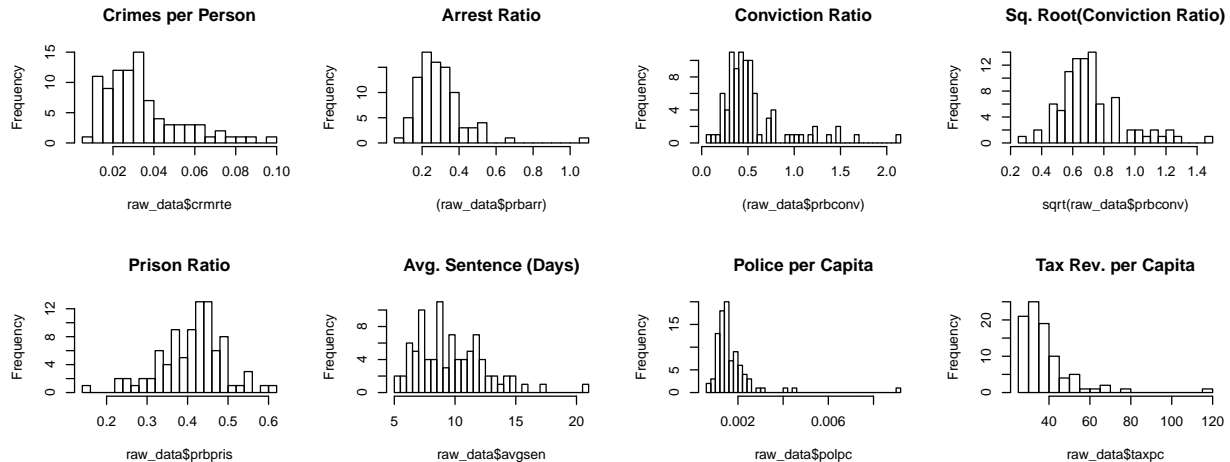
```r
hist(raw_data$density, breaks = 30, main = "People per Sq. Mile")
hist(raw_data$pctmin80, breaks = 30, main = "Percent Minority (1980)")
hist(raw_data$mix, breaks = 30, main = "Offense Mix")
hist(raw_data$pctymle, breaks = 30, main = "Percent Young Male")
```
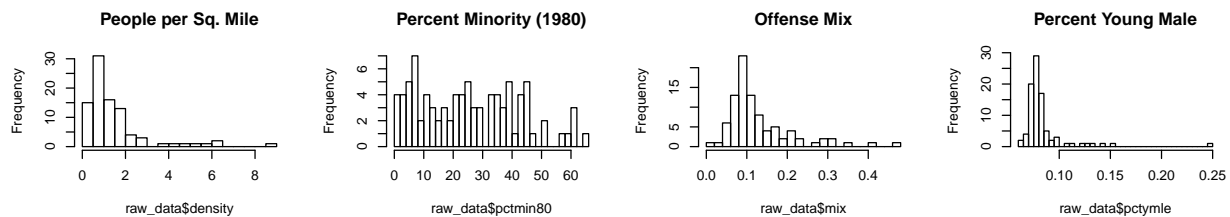

There are a handful of instances of very large outliers that tend to be associated with very small population densities (i.e., low population and/or high land area), which include: police per capita (0.9 compared to 0.5), tax revenue per capita (119, nearly double the next highest value), and percent 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.

```r
#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 sentence 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.

```
#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")
```
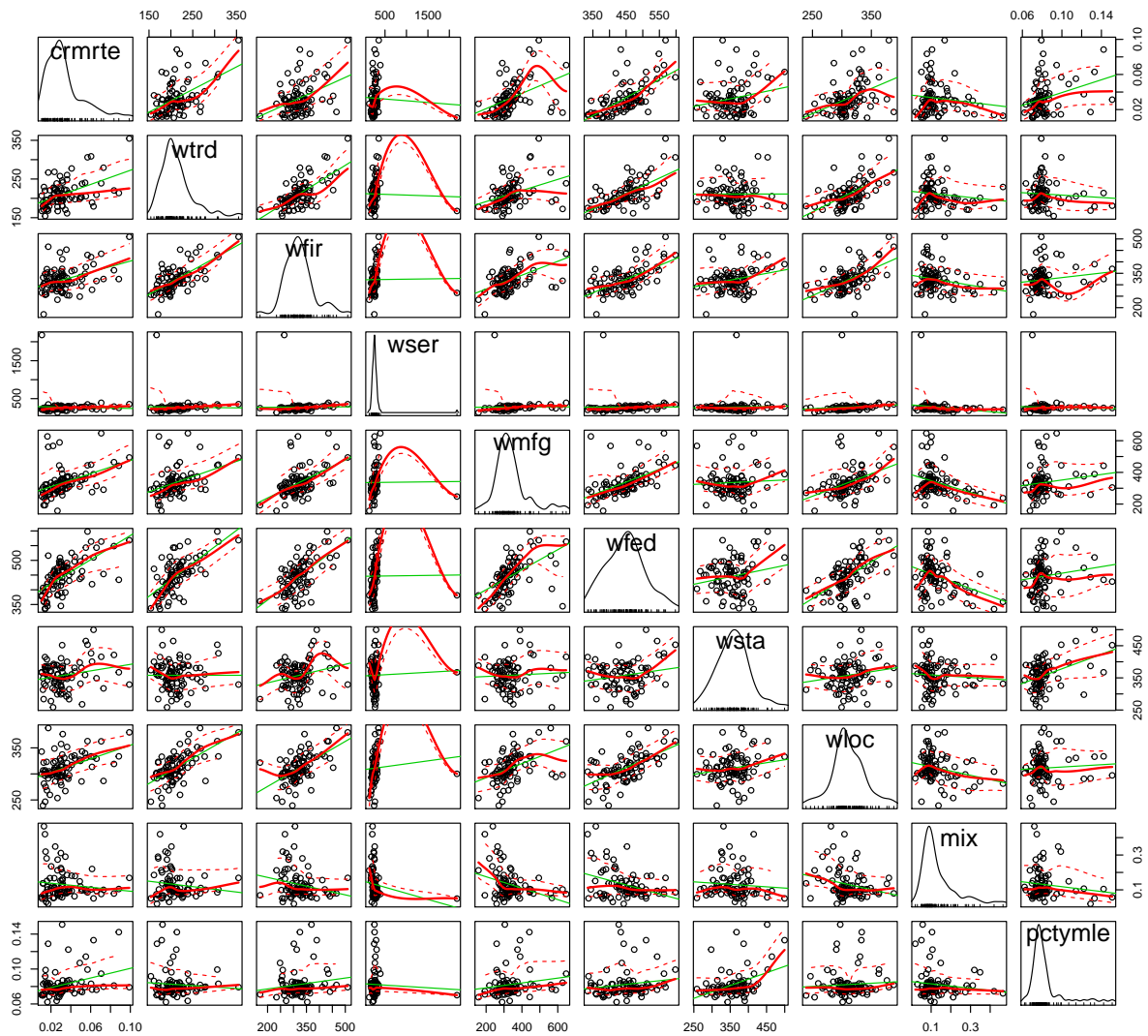


```
raw_data$wser[84] = NA
```

Scatter plot matrix Part I

```
scatterplotMatrix(raw_data[c(4:11,15:17)])
```
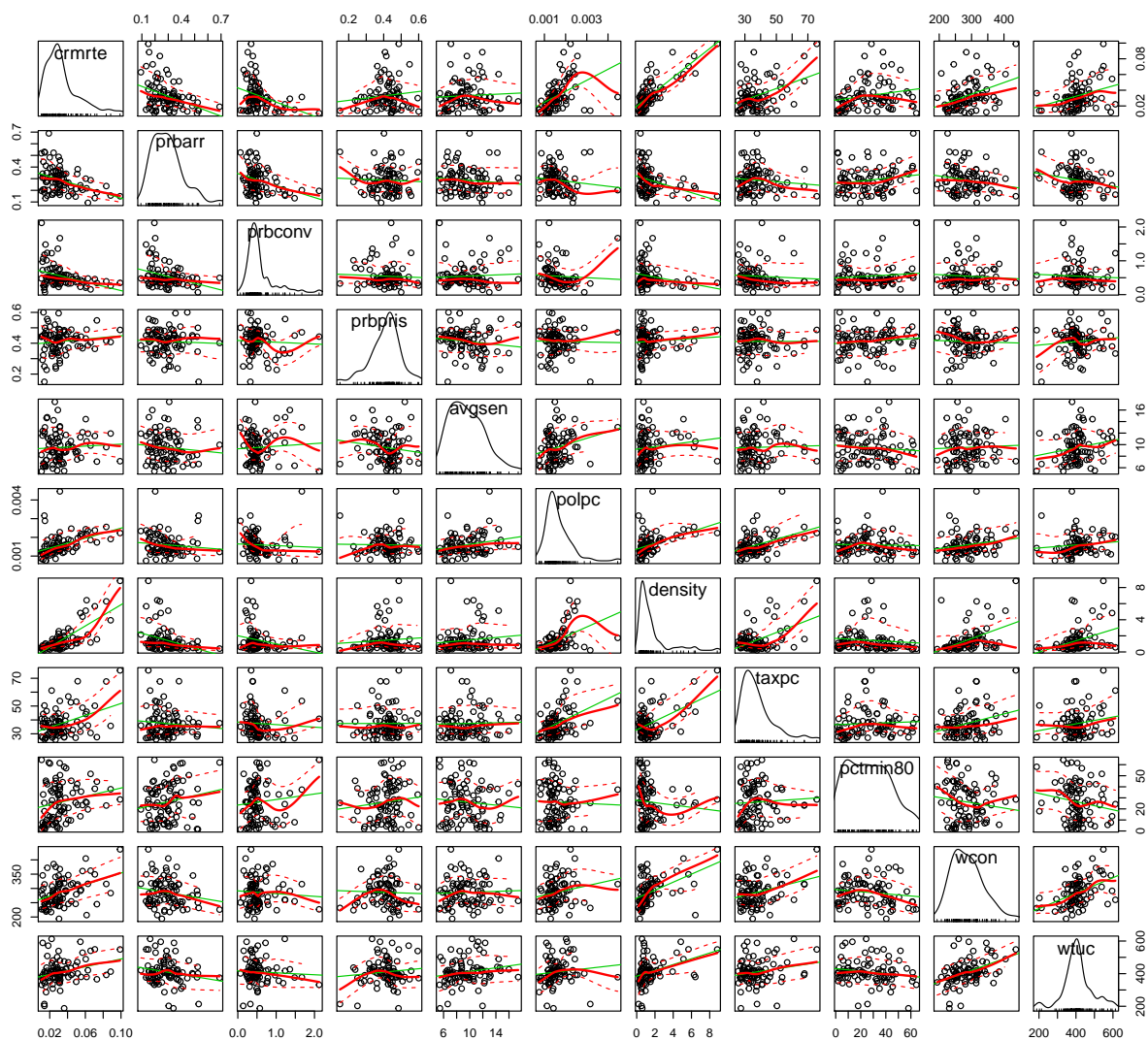
Scatterplot matrix Part II

```r
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
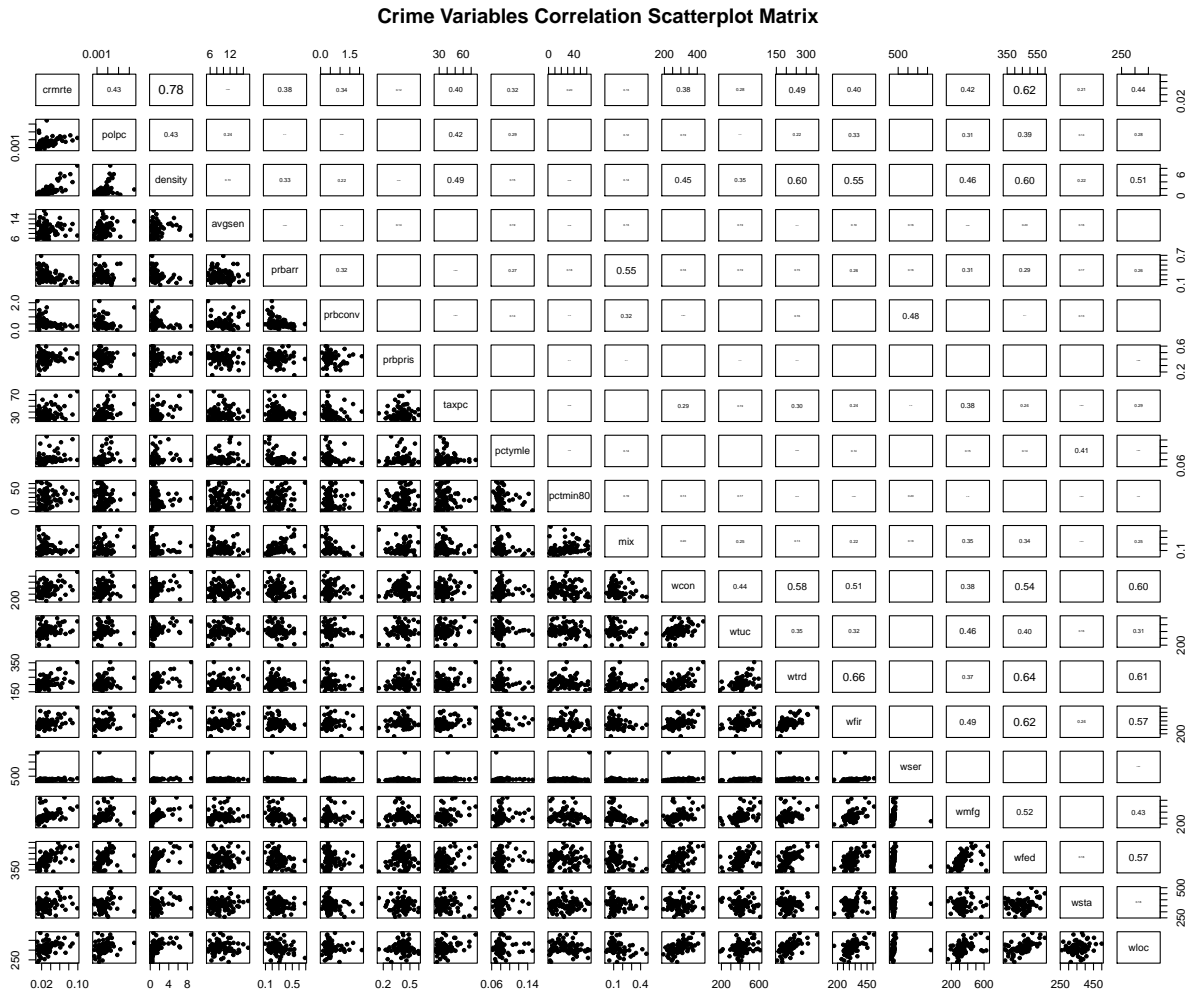
Identifying Key Determinants of Crime

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 sentence 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:

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

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

```
##
## Call:
## lm(formula = crmrte ~ polpc + avgsen + prbarr + prbconv + prbpris +
##     density + urban + (polpc * urban) + (avgsen * urban) + (prbarr *
##     urban) + (prbconv * urban) + (prbpris * urban), data = raw_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.023678 -0.005930 -0.001156  0.005014  0.028685
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0344497  0.0095740   3.598 0.000576 ***
## polpc          5.4606433  2.1765696   2.509 0.014303 *
## avgsen        -0.0003835  0.0004477  -0.857 0.394394
## prbarr        -0.0436029  0.0117272  -3.718 0.000388 ***
## prbconv       -0.0150412  0.0036207  -4.154 8.66e-05 ***
## prbpris        0.0135641  0.0136518   0.994 0.323665
## density        0.0055814  0.0016966   3.290 0.001537 **
## urban          0.5903043  0.2069074   2.853 0.005613 **
## polpc:urban   22.1806860 20.1352723   1.102 0.274213
## avgsen:urban  -0.0146873  0.0058360  -2.517 0.014014 *
## prbarr:urban  -0.0731356  0.0991793  -0.737 0.463206
## prbconv:urban -0.4027015  0.1427655  -2.821 0.006147 **
## prbpris:urban -0.7661171  0.2806822  -2.729 0.007922 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7

```
##
## Residual standard error: 0.009889 on 74 degrees of freedom
## Multiple R-squared:  0.7452, Adjusted R-squared:  0.7039
## F-statistic: 18.04 on 12 and 74 DF,  p-value: < 2.2e-16
```

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 dependent 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 differences in variance across the errors. 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]

```
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)
```
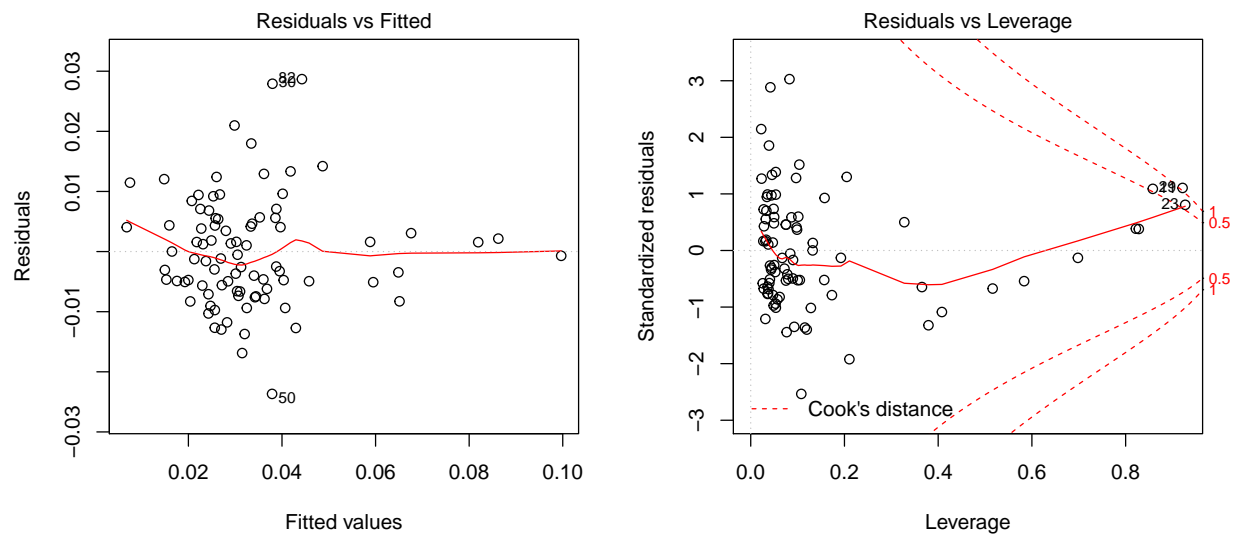
Residuals vs Fitted

Residuals vs Leverage

```r
residualPlots(model1)
```
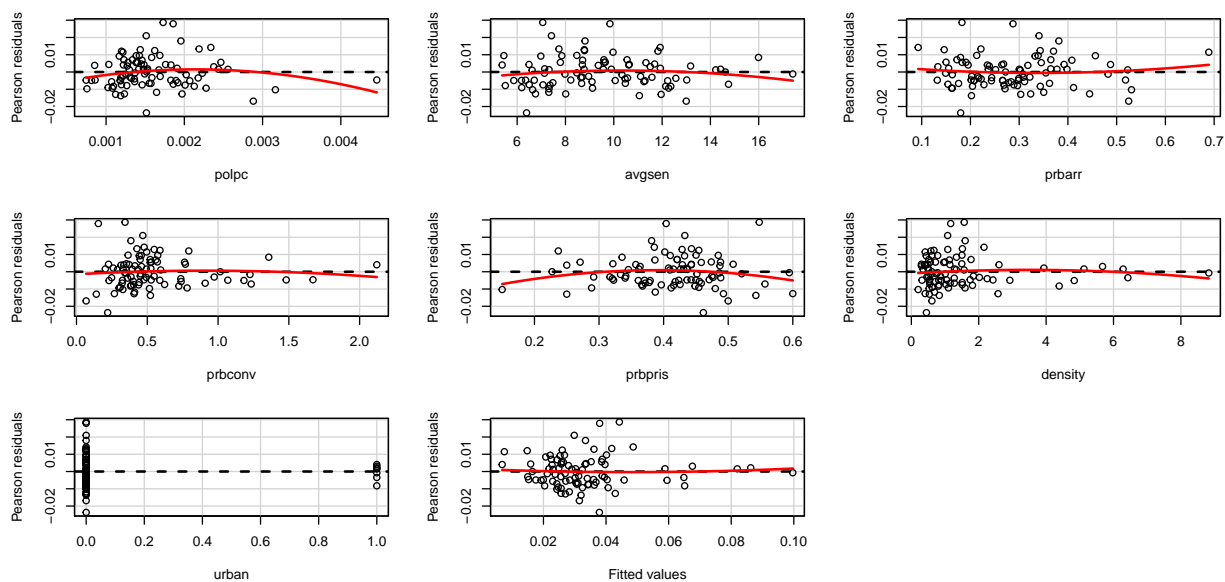


```
##              Test stat Pr(>|t|)
## polpc          -1.948    0.055
## avgsen         -0.979    0.331
## prbarr          0.684    0.496
## prbconv        -0.511    0.611
## prbpris        -1.470    0.146
## density        -1.279    0.205
## urban           0.765    0.446
## Tukey test      1.032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

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

Normal Q–Q

Series model1$residuals
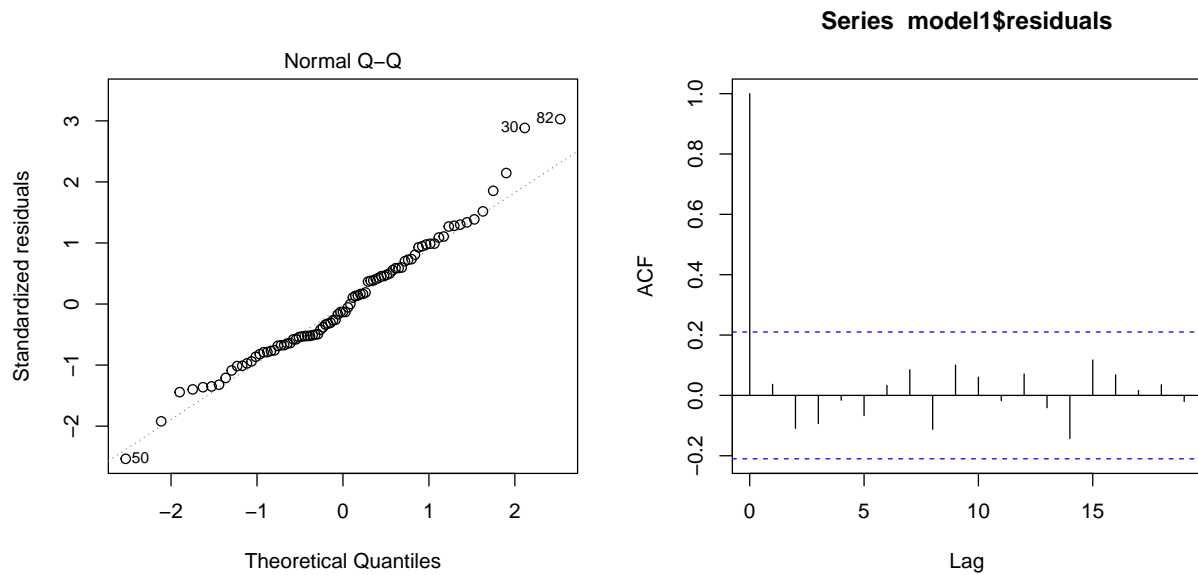
pacf(model1$residuals)

Series model1$residuals

While it would appear that interacting the *urban* indicator with our key crime covariates would absorb much of the effects of those v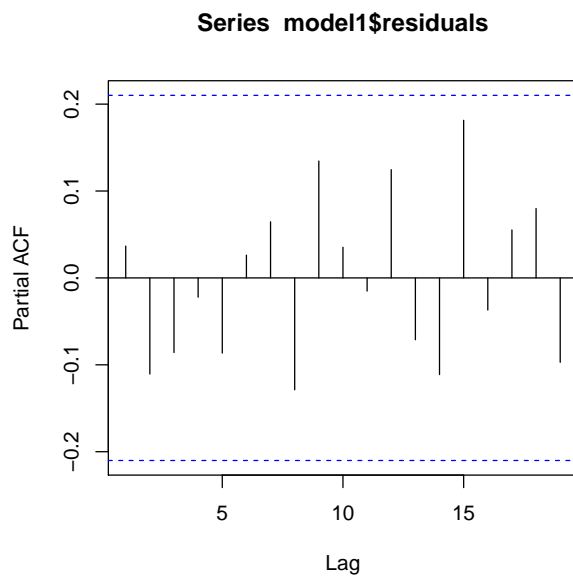ariables, we should test the bias that omitting them would introduce as well as whether the explanatory power is signifcantly improved by adding them. By running the model with omitted interaction terms, we see below that the coefficients to our raw numeric variables are for the most part unchanged, 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

11

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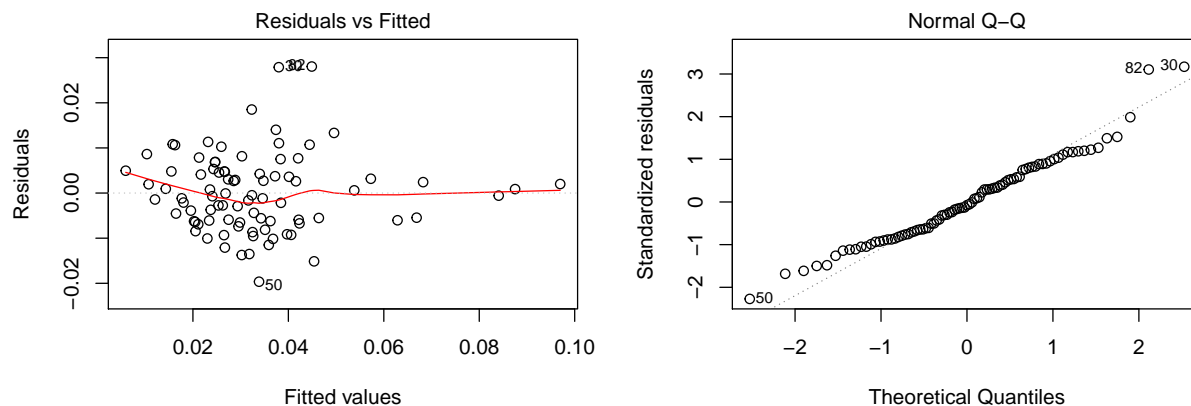
```r
anova(model1, model1_trans)
```

```
## 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
```

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

some evidence to reject the null hypothesis that model 1 has same explanatory power as model with transformed variables

Model 2 - increases accuracy without bias (controls)

Model 2: increase accuracy without increasing bias - taxpc, control for income (sq rt?) - west, central (interact each with urban? not enough obs) - pctymle, control for demographics - pctmin80 - mix? check if adds accuracy or increases bias

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

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

```
##
## Call:
## lm(formula = crmrte ~ polpc + density + (polpc * urban) + avgsen +
```

```
##      prbarr + prbconv + prbpris + taxpc + west + central + pctymle +
##      pctmin80 + mix, data = raw_data)
##
## Coefficients:
## (Intercept)         polpc        density         urban        avgsen
##    2.950e-02      5.595e+00      7.416e-03      2.691e-02     -4.427e-04
##        prbarr        prbconv        prbpris         taxpc          west
##    -4.546e-02     -1.792e-02      7.006e-03     -3.970e-05     -3.219e-03
##       central        pctymle       pctmin80           mix    polpc:urban
##    -3.902e-03      7.158e-02      3.349e-04     -2.074e-02     -1.337e+01
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + density + (polpc * urban) + avgsen +
##      prbarr + prbconv + prbpris + taxpc + west + central + pctymle +
##      pctmin80 + mix, data = raw_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.016853 -0.004179 -0.000735  0.003684  0.022270
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.950e-02  1.067e-02   2.765 0.007225 **
## polpc        5.595e+00  2.030e+00   2.756 0.007406 **
## density      7.416e-03  1.282e-03   5.786 1.73e-07 ***
## urban        2.691e-02  2.762e-02   0.974 0.333217
## avgsen      -4.427e-04  3.683e-04  -1.202 0.233321
## prbarr      -4.546e-02  1.112e-02  -4.088 0.000112 ***
## prbconv     -1.792e-02  3.135e-03  -5.715 2.32e-07 ***
## prbpris      7.006e-03  1.116e-02   0.628 0.532110
## taxpc       -3.970e-05  1.112e-04  -0.357 0.722177
## west        -3.219e-03  3.789e-03  -0.849 0.398439
## central     -3.902e-03  2.463e-03  -1.584 0.117527
## pctymle      7.158e-02  6.339e-02   1.129 0.262578
## pctmin80     3.349e-04  8.823e-05   3.796 0.000304 ***
## mix         -2.074e-02  1.341e-02  -1.547 0.126282
## polpc:urban -1.337e+01  1.303e+01  -1.026 0.308506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007858 on 72 degrees of freedom
## Multiple R-squared:  0.8435, Adjusted R-squared:  0.8131
## F-statistic: 27.72 on 14 and 72 DF,  p-value: < 2.2e-16
```

Model 3 - robustness

```
model3 = lm(crmrte ~ polpc + density + (polpc * urban) + avgsen + prbarr +prbconv +prbpris + taxpc + wes
model3
```

```
##
## Call:
```

```
## lm(formula = crmrte ~ polpc + density + (polpc * urban) + avgsen +
##     prbarr + prbconv + prbpris + taxpc + west + central + pctymle +
##     pctmin80 + mix + wcon + wtuc + wtrd + wfir + wser + wmfg +
##     wfed + wsta + wloc, data = raw_data)
##
## Coefficients:
## (Intercept)        polpc        density         urban         avgsen
##   1.538e-02     4.083e+00      6.474e-03     2.082e-02     -4.720e-04
##       prbarr       prbconv        prbpris        taxpc           west
##   -4.815e-02    -1.560e-02      3.589e-03    -4.098e-05     -4.157e-04
##      central       pctymle       pctmin80          mix           wcon
##   -3.343e-03     1.386e-01      3.701e-04    -5.977e-03      2.378e-05
##         wtuc          wtrd           wfir         wser           wmfg
##   -1.549e-06     2.204e-05     -3.937e-05    -2.482e-06      7.268e-07
##         wfed          wsta           wloc  polpc:urban
##    5.592e-05    -3.551e-05     -8.592e-07    -9.294e+00
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + density + (polpc * urban) + avgsen +
##     prbarr + prbconv + prbpris + taxpc + west + central + pctymle +
##     pctmin80 + mix + wcon + wtuc + wtrd + wfir + wser + wmfg +
##     wfed + wsta + wloc, data = raw_data)
##
## Residuals:
##        Min         1Q      Median         3Q        Max
## -0.0124783 -0.0047210 -0.0008127  0.0043319  0.0211922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.538e-02  1.835e-02   0.838 0.404976
## polpc        4.083e+00  2.163e+00   1.887 0.063826 .
## density      6.474e-03  1.454e-03   4.453 3.60e-05 ***
## urban        2.082e-02  2.897e-02   0.719 0.475113
## avgsen      -4.720e-04  3.931e-04  -1.201 0.234392
## prbarr      -4.815e-02  1.110e-02  -4.338 5.39e-05 ***
## prbconv     -1.560e-02  3.664e-03  -4.256 7.16e-05 ***
## prbpris      3.589e-03  1.136e-02   0.316 0.753243
## taxpc       -4.098e-05  1.192e-04  -0.344 0.732100
## west        -4.157e-04  4.005e-03  -0.104 0.917672
## central     -3.343e-03  2.723e-03  -1.228 0.224165
## pctymle      1.386e-01  6.804e-02   2.037 0.045888 *
## pctmin80     3.701e-04  9.403e-05   3.936 0.000212 ***
## mix         -5.977e-03  1.499e-02  -0.399 0.691523
## wcon         2.378e-05  2.685e-05   0.886 0.379130
## wtuc        -1.549e-06  1.410e-05  -0.110 0.912883
## wtrd         2.204e-05  4.412e-05   0.499 0.619230
## wfir        -3.937e-05  2.550e-05  -1.544 0.127649
## wser        -2.482e-06  5.245e-06  -0.473 0.637702
## wmfg         7.268e-07  1.368e-05   0.053 0.957795
## wfed         5.592e-05  2.829e-05   1.977 0.052547 .
## wsta        -3.551e-05  2.630e-05  -1.350 0.181785
## wloc        -8.592e-07  4.743e-05  -0.018 0.985606
```

```
## polpc:urban -9.294e+00  1.395e+01  -0.666 0.507745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007666 on 62 degrees of freedom
##    (1 observation deleted due to missingness)
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8231
## F-statistic:  18.2 on 23 and 62 DF,  p-value: < 2.2e-16
```

testing for interactions(can't keep in final lab, but good to know...):

```
model_test = lm(crmrte ~ urban + west+central+ density+taxpc, data = raw_data)
summary(model_test)
```

```
##
## Call:
## lm(formula = crmrte ~ urban + west + central + density + taxpc,
##     data = raw_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.017455 -0.007079 -0.001094  0.004180  0.040268
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.813e-02  5.299e-03   5.308 9.48e-07 ***
## urban        3.517e-03  7.208e-03   0.488  0.62693
## west        -1.309e-02  3.022e-03  -4.333 4.20e-05 ***
## central     -7.786e-03  2.784e-03  -2.797  0.00645 **
## density      9.040e-03  1.437e-03   6.289 1.53e-08 ***
## taxpc       -6.723e-05  1.355e-04  -0.496  0.62124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01053 on 81 degrees of freedom
## Multiple R-squared:  0.6838, Adjusted R-squared:  0.6643
## F-statistic: 35.04 on 5 and 81 DF,  p-value: < 2.2e-16
```

```
step <- step(model_test, ~.^2)
```

```
## Start:  AIC=-786.53
## crmrte ~ urban + west + central + density + taxpc
##
##                   Df Sum of Sq       RSS     AIC
## + central:density  1 0.0006174 0.0083643 -790.72
## + urban:central    1 0.0004951 0.0084867 -789.46
## - urban            1 0.0000264 0.0090081 -788.27
## - taxpc            1 0.0000273 0.0090090 -788.26
## + urban:density    1 0.0002601 0.0087216 -787.08
## <none>                         0.0089817 -786.53
## + west:density     1 0.0001277 0.0088540 -785.77
## + west:taxpc       1 0.0001018 0.0088799 -785.52
## + central:taxpc    1 0.0000827 0.0088991 -785.33
## + urban:taxpc      1 0.0000542 0.0089275 -785.05
## + urban:west       1 0.0000375 0.0089443 -784.89
## + density:taxpc    1 0.0000052 0.0089765 -784.58
```

```
## - central          1 0.0008673 0.0098491 -780.51
## - west             1 0.0020815 0.0110632 -770.39
## - density          1 0.0043858 0.0133675 -753.93
##
## Step:  AIC=-790.72
## crmrte ~ urban + west + central + density + taxpc + central:density
##
##                   Df  Sum of Sq       RSS     AIC
## + central:taxpc    1 0.00053877 0.0078255 -794.52
## - urban            1 0.00000083 0.0083651 -792.71
## - taxpc            1 0.00003680 0.0084011 -792.34
## <none>                           0.0083643 -790.72
## + urban:taxpc      1 0.00007842 0.0082859 -789.54
## + west:taxpc       1 0.00007828 0.0082860 -789.54
## + urban:density    1 0.00006619 0.0082981 -789.41
## + west:density     1 0.00001973 0.0083446 -788.93
## + urban:central    1 0.00000985 0.0083545 -788.83
## + urban:west       1 0.00000665 0.0083577 -788.79
## + density:taxpc    1 0.00000000 0.0083643 -788.72
## - central:density  1 0.00061744 0.0089817 -786.53
## - west             1 0.00192205 0.0102864 -774.73
##
## Step:  AIC=-794.52
## crmrte ~ urban + west + central + density + taxpc + central:density +
##     central:taxpc
##
##                   Df  Sum of Sq       RSS     AIC
## + west:taxpc       1 0.00038847 0.0074371 -796.95
## - urban            1 0.00000012 0.0078257 -796.51
## <none>                           0.0078255 -794.52
## + urban:density    1 0.00017408 0.0076515 -794.47
## + urban:taxpc      1 0.00009511 0.0077304 -793.58
## + urban:west       1 0.00005178 0.0077738 -793.09
## + density:taxpc    1 0.00005163 0.0077739 -793.09
## + urban:central    1 0.00004876 0.0077768 -793.06
## + west:density     1 0.00002474 0.0078008 -792.79
## - central:taxpc    1 0.00053877 0.0083643 -790.72
## - central:density  1 0.00107352 0.0088991 -785.33
## - west             1 0.00217121 0.0099967 -775.21
##
## Step:  AIC=-796.95
## crmrte ~ urban + west + central + density + taxpc + central:density +
##     central:taxpc + west:taxpc
##
##                 Df  Sum of Sq       RSS     AIC
## - urban          1 0.00002721 0.0074643 -798.63
## <none>                         0.0074371 -796.95
## + urban:taxpc    1 0.00009141 0.0073457 -796.02
## + urban:density  1 0.00004913 0.0073879 -795.52
## + density:taxpc  1 0.00002703 0.0074100 -795.26
## + urban:west     1 0.00001450 0.0074226 -795.12
## + west:density   1 0.00000463 0.0074324 -795.00
## + urban:central  1 0.00000011 0.0074370 -794.95
## - west:taxpc     1 0.00038847 0.0078255 -794.52
```

```
## - central:taxpc    1 0.00084896 0.0082860 -789.54
## - central:density  1 0.00124953 0.0086866 -785.43
##
## Step:  AIC=-798.63
## crmrte ~ west + central + density + taxpc + central:density +
##     central:taxpc + west:taxpc
##
##                   Df  Sum of Sq       RSS      AIC
## <none>                         0.0074643 -798.63
## + urban           1 0.00002721 0.0074371 -796.95
## + density:taxpc   1 0.00002600 0.0074383 -796.93
## + west:density    1 0.00000537 0.0074589 -796.69
## - west:taxpc      1 0.00036138 0.0078257 -796.51
## - central:taxpc   1 0.00082356 0.0082878 -791.52
## - central:density 1 0.00122506 0.0086893 -787.41
```

```r
step_output <- step$anova
write.csv(step_output, "Step2variables2.csv")
```

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

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