A Regression Analysis
FINAL
W203 Lab 3: Reducing Crime in North Carolina

Anusha Praturu, Stephanie Mather, Thanh Le, Laura Chutny

April 15, 2019

Abstract

To support the Governor's bid for re-election, we investigate possible drivers for the rate of crime in North Carolina using a data set from 1987. Our intent is to answer the research question: What are the top 3 determinants of crime rate that our political campaign will propose policy to address to improve the lives of North Carolinians? Through our analysis we identified three hallmark policy recommendations to decrease crime rate and we provide statistical evidence to support our recommendations. Our policy recommendations are to address population density, increase the rate of conviction in the state, and close the wage gap between the high and low earners. We have also identified several directions in which we could further our analysis. In following our recommendations for policy action and further study, we are confident the Governor's campaign policy will help to effectively reduce crime in North Carolina.

Contents

List of Figures

List of Tables

1. Introduction

This statistical investigation aims to provide insights that will help shape policy for our Governor's political campaign. We examine the 1987 snapshot of crime statistics from panel data for select counties in North Carolina. This data was first used in a study by Cornwell and Trumball, (C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data", *Review of Economics and Statistics* **76**, pp.360-366).

Without referencing the analysis conducted in the paper, but using the 1987 data, we propose this research question:

What are the top 3 determinants of crime rate that our political campaign will propose policy to address to improve the lives of North Carolinians?

This quantitative analysis examines this subset of data to understand the underlying drivers of crime rate and how these factors can be used to inform the policy platform for our political campaign: what are the determinants of crimes in North Carolina? We conduct the following analyses:

- Explore potential relationships between crime rate and population density, wages, tax revenue, police per capita, conviction and arrest rate using both bivariate and a multivariate methods.
- Inspect region-specific aspects of the variables to assess their potential effects on crime rates.
- Check variable outliers that could skew our regression models and determine whether to remove or keep based on data quality and consistency.
- Discuss omitted variables and what effect this may have on our conclusions.
- Finally, we utilise the results of the statistical models to provide policy guidance for the upcoming election campaign.

The details of the analysis and our findings are discussed in the report below.

2. Initial Data Cleaning

2.1 Data Import

Our data source is the CSV file `crime_v2.csv` from C. Cornwell and W. Trumball (ibid.). There are 97 rows of data and 25 variables in the original file. The variable descriptions are as shown in Table 1.

Variable	**Description**
county	county identifier
year	1987
Dependent Variable	
crmrte	crimes committed per person
Independent Variable	*Judiciary and Crime*
polpc	police per capita
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	average sentence in days
mix	offense mix: face-to-face/other
Independent Variable	*Demographic*
density	people per square mile
pctmin80	perc. minority, 1980
pctymle	percent young male
Independent Variable	*Demographic(Region)*
west	=1 if in western N.C.
central	=1 if in central N.C.
urban	=1 if in SMSA
Independent Variable	*Tax and Wages*
taxpc	tax revenue per capita
wcon	weekly wage, construction
wtuc	wkly wge, trns, util, commun
wtrd	wkly wge, whlesle, retail trade
wfir	wkly wge, fin, ins, real est
wser	wkly wge, service industry
wmfg	wkly wge, manufacturing
wfed	wkly wge, fed employees
wsta	wkly wge, state employees
wloc	wkly wge, local gov emps

Table 1: Table of Variables

```
> #library installation
> library(car)
> library(corrplot)
> library(ggplot2)
> library(lmtest)
> library(psych)
> library(sandwich)
> library(stargazer)
> library(xtable)
```

```
> # Read in data and have a brief peek at it
> crime_df = read.csv("crime_v2.csv")
> str(crime_df)
```

```
'data.frame':   97 obs. of  25 variables:
 $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
 $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
 $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
 $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
 $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
 $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
 $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
 $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
 $ density : num  2.423 1.046 0.413 0.492 0.547 ...
 $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
 $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
 $ central : int  1 1 0 1 0 0 0 0 0 0 ...
 $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
 $ wcon    : num  281 255 227 375 292 ...
 $ wtuc    : num  409 376 372 398 377 ...
 $ wtrd    : num  221 196 229 191 207 ...
 $ wfir    : num  453 259 306 281 289 ...
 $ wser    : num  274 192 210 257 215 ...
 $ wmfg    : num  335 300 238 282 291 ...
 $ wfed    : num  478 410 359 412 377 ...
 $ wsta    : num  292 363 332 328 367 ...
 $ wloc    : num  312 301 281 299 343 ...
 $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
 $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

2.2 Data Cleansing

Our initial data exploration found several values that were measurement or recording errors. These data points were justifiably removed to ensure the analysis of the data was consistent and more representative of reality.

2.2.1 *Removing Empty Rows*

The last six rows of the data set contained all NAs. Without the original data source, one cannot determine whether these were residual errors from the original import file or due to other unknown factors. The removal of these rows will not affect our analysis.

```
> tail(crime_df,7)
```

```
   county year    crmrte   prbarr     prbconv  prbpris avgsen      polpc
91    197   87 0.0141928 0.207595 1.182929993 0.360825  12.23 0.00118573
92     NA   NA        NA       NA                   NA     NA         NA
93     NA   NA        NA       NA                   NA     NA         NA
94     NA   NA        NA       NA                   NA     NA         NA
95     NA   NA        NA       NA                   NA     NA         NA
96     NA   NA        NA       NA                   NA     NA         NA
97     NA   NA        NA       NA           `       NA     NA         NA
    density    taxpc west central urban pctmin80    wcon     wtuc    wtrd
91 0.889881 25.95258    1       0     0  5.46081 314.166 341.8803 182.802
92       NA       NA   NA      NA    NA       NA      NA       NA      NA
93       NA       NA   NA      NA    NA       NA      NA       NA      NA
94       NA       NA   NA      NA    NA       NA      NA       NA      NA
95       NA       NA   NA      NA    NA       NA      NA       NA      NA
96       NA       NA   NA      NA    NA       NA      NA       NA      NA
97       NA       NA   NA      NA    NA       NA      NA       NA      NA
       wfir     wser   wmfg   wfed   wsta   wloc        mix    pctymle
91 348.1432 212.8205 322.92 391.72 385.65 306.85 0.06756757 0.07419893
92       NA       NA     NA     NA     NA     NA         NA         NA
93       NA       NA     NA     NA     NA     NA         NA         NA
94       NA       NA     NA     NA     NA     NA         NA         NA
95       NA       NA     NA     NA     NA     NA         NA         NA
96       NA       NA     NA     NA     NA     NA         NA         NA
97       NA       NA     NA     NA     NA     NA         NA         NA
```

```
> # Note the NA rows at the bottom. Removing these will not change the data.
> crime_df <- crime_df[complete.cases(crime_df), ]
> # checking remaining data for NAs
> sapply(crime_df,function(x) sum(is.na(x)))
```

```
  county     year   crmrte   prbarr  prbconv  prbpris   avgsen    polpc
       0        0        0        0        0        0        0        0
 density    taxpc     west  central    urban pctmin80     wcon     wtuc
       0        0        0        0        0        0        0        0
    wtrd     wfir     wser     wmfg     wfed     wsta     wloc      mix
       0        0        0        0        0        0        0        0
 pctymle
       0
```

2.2.2 *Incorrect Data Type on Import*

The variable `prbconv` representing Probability of Conviction after arrest had imported incorrectly

as a Factor class due to the inclusion of a single quote (‘) in one of the blank original NA rows which has now been removed. The conversion of `prbconv` back to numerical class results in no loss of data.

```
> # convert to numeric
> crime_df['prbconv'] <- as.numeric(as.character(crime_df$prbconv))
> summary(crime_df['prbconv'])
```

```
    prbconv
 Min.   :0.06838
 1st Qu.:0.34541
 Median :0.45283
 Mean   :0.55128
 3rd Qu.:0.58886
 Max.   :2.12121
```

Convert categorical variables `west`, `central`, and `urban` to factor data type:

```
> crime_df$west <- factor(crime_df$west)
> crime_df$central <- factor(crime_df$central)
> crime_df$urban <- factor(crime_df$urban)
```

2.2.3 *Erroneous Data / Outliers*

One of the wages in the Service Wage data is 10 times larger than any other data. There is no other data for this county to suggest a high income (i.e tax revenue shows no anomaly), so we feel justified in removing this data point as potentially erroneous. All other data is kept for the county.

```
> crime_df$wser[84]=NA
```

Population density variable `density` has an erroneous value of 0.00002 which is equivalent to 1 person per 50,000 square mile. North Carolina is only 53,800 square miles, confirming this must be a data error. The data point has been removed by bottom coding the population densities to 0.001 people per square mile. Values more extreme than the bottom limit are removed from the data set.

```
> summary(crime_df['density'])
```

```
    density
 Min.   :0.00002
 1st Qu.:0.54741
 Median :0.96226
 Mean   :1.42884
 3rd Qu.:1.56824
 Max.   :8.82765
```

```
> #subset dataframe to population densities above 0.001 people per square mile
> crime_df <- subset(crime_df, crime_df$density > 0.001)
> summary(crime_df['density'])
```

```
    density
 Min.   :0.3006
```

```
 1st Qu.:0.5519
 Median :0.9792
 Mean   :1.4447
 3rd Qu.:1.5693
 Max.   :8.8277
```

Probability values greater than 1 do not make mathematical sense, however they are present in the variables Probability of Arrest, `prbarr` and Probability of Conviction, `prbconv`. This is likely of a result of them being ratios, rather than true probabilities. Perhaps there were multiple convictions and arrests for a single offense, e.g. a robbery conducted by 3 people or arrest and convictions fell across a sample year. The Probability of Prison Sentence, `prbpris` is between 0 and 1. These values are left in the data set, because although they may lead to an overestimation of the probability of arrest and conviction, the error is consistent across all data points and removing a few of the lines would introduce skew away from arrest or conviction.

2.2.4 *Duplicate Entry*

There is a duplicate entry for county 193. This should be removed so that it does not have additional weight in the results.

```
> subset(aggregate(year ~ county, crime_df, function(x) length(x)), year >1)
```

```
   county year
87    193    2
```

Remove duplicate entry:

```
> crime_df = crime_df[!duplicated(crime_df),]
```

2.2.5 *Manual Entry*

County 115 has three extreme outliers for the `prbconv`, `prbpris` and `mix` variables. A manual entry error is possible as all other entries have 6 significant figures: for each variable this county only has 1-2. Due to the numerous extreme outliers and the possible data quality issue, this county is removed from the study.

This leaves us with 88 out of 100 counties included in our models.

```
> which(crime_df$county == 115)
```

```
[1] 51
```

```
> crime_df = crime_df[-c(51),]
> dim(crime_df)
```

```
[1] 88 25
```

```
> summary(crime_df$prbpris)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2273  0.3654  0.4222  0.4126  0.4560  0.6000
```

3. Exploratory Data Analysis (EDA) of Variables

The data exploration focused on possible policy areas for the political campaign. Crime rate has been chosen as our dependent variable, representing the number of crimes committed per person in each county, as it is likely to be an area of focus of the Governor's constituents. The explanatory variables were selected using a combination of factors, including EDA showing high correlation with crime rate, and the fact that the selected explanatory variables can be concretely addressed with policy recommendations.

3.1 Dependent Variable, Crime Rate: `crmrte`

We start our exploratory data analysis (EDA) of our Dependent Variable, Crime Rate, with summary the statistics:

```
> (summary(crime_df$crmrte))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01062 0.02201 0.03002 0.03405 0.04088 0.09897
```

```
> cat("The summary statistics variables show a range of values from",
+       round(min(crime_df$crmrte),5)*100,"to", round(max(crime_df$crmrte),5)*100,
+   "crimes per hundred people,","with a mean crime rate of",
+       round(mean(crime_df$crmrte),5)*100,"crimes per hundred people.")
```

The summary statistics variables show a range of values from 1.062 to 9.897 crimes per hundred people, with a mean crime rate of 3.405 crimes per hundred people.

Next we check the normality assumption for the crime rate variable in the data set by viewing the histogram (Figure 1a). The crime rate is roughly normal with a positive skew. To correct the positive skew we applied a natural logarithmic transformation to crime rate as seen in Figure 1b. This means our model will provide prediction in terms of percentage increase in crime, which is more practical when addressing policy concerns. The log of crime rate follows the normal distribution well in Figure 1c, thus it will be used as the dependent variable for our regression model.

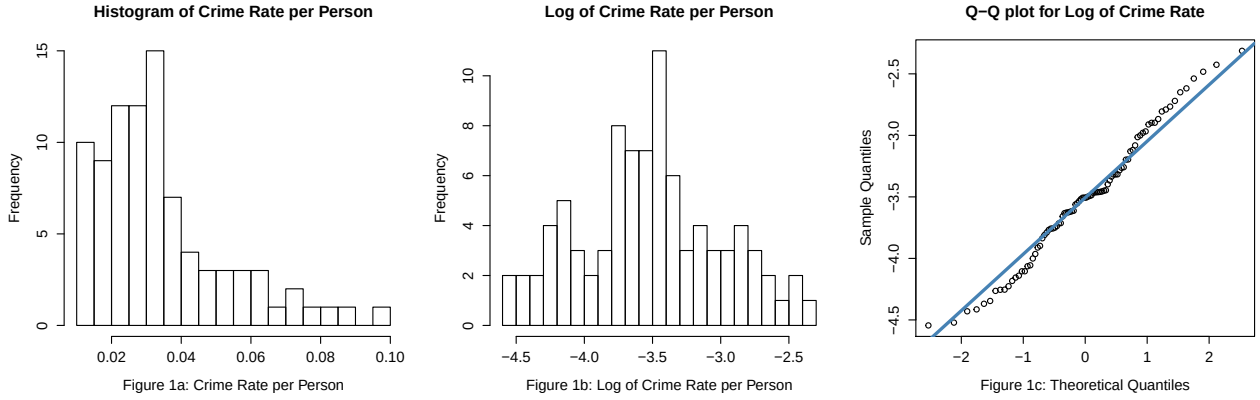

Figure 1a: Crime Rate per Person

Figure 1b: Log of Crime Rate per Person

Figure 1c: Theoretical Quantiles

Figure 1: Transform of Crime Rate as Dependent Variable

3.2 Independent Variable Identification

A heat map plot identifies variables which are most strongly correlated with crime rate and each other (negatively or positively) is in Figure 2. There were no pairs with perfect collinearity, but several variables have correlation. Note, Figure 2 does not check for cross correlation for the Region variables as they are categorical data and their relationship to other variables is explored in Section 3.4.

```
> #Add the log(crimerate) as a separate data column to the dataset
> #so we can check variables correlation to the log of crmrte.
> crime_df$logcrmrte <- log(crime_df$crmrte)
```

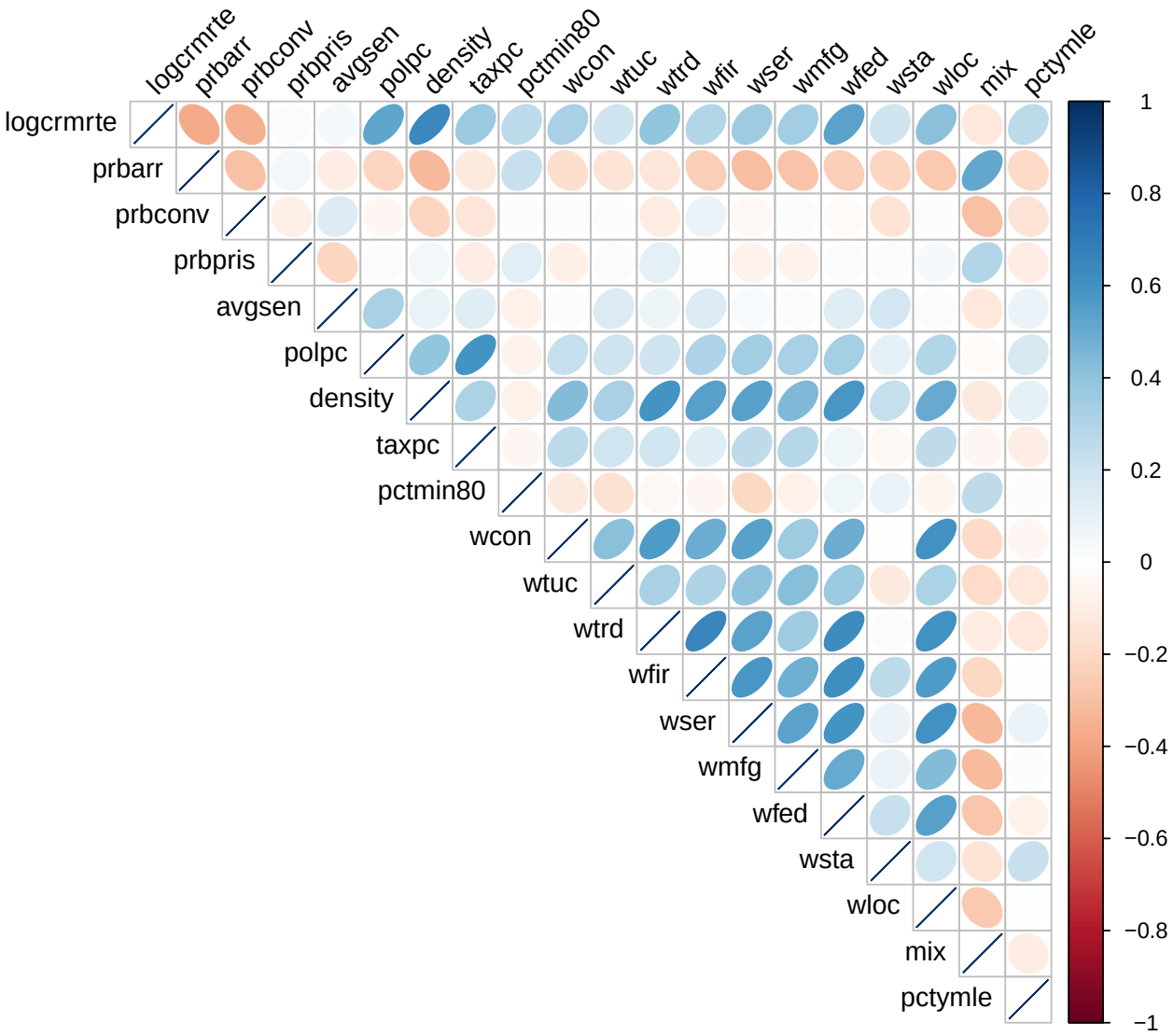

Figure 2: Variable Cross Correlation Plot

Obvious in Figure 2 is the positive correlation between crime rate and population density suggesting this may be major component of our crime rate prediction.

3.3 Independent Variables: Judicary and Crime

This section will look at independent variables that fall in the categories describing crimes, policing and the effect of the judiciary system:

1. Probability of Arrest (`prbarr`): the ratio of arrests to offenses
2. Probability of Conviction (`prbconv`): the ratio of convictions to arrests
3. Probability of being sentenced to Prison (`prbpris`): the proportion of convictions resulting in prison sentences to total convictions
4. Average Sentence in Days (`avgsen`): average prison sentence in days is a proxy for sanction severity
5. Offense Mix (`mix`): ratio of face-to-face crimes to other crime types
6. Police per Capita (`polpc`): the number of police per capita

We theorize that crime rates will decrease with an increased probability of arrest, conviction and greater likelihood and severity of punishment (sentencing). Figure 3 is a matrix of the Judiciary and Crime variable histograms (diagonal) , bivariate scatterplots (lower half) and correlations (upper half). It also includes the dependent variable log(Crime Rate) for comparison.

Figure 3 suggests the theory is only true for increased arrest and convictions. Prison sentences may increase crime rates as it has a positive correlation with crime. The histograms also show a negative skew expected with more numerous low level crimes and the effect of the zero-cutoff for real numbers. This is corrected by using the natural log of `prbarr`/`prbcon`/`prbpris`. This is practical for a policy standpoint as a percentage increase in arrests etc. can lead to a similar percentage decrease in crimes. Figure 4 shows the transformed variables for arrest, conviction and prison sentence.

```
> jud_crm <- crime_df[ ,c('crmrte','prbarr','prbconv','prbpris','avgsen',
+                         'mix', 'polpc')]
> jud_crm$logcrmrte <- log(jud_crm$crmrte)
> pairs.panels(jud_crm[,-1], scale=FALSE, density=TRUE, digits=2, method="pearson",
+           hist.col='blue',rug=FALSE,breaks=20)
```

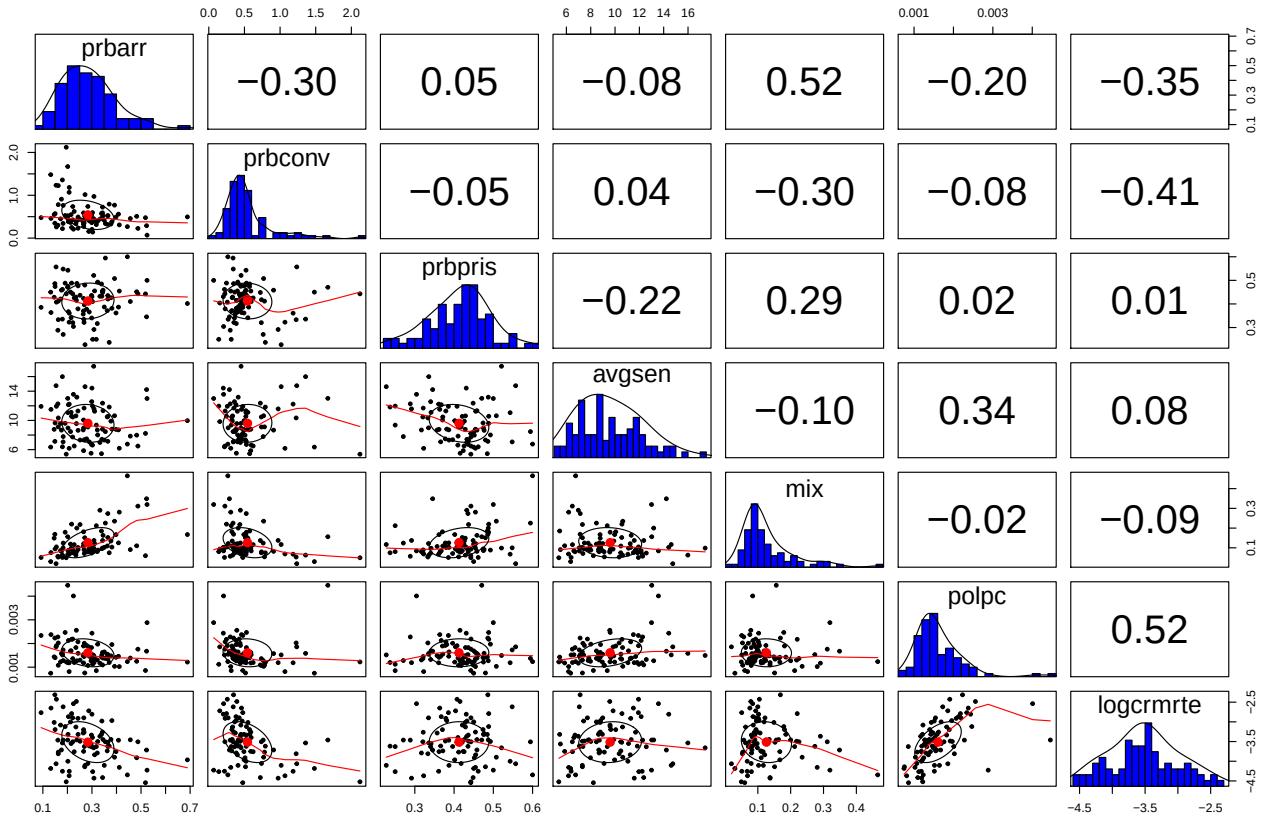

Figure 3: Judiciary and Crime variables vs log(Crime Rate)

There is possible collinearity between the `prbarr`/`prbcon`/`prbpris` variables due to their ratio relationship. However, the low VIF (explained below) confirms the visual assessment of low collinearity and we can keep them in our analysis.

Average sentence shows a very weak correlation with crime rate as shown in Figure 3 and has a non-normal distribution so it is not included in our model.

There is a possibility that the type of crimes being committed in a county affects the arrest rate, i.e. face-to-face crime leads to a greater identification rate and thus arrest The EDA shows crime *mix* is not correlated with arrest rate ($R^2 < 50\%$). The correlation plot against crime rate also showed a low correlation. Due to these two factors and our inability to influence crime mix, `mix` will not be further included in our base model.

A variance inflation factor (VIF) detects multicollinearity in regression analysis; it estimates how much the variance (standard error squared) of a coefficient is inflated due to multicollinearity in the model. The VIF is always a value equal to or greater than 1. Typically, a VIF ≥ 10 is statistically significant; however, in weaker regression models, a VIF ≥ 2.5 may be cause for concern. A VIF of 1.5 denotes the variance of that coefficient is 50% bigger than what one would expect if there was no multicollinearity.

```
> cat("Judicial Variables VIF:")
```

```
Judicial Variables VIF:
```

```
> vif(lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris), data=crime_df))
```

```
log(prbarr) log(prbconv) log(prbpris)
   1.103734     1.106059     1.002735
```

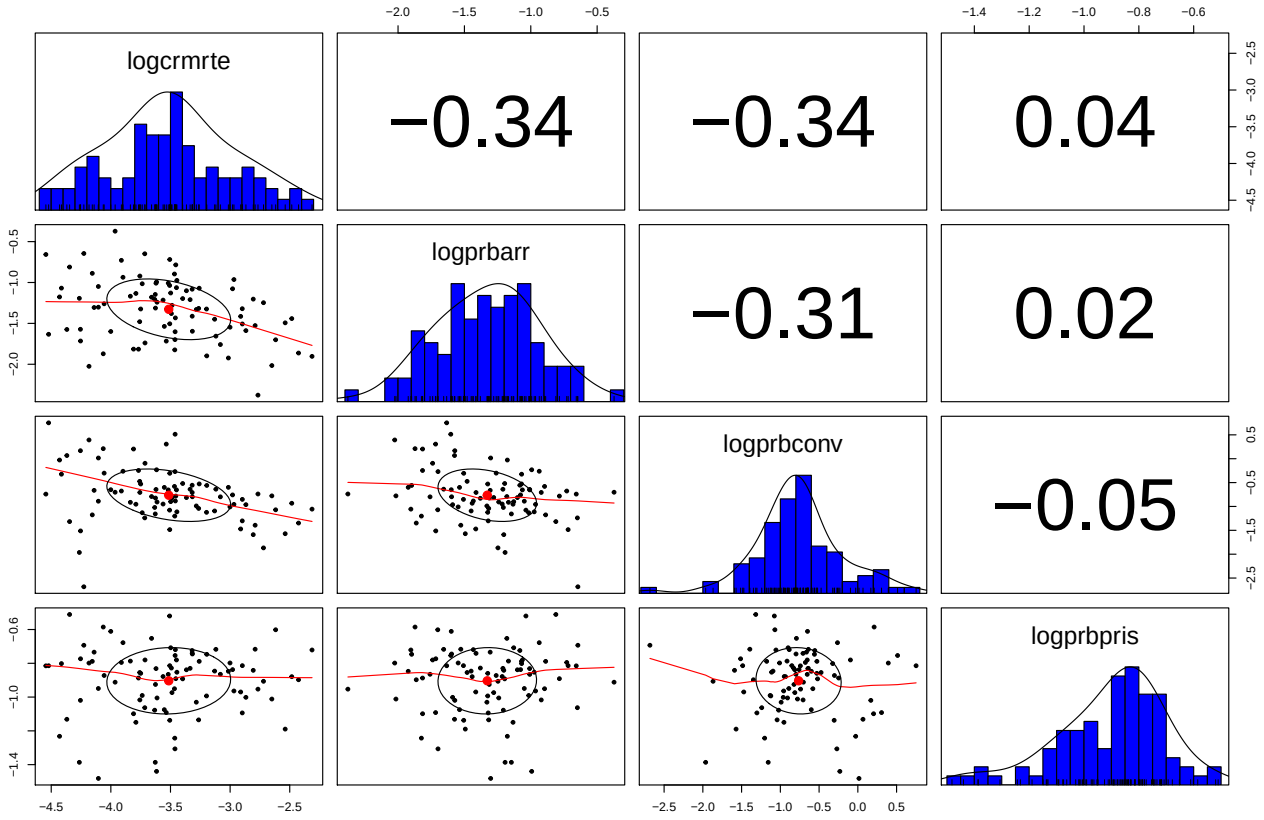

Figure 4: Transformed Judiciary and Crime variables vs Crime Rate

3.3.1 Crimes Committed vs. Police Per Capita

Police presence (police per capita), represented by `polpc`, is one choice for a main variable that seems obvious at first and shows a strong initial correlation with crime rate. It is hypothesized that crime reduces with increased police presence. Figure 5 shows the results of the bivariate linear model and standardized residuals.

```
> model1 <- lm(log(crime_df$crmrte) ~ log(crime_df$polpc))
> cat("R Squared: ", round(summary(model1)$r.square,5),"\n")
```

```
R Squared:  0.36387
```

```
> coeftest(model1, vcov=vcovHC)
```

```
t test of coefficients:

                Estimate Std. Error t value  Pr(>|t|)
```

```
(Intercept)           2.86508    1.35042  2.1216   0.03675 *
log(crime_df$polpc)   0.98355    0.20520  4.7932 6.807e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> cat("Correlation: ",round(cor(log(crime_df$polpc), log(crime_df$crmrte)),5))

Correlation:  0.60322
```

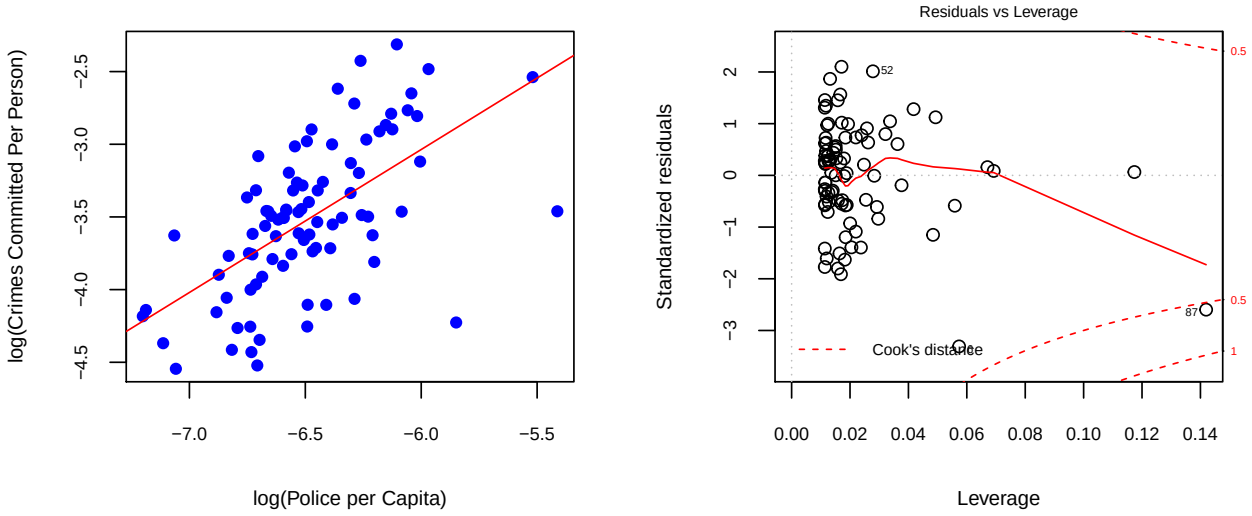

Figure 5: Crime Rate vs Police per Capita

As seen in Figure 5, there appears to be a reasonable correlation between police per capita and the crime rate in counties across North Carolina. However the correlation that exists is in the opposite direction to what was expected, i.e. the data suggest that more police are correlated with a higher crime rate. Thus, we will likely not be recommending an investment in increasing police forces as part of our policy platform without further analysis.

From the residuals graph, we also see there is one data point (outlier) that has some influence on the model. However, given that its Cook's distance < 1, there is no cause for alarm. It also does not show signs of being an erroneous data point, so it is left in the model.

We proposed to consider the following independent judicial and crime variables in our models: `prbarr` and `prbconv`.

3.4 Independent Variables: Demographic

1. Population Density (`density`)
2. Percentage of Young males (`pctymle`)
3. Percentage of minorities (`pctmin80`) - note this data is from 1980, seven years prior to the rest of the data in the model
4. County in West N.C. (`west`) - if west = 1
5. County in Central N.C. (`central`) - if central = 1
6. County is a Standard Metropolitan Statistical Area (SMSA)* in N.C. (`urban`) - if urban = 1

Note the definition for SMSA has changed since 1987. SMSA's are now known as as [Metropolitan Statistical Areas] (https://www.census.gov/programs-surveys/metro-micro/about.html).

The demographics of a county and state are important influences on crime rate, however they are not easily influenced through policy, especially in the short term. Demographics are often related to affluence/poverty as well as industrial base and may be highly variable within the county's suburban, rural and urban areas. Changes to policy with an industrial or poverty focus may require a longer period of time to see effects in regards to demographics, possibly in the range of decades.

3.4.1 Density and Crime Rate by Region

The information in the sample set location data allows us to to ask if different policies could be promoted within different county clusters. Location could have a possible clustering affect on the data which is an important violation of the Random Sampling OLS and MLR Assumptions (as discussed in Section 6). Note: The definition for code `urban` in the data is that the county is in a Standard Metropolitan Statistical Area, as an indicator variable to signify county as 1=urban or 0=rural. Counties are also identified as being in the West or Central areas, The rest are unclassified.

Figure 6 shows how counties are distributed by region and the count of SMSA in each. Interestingly, in this data set, each county is weighted equally, which means rural data is essentially weighted more on a population basis.

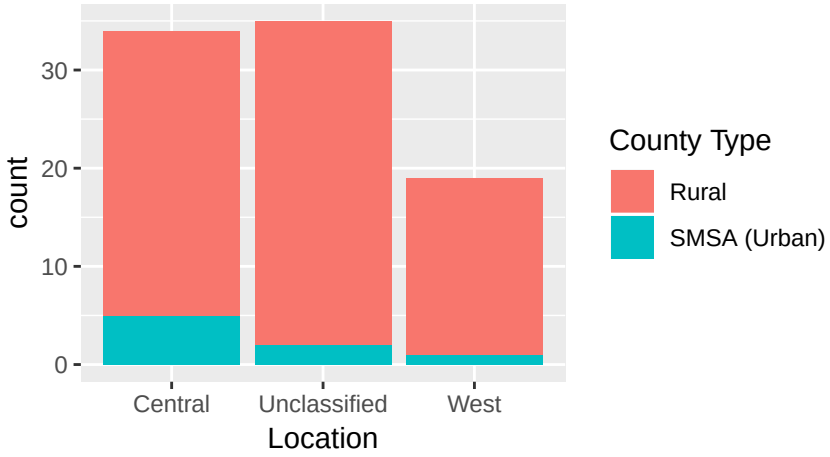

Figure 6: Distribution of Counties by Region

Next, we will see if there is a clustering effect on crime rate based on location. Upon visual inspection of the bubble chart in Figure 7 below, it does not appear that there is any undue clustering of crime by location when plotted against police per capita, although there is more crime in higher density SMSA locations. The effect of clustering is further reduced by the large amount of unclassified counties.

To maintain applicability of the central limit theorem we cannot group by location as the sample sizes would become too small. In addition, the limited available description of the meaning of the location parameters directs us away from using it a key parameter in our policy suggestions. Finally, a county being classified as 'SMSA or urban' is omitted from the analysis as we are using density as

a more granular proxy for being in a city. Thus, we will not use the location variables, but rely on population density instead.

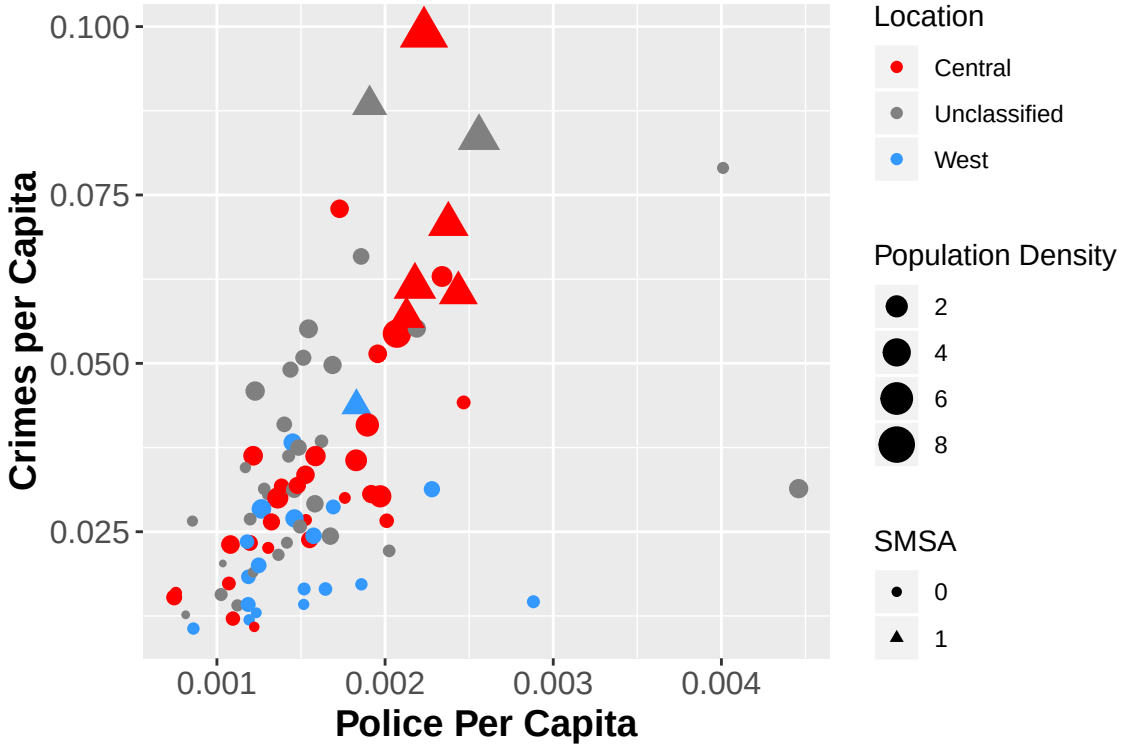

Figure 7: Crime Rate vs Police Per Capita by Location

3.4.2 Non-location based Demographic Variables Investigation

The normality check for `density`, `pctymle`, and `pctmin80` suggest a log is a suitable transformation for these variables. This is also practical from a policy standpoint as a percentage reduction in crime could be attributed to a percentage change in population density, young males or minority population proportions.

As evidenced in the plot in Figure 8, percentage young male and minorities have a weaker correlation with crime rate than population density does. For this reason they will not be included in the base model. Density has a strong correlation.

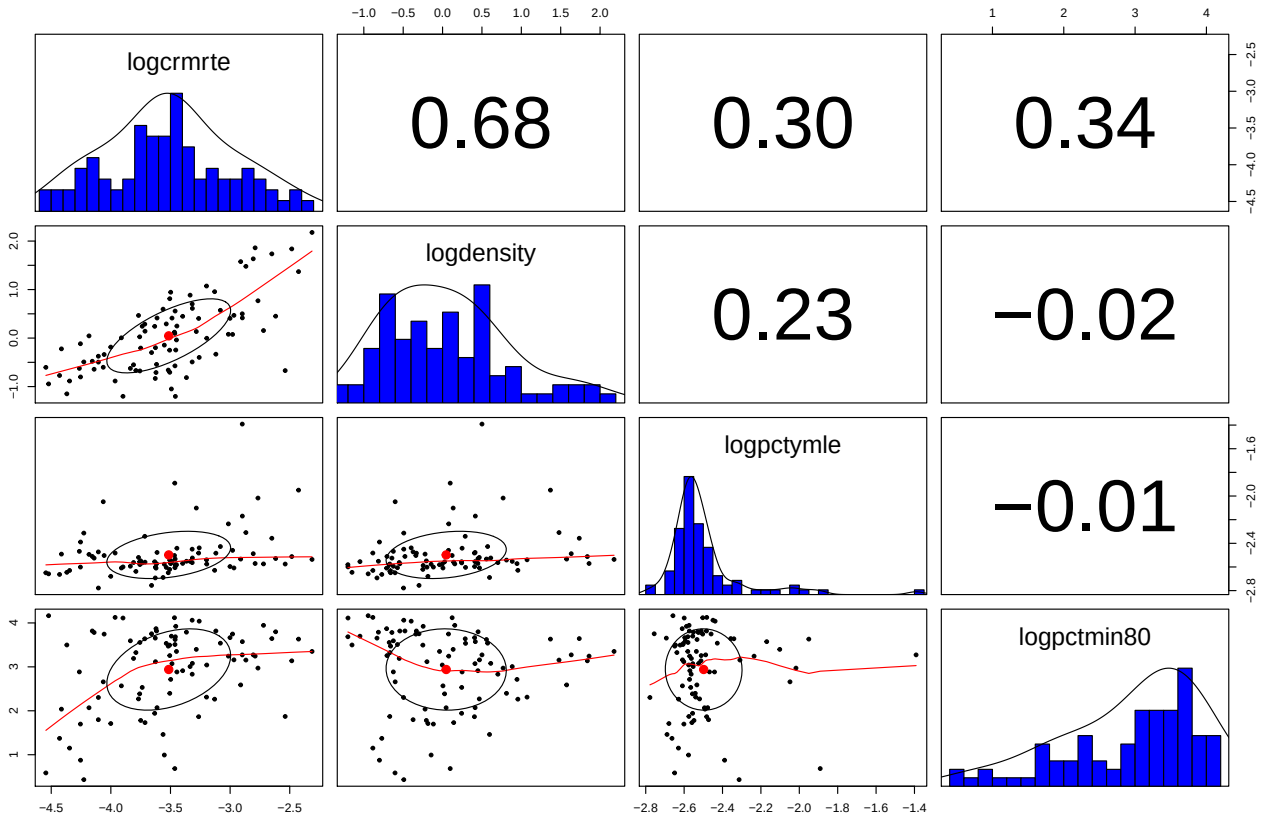

Figure 8: Demographic variables vs Crime Rate

3.4.3 Crime Rate vs. Population Density

In the model below, we look at Population Density and its relationship to crime rate following our discovery of their high correlation.

```
> model3 = lm(log(crime_df$crmrte) ~ log(crime_df$density))
> model3
```

```
Call:
lm(formula = log(crime_df$crmrte) ~ log(crime_df$density))

Coefficients:
          (Intercept)  log(crime_df$density)
              -3.5359                 0.4643
```

```
> cat("R Squared: ", round(summary(model3)$r.squared,5))
```

```
R Squared:  0.4663
```

```
> cat("Pvalue for Slope ", summary(model3)$coefficients[2,4])
```

```
Pvalue for Slope  2.329455e-13
```

```
> cat("Correlation: ", round(cor(log(crime_df$density), log(crime_df$crmrte)),5))

Correlation:  0.68286
```

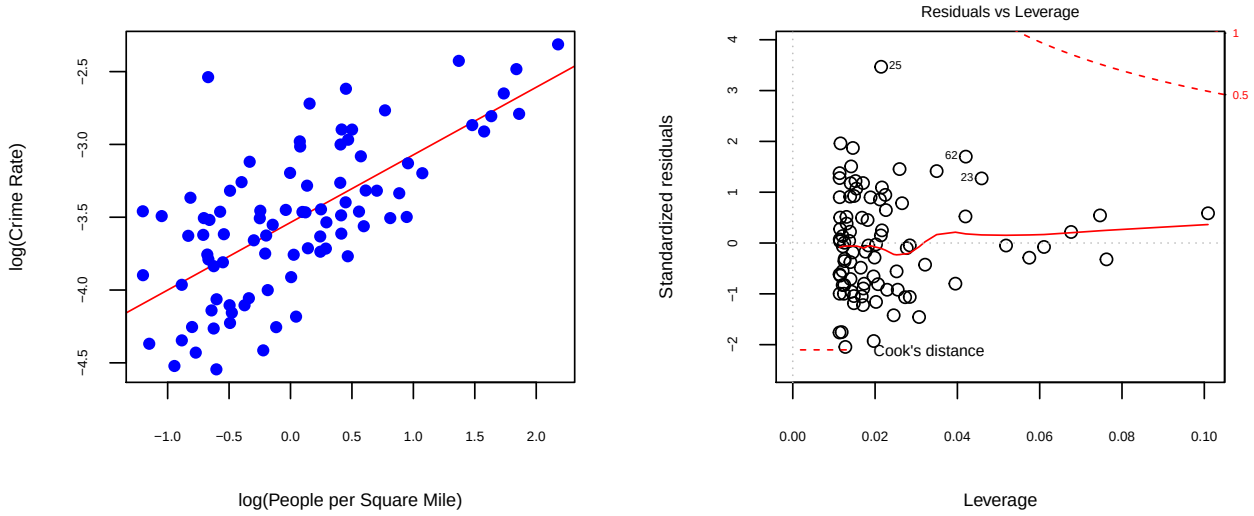

Figure 9: Crime Rate vs Population Density

Figure 9 shows the relationship between population density and crime rate showing the strongest correlation of any variable studied thus far, at 0.683. Further, a linear model yields an extremely small p-value, well below 0.05, allowing us to reject the null hypothesis ($H_0 : \beta_1 = 0$) and say that there exists a relationship between population density and crime rate. We cannot yet determine if this relationship is causal, but it could lead to policy recommendations having to do with addressing population density with the intent of reducing crime. The Residuals vs Leverage graph also confirms that the data in this analysis falls within the Cook's distance, indicating no high influence or leverage of any particular data point.

3.5 Tax and Wage Variables

Tax and wages are areas of economic policy that could be leveraged in the campaign to decrease crime rates. Tax is a key driver of available funds for policing or other crime reduction measures whereas wages are indicative of affluence in a county. Policies could be enacted in respect to minimum wage (proxied by the service wage in our data set) or government employees' wages. We hypothesize that as wages and tax revenue increase, crime rate will decrease.

3.5.1 Tax

The histogram of tax shows a positive skew and is partially corrected by using a log transform as shown in Figure 10. A positive skew (even after transform) is common for tax and wage data due to the presence of large positive outliers. The log transformation is also practical from a policy standpoint as we can then discuss percentage changes in tax having an effect on crime rate. The deviation from normality is accepted to keep practicality of a percentage change in tax rather than employing more complex modelling techniques.

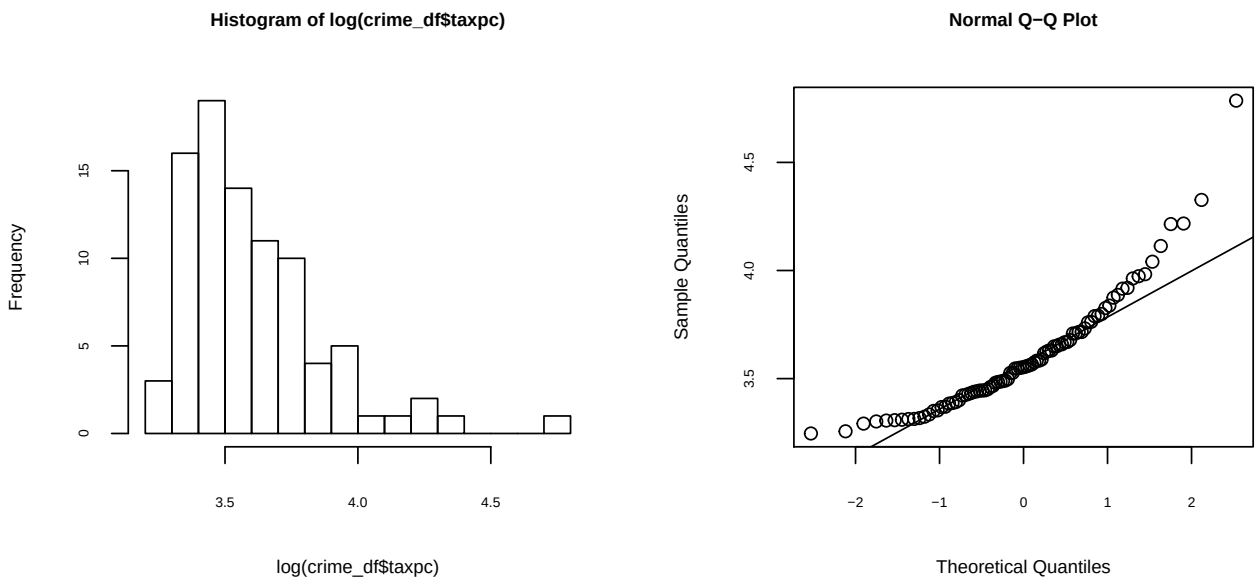

Figure 10: Transform for Tax Revenue per capita

Upon review of Figure 10, following the log transformation of variable `taxpc`, we can see the one extreme outlier for the log(`taxpc`) distribution. As can be seen in Figure 11 below, although the outlier has large leverage, it has a small residual and thus a Cook's distance < 0.5 which stops it from unduly influencing a simple linear model. The scale-location plot in Figure 11 does show some deviation from homoskedasticity. The HC covariance model will be utilised in our models to correct for this heteroskedasticity.

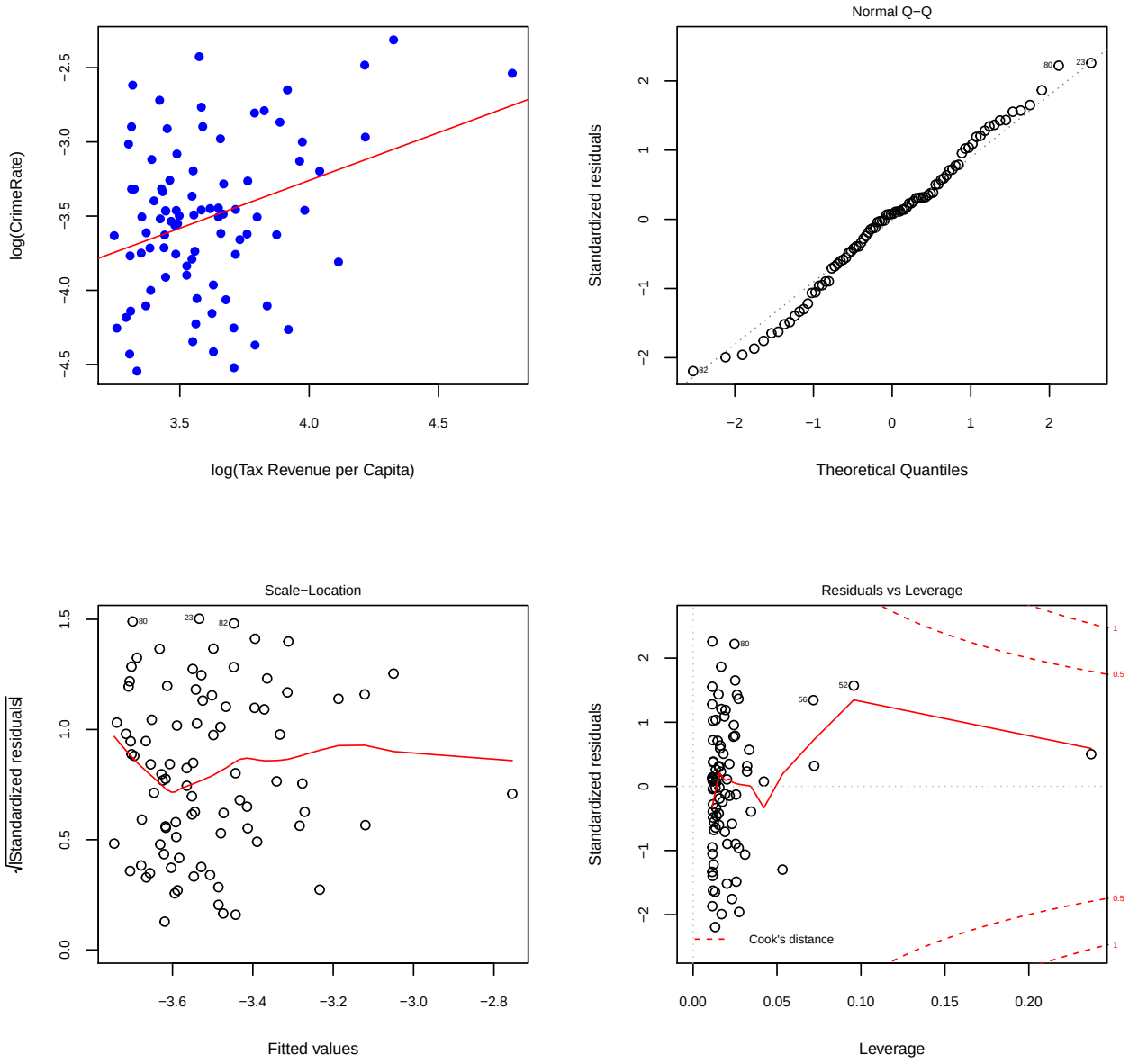

Figure 11: log(Crime Rate) vs log(Tax Revenue per capita) linear model

An analysis of the coefficients in the relationship shown in Figure 11 between tax revenue and crime rate across North Carolina is shown below.

```
> model2 = lm(log(crime_df$crmrte) ~ log(crime_df$taxpc))
> coeftest(model2, vcov = vcovHC)
```

```
t test of coefficients:

                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)         -5.83489    0.73049 -7.9877 5.602e-12 ***
log(crime_df$taxpc)  0.64376    0.20315  3.1690  0.002119 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> cat("R Squared: ", round(summary(model2)$r.squared,5))

R Squared:  0.10915
> cat("Pvalue for Slope ", round(coeftest(model2, vcov = vcovHC)[2,4],6))

Pvalue for Slope  0.002119
> cat("Correlation: ",round(cor(log(crime_df$taxpc), log(crime_df$crmrte)),5))

Correlation:  0.33038
```

From the analysis of the linear model we see a reasonable correlation between the log of crime rate and the log of tax revenue per capita, with our model yielding a p-value for the slope < 0.05. Thus we can reject the null hypothesis (that the slope is 0) and use the supporting evidence that there exists a relationship between tax revenue and crime rate. While we cannot yet determine if this relationship is causal, we will include this variable in our multiple regression model and these findings may lead to policy recommendations related to tax rates. Note that all data in this analysis fall with in the Cook's distance indicating low leverage and no significant outliers.

3.5.2 Wage Variables

Observing the cross correlations shown in Figure 2, we see that wages unexpectedly show a positive correlation with crime rate: as wages increase, crime also increases. This could be explained by an increased reporting rate in affluent areas. Wages show a strong correlation with population density, suggesting counties with greater population have more economic opportunities and higher wages. The service wage shows a negative correlation with minorities, suggesting communities with higher minority percents are more likely to be paid minimum wage in the service industry. The state wages are also influenced by percent young male. This could be the impact of military service wages or other factors.

Because of the correlation with density most of the wage variables will not be included in the Base model but they will be considered as a major covariates in the Revised model. As a proxy for minimum wage, we will include the service sector wage (`wser`) in the Base model, as this is one area that we may be able to effectively apply policy to affect wages.

Several options were explored to aggregate the wage variables in the data set.

- Option 1: Average: We discounted using an average wage rate for each county as we are unable to know what proportion of the population in each county is paid in which category.
- Option 2: Single: We will use Service Wage `wser` as a proxy for minimum wage in our Base model. This will allow us to test our hypothesis if raising the minimum wage will reduce crime.
- Option 3: Simplified: Choose three wage variables to represent all wages across the county, with the reduced number of variables increasing parsimony in our model.
- Option 4: All: Include all wage variables in the model individually.

Option 3 was explored by looking at the collinearity between the wages in each county. If the collinearity is strong then we can employ a simplified wage model (Option 3) because we can assume

that one wage is close to being simply a linear combination of the other wages. However, referring to the correlation plot in Figure 12 below does not suggest even moderate collinearity (~80%).

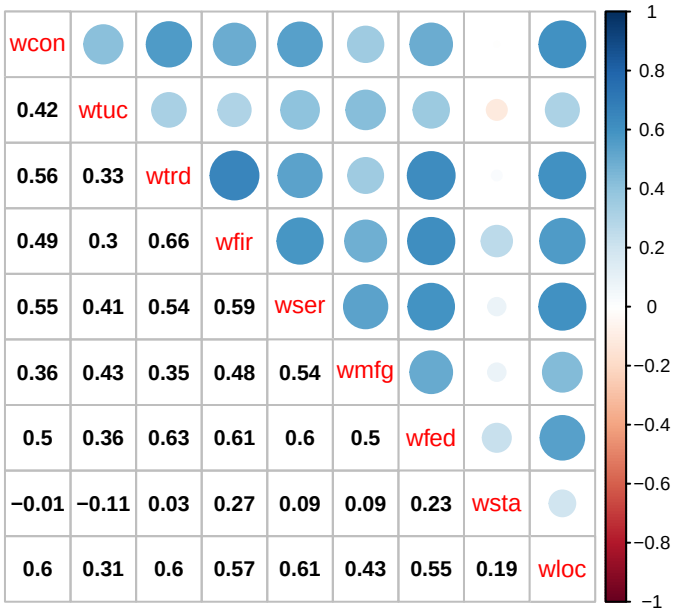

Figure 12: Correlation Plot for Wage Variables

To check for multicollinearity, we use variance inflation factor (VIF). The VIF represents the proportion of variance in one predictor explained by all the other predictors in the model. Smaller VIF indicates less multicollinearity. All VIF values are less than 2.5, which indicates little multicollinearity. Based on the combined outcome of the correlation plot and VIF, we will add all the wage predictors to our 'Revised' model. We will continue to use service wage (Option 2) in the base model and all wages in all other models (Option 4). This may decrease parsimony so a joint significance test will be conducted, see section 4.3.1.

```
> mod_wage2 = lm(log(crime_df$crmrte) ~ log(crime_df$wcon) + log(crime_df$wtuc)
+               + log(crime_df$wtrd) + log(crime_df$wfir) + log(crime_df$wser)
+               + log(crime_df$wmfg) + log(crime_df$wfed) + log(crime_df$wsta)
+               + log(crime_df$wloc))
> cat("VIF values for all wage variables:")
```

```
VIF values for all wage variables:
```

```
> vif(mod_wage2, vcov=vcovHC)
```

```
log(crime_df$wcon) log(crime_df$wtuc) log(crime_df$wtrd)
          1.881374           1.387818           2.423598
log(crime_df$wfir) log(crime_df$wser) log(crime_df$wmfg)
          2.262446           2.122328           1.802861
log(crime_df$wfed) log(crime_df$wsta) log(crime_df$wloc)
          2.170620           1.266077           2.211505
```

Finally, before we use the wage variables, we check the normality of the data. For wages we have applied a log transformation, this allows us to avoid discussing absolute dollar values for the effect

of changes to wages, and allows us to talk about % changes instead. We will use the QQ-plots of all 9 wage variables to check for normality, along with the Shapiro-Wilk's test.

```
> par(mfrow=c(3,3), ps=7, cex.axis=1.2, cex.lab=1.4,cex.main=1.4)
> qqnorm(log(crime_df$wser), main="Service Wage")
> qqline(log(crime_df$wser))
> qqnorm(log(crime_df$wcon), main="Construction Wage")
> qqline(log(crime_df$wcon))
> qqnorm(log(crime_df$wtuc), main="Trans/Util/Commun Wage")
> qqline(log(crime_df$wtuc))
> qqnorm(log(crime_df$wtrd), main="Whlsle Retail Wage")
> qqline(log(crime_df$wtrd))
> qqnorm(log(crime_df$wfir), main="Fin Ins RealEst Wage")
> qqline(log(crime_df$wfir))
> qqnorm(log(crime_df$wmfg), main="Manuf Wage")
> qqline(log(crime_df$wmfg))
> qqnorm(log(crime_df$wfed), main="Fed Gov Wage")
> qqline(log(crime_df$wfed))
> qqnorm(log(crime_df$wsta), main="State Gov Wage")
> qqline(log(crime_df$wsta))
> qqnorm(log(crime_df$wloc), main="Local Gov Wage")
> qqline(log(crime_df$wloc))
```

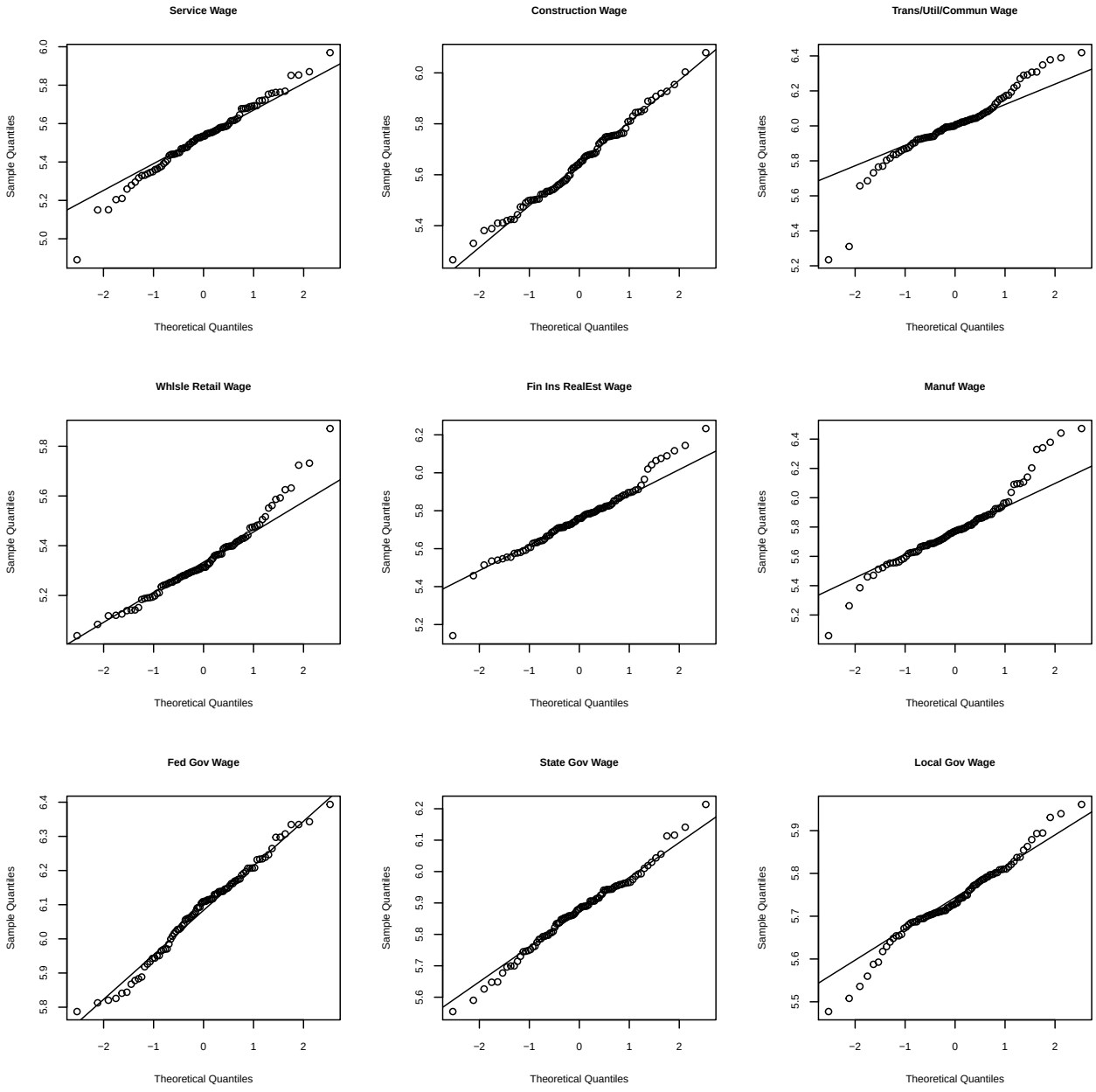

Figure 13: QQPlot of Log of Wage Variables

```
> wser_st <- shapiro.test(log(crime_df$wser))
> wcon_st <- shapiro.test(log(crime_df$wcon))
> wtuc_st <- shapiro.test(log(crime_df$wtuc))
> wtrd_st <- shapiro.test(log(crime_df$wtrd))
> wfir_st <- shapiro.test(log(crime_df$wfir))
> wmfg_st <- shapiro.test(log(crime_df$wmfg))
> wfed_st <- shapiro.test(log(crime_df$wfed))
> wsta_st <- shapiro.test(log(crime_df$wsta))
> wloc_st <- shapiro.test(log(crime_df$wloc))
```

```
> names <- c("wser","wcon","wtuc","wtrd","wfir","wmfg","wfed","wsta","wloc")
> sts <- c(round(as.numeric(wser_st[2]),5),round(as.numeric(wcon_st[2]),5),
+       round(as.numeric(wtuc_st[2]),8),round(as.numeric(wtrd_st[2]),5),
+       round(as.numeric(wfir_st[2]),5),round(as.numeric(wmfg_st[2]),5),
+       round(as.numeric(wfed_st[2]),5),round(as.numeric(wsta_st[2]),5),
+       round(as.numeric(wloc_st[2]),5))
> both <- do.call(rbind.data.frame, Map('c',names,sts))
> colnames(both) <- c("Wage_Cat","SW_pvalue")
> print(xtable(both, caption = "Shapiro-Wilks test for Normality of Wage Variables"),
+       include.rownames=FALSE)
```

% latex table generated in R 3.5.1 by xtable 1.8-2 package % Wed Apr 17 20:22:07 2019

Wage_Cat	SW_pvalue
wser	0.29316
wcon	0.8606
wtuc	3.811e-05
wtrd	0.00641
wfir	0.01702
wmfg	0.0017
wfed	0.4236
wsta	0.66582
wloc	0.13189

Table 2: Shapiro-Wilks test for Normality of Wage Variables

Note that for the Shapiro-Wilks test, the null hypothesis is that the data is normally distributed. Thus, for a significance level of $\alpha = 0.05$, any p-value less than 0.05 typically mean the data is not normal. From the QQ plots in Figure 13 and the Shapiro-Wilks test results in Table 2, we can see that log(`wtuc`), log(`wtrd`), log(`wfir`) and log(`wmfg`) are not normally distributed. However, we will make use of the Central Limit Theorem and assert that as the data set gets larger (with $n > 30$), this degree of non-normality can be tolerated.

3.6 Summary of Variables

Table 3 is a summary of all the variables and their transforms that may be used in the Regression modelling in Section 4.

Variable	Description	Transformation	Model
county	County Identifier		
year	1987	N/A	Not Included (constant)
Dependent			
crmrte	Crimes Committed / Person	log (fix neg. skew/inc. practicality)	All (dependent variable)
Independent	*Judiciary and Crime*		
prbarr	'Probability' of Arrest	log (fix neg. skew)	All (hypothesized key judiciary efficiency measure)
prbconv	'Probability' of Conviction	log (fix neg. skew)	All (hypothesized key judiciary efficiency measure)
prbpris	'Probability' of Prison Sentence	log (fix neg. skew)	None (low corr. w/ crime rate)
avgsen	Average Sentence in Days	log (fix neg. skew)	None (low corr. w/ crime rate)
mix	Offense Crime Mix	log (fix neg. skew)	None (low corr. w/ crime rate)
Independent	*Demographic*		
polpc	Police per Capita	log (fix neg. skew/inc. practicality)	None (cannot reject null H_0 (increasing police decreases crime). See sec. 5.)
density	People per square mile	log (fix neg. skew/inc. practicality)	All (high corr. w/ crime rate, hypothesized major contributor to crime rt.)
pctmin80	Percent Minority, 1980	log (inc. practicality)	Revised (hypothesized minor demographic contributor of crime)
pctymle	Percent Young Male	log (fix neg. skew/inc. practicality)	Revised (hypothesized minor demographic contributor of crime)
Independent	*Demographic(Region)*		
west	=1 if in Western N.C.	N/A	Not Included (not enough sample points to satisfy CLT)
central	=1 if in Central N.C.	N/A	Not Included (not enough sample points to satisfy CLT)
urban	=1 if in SMSA	N/A	Not Included (not enough sample points to satisfy CLT)
Independent	*Tax and Wages*		
taxpc	Tax Revenue per Capita	log (financial transformation)	All (hypothesized key wealth measure)
wcon	Construction Wage	log (inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wtuc	Transport/Util./Comm. Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wtrd	Wholesale/Retail Trade Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wfir	Finance/Ins./Real Estate Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wser	Service Industry Wage	(inc. practicality)	All (hypothesized major wealth contributor of crime, proxy for min. wage)
wmfg	Manufacturing Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wfed	Federal Gov't Employees Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wsta	State Gov't Employees Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)
wloc	Local Gov't Employees Wage	(inc. practicality)	Revised (hypothesized minor wealth contributor of crime)

Table 3: Table of Variables for Regression Modeling

4. Regression Modeling - Multivariate Analyses

4.1 Base Model

The Base model aims to seek out the main determinants of crime rate in North Carolina and into account both judicial and demographic system variables to come up with a causal explanation of crime rate. From our EDA and analyses above, we have identified population density as the variable with the strongest correlation with crime rate. The tax revenue per capita, service wage (as a proxy for minimum wage), and the 'probabilities' of arrest and conviction are hypothesized as being most actionable by policy and thus explore their effect here.

The base model, `modelA`, proposed as follows:

$$log(crmrte_A) = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(taxpc) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(prbarr) + \beta_5 \cdot log(wser) + u$$

```
> modelA <- lm(log(crime_df$crmrte) ~ log(crime_df$density)  + log(crime_df$taxpc)
+              + log(crime_df$prbconv) + log(crime_df$prbarr)  + log(crime_df$wser))
> cat("Base Model A R-Squared is: ", round(summary(modelA)$r.squared,5))
```

```
Base Model A R-Squared is:  0.57688
```

```
> cat("Base Model A adjusted R-Squared is: ", round(summary(modelA)$adj.r.squared,5))
```

```
Base Model A adjusted R-Squared is:  0.55076
```

```
> cat("Coefficients with t-tests for Base Model A are:")
```

```
Coefficients with t-tests for Base Model A are:
```

```
> coeftest(modelA, vcov=vcovHC)
```

```
t test of coefficients:

                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)           -2.985646   1.701028 -1.7552   0.08301 .
log(crime_df$density)  0.404392   0.085855  4.7102 1.011e-05 ***
log(crime_df$taxpc)    0.394983   0.198609  1.9887   0.05011 .
log(crime_df$prbconv) -0.211295   0.134547 -1.5704   0.12022
log(crime_df$prbarr)  -0.324092   0.139434 -2.3243   0.02261 *
log(crime_df$wser)    -0.463045   0.303243 -1.5270   0.13066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> modelA$AIC <- AIC(modelA)
> cat("Base Model A AIC is: ",modelA$AIC)
```

```
Base Model A AIC is:  68.17718
```

```
> cat("Base Model A VIF are:")
```

```
Base Model A VIF are:
```

```
> vif(modelA)
```

```
log(crime_df$density)   log(crime_df$taxpc) log(crime_df$prbconv)
             1.926112              1.171456              1.312143
 log(crime_df$prbarr)    log(crime_df$wser)
             1.388606              1.736678
```

```
> se.modelA = sqrt(diag(vcovHC(modelA)))
> cat("Wald Test for Base Model A is:")
```

```
Wald Test for Base Model A is:
```

```
> waldtest(modelA, vcov=vcovHC)
```

```
Wald test

Model 1: log(crime_df$crmrte) ~ log(crime_df$density) + log(crime_df$taxpc) +
    log(crime_df$prbconv) + log(crime_df$prbarr) + log(crime_df$wser)
Model 2: log(crime_df$crmrte) ~ 1
  Res.Df Df      F    Pr(>F)
1     81
2     86 -5 32.165 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For this Base model, the adjusted R-squared value is 0.55: approximately 55% of the variation in crime rate can be explained by population density, tax revenue per capita, the 'probabilities' of conviction and arrest and the service wage (as a proxy for minimum wage). Note for this model the Wald test is used to calculate the F statistic, allowing use of the heteroskedastic-robust covariance matrix.

Interpreting the coefficients:

- For a 1% population density increase, crime rate increases by 0.4%.
- For a 1% tax revenue per capita increase, crime rate increases by 0.4%.
- For a 1% conviction to arrest ratio increase, crime rate decreases by 0.2%.
- For a 1% arrests to offenses ratio increase, crime rate decreases by 0.3%.
- For a 1% service wage increase, crime rate decreases by 0.5%.

To check MLR assumptions for this model, the residuals are plotted in Figure 14:

```
> par(mfrow=c(2,2), ps=7, cex.axis=1.2, cex.lab=1.4,cex.main=1.4)
> plot(modelA, cex.main = 0.9, cex.lab = 0.9, cex.axis = 0.9, cex.sub = 0.5)
```

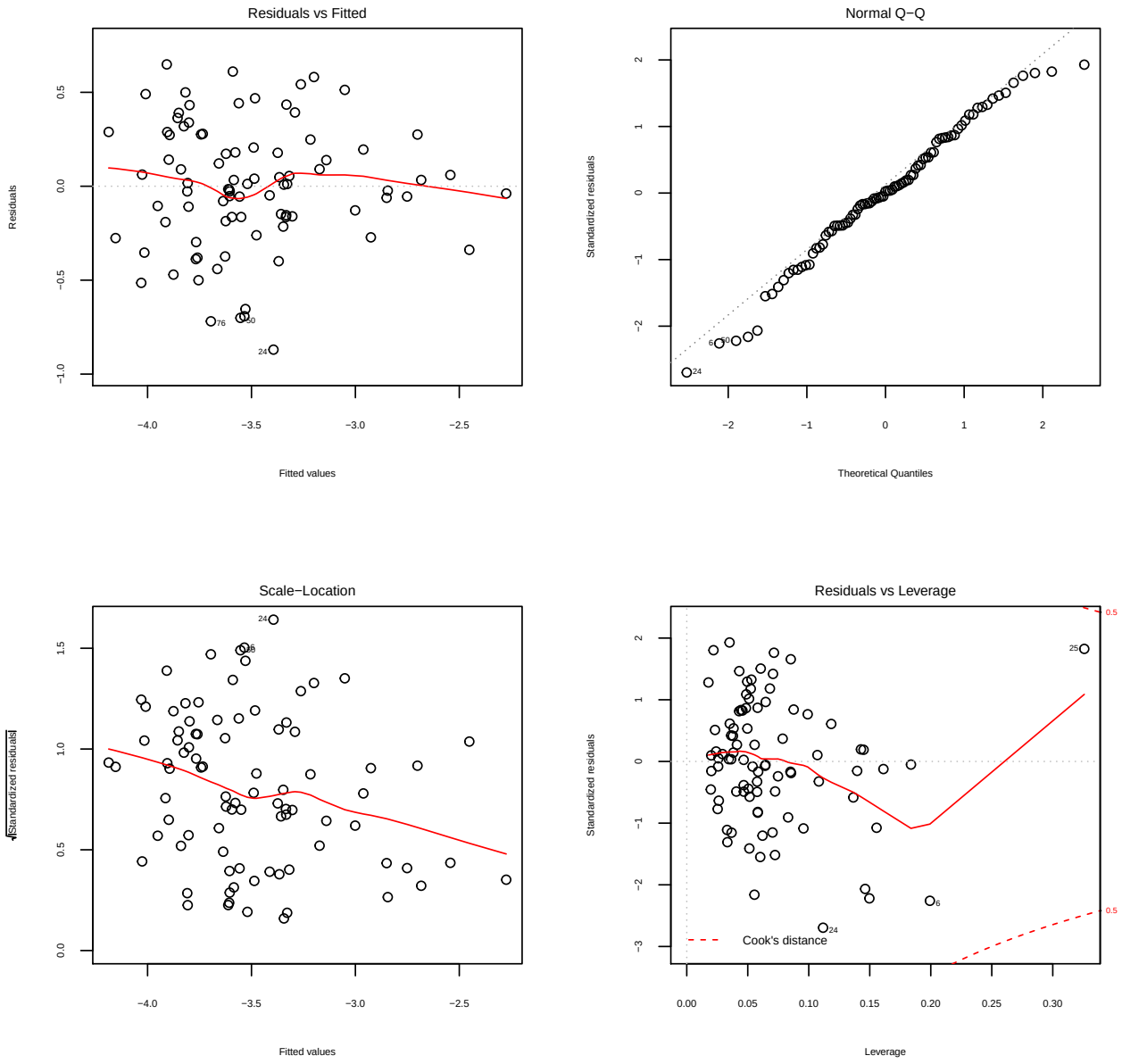

Figure 14: Base Model Residual Plots for OLS Assumption Evaluation Model A

Note the following significant points for ModelA:

- The residuals are reasonably normally distributed, as seen in the QQ plot.
- The Conditional Mean of the residuals is reasonably close to zero.
- The model scale-location plot and the Breusch-Pagan test (following) show we must reject homoskedasticity. This is accounted for by using heteroskedastic-robust standard errors.

We can say that this model meets the MLR assumptions and thus has consistent and 'best' linear unbiased estimators.

```
> bptest(modelA)
```

```
    studentized Breusch-Pagan test

data:  modelA
BP = 21.296, df = 5, p-value = 0.000712
```

4.2 Revised Model - Major Covariates

The Revised model, `modelB`, includes the variables from ModelA, with the addition of the other wage variables. These additional explanatory variables were identified during our EDA.

$$\begin{aligned} log(crmrte_B) = \beta_0 &+ \beta_1 \cdot log(density) + \beta_2 \cdot log(taxpc) + \beta_3 \cdot log(prbconv) \\ &+ \beta_4 \cdot log(prbarr) + \beta_5 \cdot log(wser) + \beta_6 \cdot log(wcon) + \beta_7 \cdot log(wtuc) \\ &+ \beta_8 \cdot log(wtrd) + \beta_9 \cdot log(wfir) + \beta_{10} \cdot log(wmfg) + \beta_{11} \cdot log(wfed) \\ &+ \beta_{12} \cdot log(wsta) + \beta_{13} \cdot log(wloc) + u \end{aligned}$$

```
> modelB <- lm(log(crime_df$crmrte) ~  log(crime_df$density) + log(crime_df$wser)
+              + log(crime_df$prbarr) + log(crime_df$prbconv) + log(crime_df$taxpc)
+              + log(crime_df$wcon) + log(crime_df$wtuc) + log(crime_df$wtrd)
+              + log(crime_df$wfir) + log(crime_df$wmfg) + log(crime_df$wfed)
+              + log(crime_df$wsta) + log(crime_df$wloc))
> cat("Revised Model B R-Squared is: ", round(summary(modelB)$r.squared,5))
```

```
Revised Model B R-Squared is:  0.636
```

```
> cat("Revised Model B adjusted R-Squared is: ", round(summary(modelB)$adj.r.squared,5))
```

```
Revised Model B adjusted R-Squared is:  0.57118
```

```
> cat("Coefficients with t-tests for Model B are:")
```

```
Coefficients with t-tests for Model B are:
```

```
> coeftest(modelB, vcov=vcovHC)
```

```
t test of coefficients:

                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -9.145865   4.309436 -2.1223 0.037207 *
log(crime_df$density)  0.296996   0.104099  2.8530 0.005631 **
log(crime_df$wser)    -0.728914   0.363894 -2.0031 0.048880 *
log(crime_df$prbarr)  -0.367067   0.154589 -2.3745 0.020204 *
log(crime_df$prbconv) -0.269649   0.144235 -1.8695 0.065561 .
log(crime_df$taxpc)    0.394117   0.238501  1.6525 0.102732
log(crime_df$wcon)     0.013589   0.326058  0.0417 0.966871
log(crime_df$wtuc)    -0.068580   0.379156 -0.1809 0.856966
log(crime_df$wtrd)    -0.119920   0.470283 -0.2550 0.799444
log(crime_df$wfir)    -0.361330   0.438585 -0.8239 0.412705
```

```
log(crime_df$wmfg)     0.108730   0.206019  0.5278 0.599262
log(crime_df$wfed)     1.206860   0.574032  2.1024 0.038965 *
log(crime_df$wsta)    -0.170385   0.323453 -0.5268 0.599950
log(crime_df$wloc)     0.630369   0.702750  0.8970 0.372666
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> modelB$AIC <- AIC(modelB)
> cat("Revised Model B AIC is: ", modelB$AIC)
```

```
Revised Model B AIC is:  71.084
```

```
> se.modelB = sqrt(diag(vcovHC(modelB)))
```

This Revised Model includes the effect of all the remaining wages. Note that each wage grouping was added independently. We cannot properly sum nor average the wages, as we have no information to indicate which wage grouping may be more or less represented in each county, thus we included each wage as a separate variable.

Comparing this Revised model to our Base model, the adjusted R-squared has improved from 0.55 to 0.57: by adding all the wage variables we have increased our ability to explain the variation in crime rate from 55% to 57%. When we analyse wages, depending on the specific wage, the crime rate effect varies, from a decline of 0.73% in crime rate for a 1% increase in service sector weekly wages to an increase of 1.2% in crime rate for a 1% increase in federal government weekly wages.

Checking on MLR assumptions for our models, see Figure 15:

```
> par(mfrow=c(2,2), ps=7, cex.axis=1.2, cex.lab=1.4,cex.main=1.4)
> plot(modelB, cex.main = 0.9, cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.0)
```

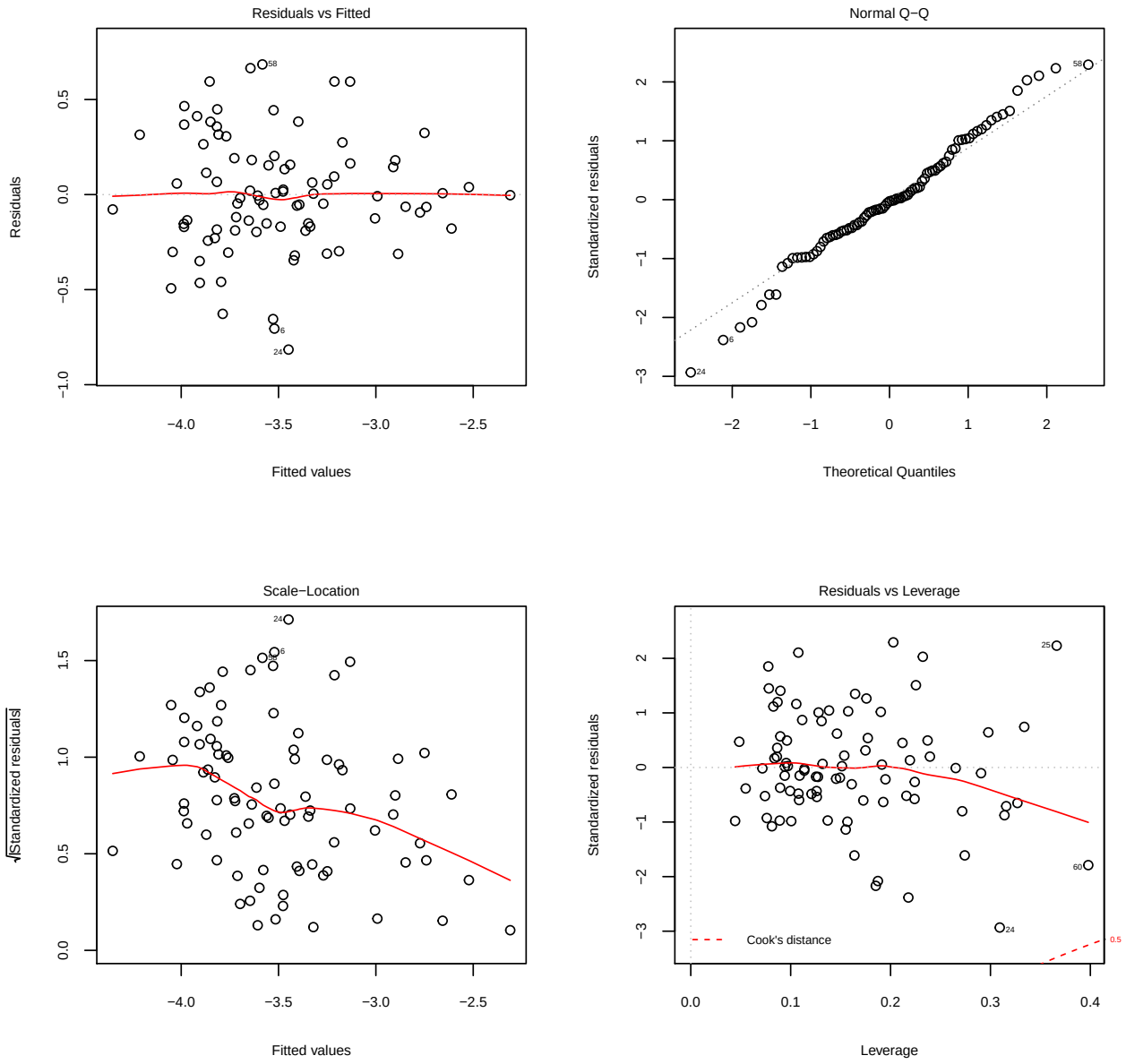

Figure 15: Revised Model Residual Plots for OLS Assumption Evaluation for Model B

With this model, we note significant changes:

- MLR4: The inclusion of more explanatory variables improves the mean residual, making it closer to zero.
- MLR5 Reviewing the scale-location plot, this model appears to be heteroskedastic. The Breusch-Pagan test below allows us to reject homoskedasticity. Our choice of heteroskedastic-robust standard errors is thus not too conservative. Therefore the specifications for Revised Model B are consistent and represent the Best Linear Unbiased Estimators for effect of the independent variables on crime rate.

```
> bptest(modelB)


    studentized Breusch-Pagan test

data:  modelB
BP = 28.821, df = 13, p-value = 0.00694
```

4.3.1 Test for combined significance of wage variables

The assumption that the service wage is a proxy for minimum wage and thus attributing all the major economic benefit to this one sector of the economy is an assumption worth further exploration. As can be seen from our analysis, the wage variables are not individually significant contributors to the crime rate in the Revised model. However, we want to check to see if they have joint significance due to their likely importance in the political arena. We test here using exclusion restriction (Wald test), via the Linear Hypothesis function. In this case the test results have an F value of 0.12, and we cannot reject the null hypothesis that the wage variables are jointly insignificant, suggesting that their inclusion has no effect on the model nor the explanation of the crime rate.

```
> linearHypothesis(modelB, c('log(crime_df$wcon)=0', 'log(crime_df$wtuc)=0',
+                            'log(crime_df$wtrd)=0', 'log(crime_df$wfir)=0',
+                            'log(crime_df$wser)=0', 'log(crime_df$wmfg)=0',
+                            'log(crime_df$wfed)=0', 'log(crime_df$wsta)=0',
+                            'log(crime_df$wloc)=0'), vcov=vcovHC)

Linear hypothesis test

Hypothesis:
log(crime_df$wcon) = 0
log(crime_df$wtuc) = 0
log(crime_df$wtrd) = 0
log(crime_df$wfir) = 0
log(crime_df$wser) = 0
log(crime_df$wmfg) = 0
log(crime_df$wfed) = 0
log(crime_df$wsta) = 0
log(crime_df$wloc) = 0

Model 1: restricted model
Model 2: log(crime_df$crmrte) ~ log(crime_df$density) + log(crime_df$wser) +
    log(crime_df$prbarr) + log(crime_df$prbconv) + log(crime_df$taxpc) +
    log(crime_df$wcon) + log(crime_df$wtuc) + log(crime_df$wtrd) +
    log(crime_df$wfir) + log(crime_df$wmfg) + log(crime_df$wfed) +
    log(crime_df$wsta) + log(crime_df$wloc)

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
```

```
1     82
2     73  9 1.6305 0.1225
```

4.4 Inclusive Model

The Inclusive Model, `modelC`, includes all the variables in Model B plus the demographic variables `pctmin80` and `pctymle` as follows:

$$\begin{aligned} log(crmrte_C) = \beta_0 &+ \beta_1 \cdot log(density) + \beta_2 \cdot logg(taxpc) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(prbarr) + \beta_5 \cdot log(wcon) \\ &+ \beta_6 \cdot log(wtuc) + \beta_7 \cdot log(wtrd) + \beta_8 \cdot log(wfir) + \beta_9 \cdot log(wser) + \beta_{10} \cdot log(wmfg) \\ &+ \beta_{11} \cdot log(wfed) + \beta_{12} \cdot log(wsta) + \beta_{13} \cdot log(wloc) \\ &+ \beta_{14} \cdot log(pctmin80) + \beta_{15} \cdot log(pctymle) + u \end{aligned}$$

This model contains most variables presented in the original data set with the following exceptions:

- `west/central/urban`: as this is strongly correlated with the population density, see the discussion in Section 3.4.1
- `mix`: as there is no information that would give us an actionable item nor do we know how this was developed
- `prbpris` and `avgsen`: these variables are strongly correlated with `prbconv` and `prbarr` variables and do not add any new information

```
> modelC <- lm(log(crime_df$crmrte) ~  log(crime_df$density) + log(crime_df$wser)
+              + log(crime_df$prbarr) + log(crime_df$prbconv)
+              + log(crime_df$taxpc) + log(crime_df$wcon) + log(crime_df$wtuc)
+              + log(crime_df$wtrd) + log(crime_df$wfir) + log(crime_df$wmfg)
+              + log(crime_df$wfed) + log(crime_df$wsta) + log(crime_df$wloc)
+              + log(crime_df$pctmin80) + log(crime_df$pctymle))
> cat("Inclusive Model C R-Squared is: ", round(summary(modelC)$r.squared,5))
```

```
Inclusive Model C R-Squared is:  0.78149
```

```
> cat("Inclusive Model C adjusted R-Squared is: ", round(summary(modelC)$adj.r.squared,5))
```

```
Inclusive Model C adjusted R-Squared is:  0.73532
```

```
> cat("Coefficients with t-tests for Inclusive Model C are:")
```

```
Coefficients with t-tests for Inclusive Model C are:
```

```
> coeftest(modelC, vcov=vcovHC)
```

```
t test of coefficients:

                        Estimate Std. Error t value  Pr(>|t|)
(Intercept)            -6.616600   4.036743 -1.6391  0.105618
log(crime_df$density)   0.281652   0.089924  3.1321  0.002521 **
log(crime_df$wser)     -0.506383   0.378597 -1.3375  0.185320
```

```
log(crime_df$prbarr)   -0.406536   0.126668 -3.2095  0.001997 **
log(crime_df$prbconv)  -0.317017   0.112807 -2.8103  0.006392 **
log(crime_df$taxpc)     0.276086   0.332413  0.8305  0.409013
log(crime_df$wcon)      0.083445   0.253070  0.3297  0.742574
log(crime_df$wtuc)      0.110061   0.351355  0.3132  0.755011
log(crime_df$wtrd)     -0.114730   0.349690 -0.3281  0.743808
log(crime_df$wfir)     -0.226785   0.415559 -0.5457  0.586959
log(crime_df$wmfg)      0.137927   0.215146  0.6411  0.523530
log(crime_df$wfed)      0.782241   0.386420  2.0243  0.046699 *
log(crime_df$wsta)     -0.342639   0.267666 -1.2801  0.204678
log(crime_df$wloc)      0.249278   0.655542  0.3803  0.704886
log(crime_df$pctmin80)  0.216939   0.048816  4.4440 3.183e-05 ***
log(crime_df$pctymle)   0.285726   0.221949  1.2873  0.202154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> se.modelC = sqrt(diag(vcovHC(modelC)))
> modelC$AIC <- AIC(modelC)
> cat("Inclusive Model C AIC is: ", modelC$AIC)
```

```
Inclusive Model C AIC is:  30.68651
```

The Inclusive model has an adjusted R-squared of 0.74; approximately 74% of the variation in crime rate can be explained by population density, tax revenue per capita, the 'probabilities' of conviction and arrest, all wages and percent minority and percent young male. Adding in the percent young male and percent minority have increased the Adjusted R-squared from 0.57 in Model B to 0.74 - with a concomitant decrease in AIC from 71 to 31. This would indicate that these two demographic variables have a large effect on the goodness of fit of the model.

Interpreting the coefficients:

- For a 1% population density increase, crime rate increases by 0.3%.
- For a 1% tax revenue per capita increase, crime rate increases by 0.3%.
- For a 1% conviction to arrest ratio increase, crime rate decreases by 0.3%.
- For a 1% arrests to offenses ratio increase, crime rate decreases by 0.4%.
- For a 1% service wage increase, crime rate decreases by 0.5%.
- For a 1% percent minorities increase, crime rate increases by 0.2%.
- For a 1% percent young males increase, crime rate increases by 0.3%.

The remaining wage variables have a varying effect on crime rate in this model: for a 1% increase in federal wages, the crime rate increases by 0.8% and at the other end of the spectrum for a 1% increase in state wages, there is a 0.3% decrease in crime rate.

MLR assumptions for the model: see the residuals for the Inclusive Model in Figure 16:

```
> par(mfrow=c(2,2), ps=7, cex.axis=1.2, cex.lab=1.4,cex.main=1.4)
> plot(modelC, cex.main = 1.2, cex.lab =1.2, cex.axis = 1.2, cex.sub = 1.2)
```

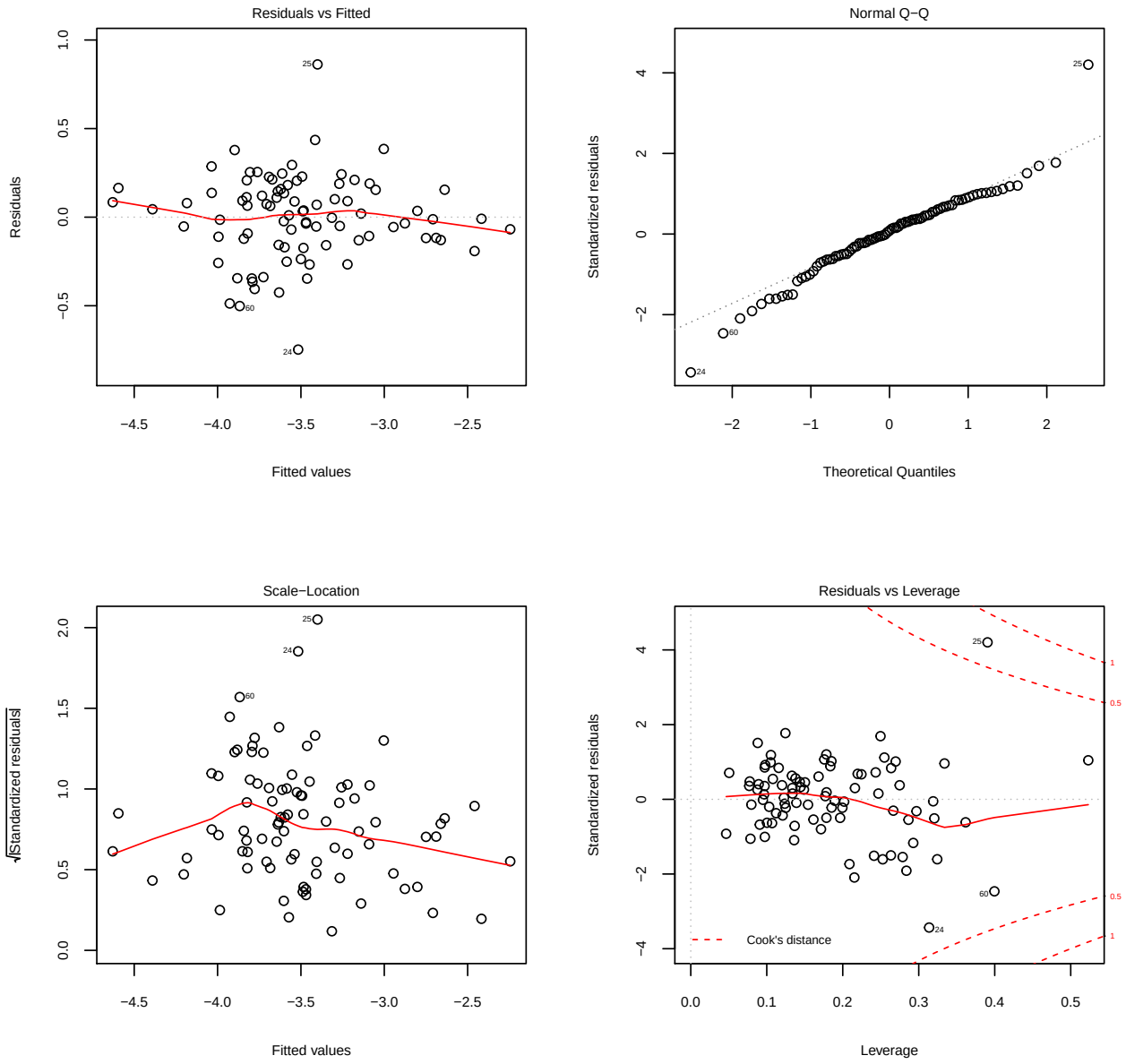

Figure 16: Inclusive Model Residual Plots for OLS Assumption Evaluation Model C

Note significant changes for this model in Figure 16:

- Considering the conditional mean of the residuals, the inclusion of more explanatory variables has increased the mean residual, moving it away from zero. These variables appear to have introduced endogeneity into our model suggesting the additional variables add more error than is useful, thus they may not be main determinants of crime.

- This model appears to also have worsened the scale-location plot and the Breusch-Pagan test (following) shows we must reject homoskedasticity. This is accounted for by using heteroskedastic-robust standard errors.

```
> bptest(modelC)
```

```
    studentized Breusch-Pagan test

data:  modelC
BP = 35.232, df = 15, p-value = 0.002279
```

4.5 Final Model

The Final model, `modelD`, includes all the variables in Model A, with the addition of percent minority and percent young males included.

$$log(crmrte_D) = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(taxpc) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(prbarr) + \beta_5 \cdot log(wser) + \beta_6 \cdot log(pctmin80) + \beta_7 \cdot log(pctymle) + u$$

```
> modelD <- lm(log(crime_df$crmrte) ~ log(crime_df$density) + log(crime_df$wser)
+              + log(crime_df$prbarr) + log(crime_df$prbconv) + log(crime_df$taxpc)
+              + log(crime_df$pctmin80) + log(crime_df$pctymle))
> cat("Final Model D R-Squared is: ", round(summary(modelD)$r.squared,5))
```

```
Final Model D R-Squared is:  0.74884
```

```
> cat("Final Model D adjusted R-Squared is: ", round(summary(modelD)$adj.r.squared,5))
```

```
Final Model D adjusted R-Squared is:  0.72659
```

```
> cat("Coefficients with t-tests for Final Model D are:")
```

```
Coefficients with t-tests for Final Model D are:
```

```
> coeftest(modelD, vcov=vcovHC)
```

```
t test of coefficients:

                        Estimate Std. Error t value  Pr(>|t|)
(Intercept)            -4.436475   1.436083 -3.0893  0.002769 **
log(crime_df$density)   0.346899   0.071199  4.8722 5.576e-06 ***
log(crime_df$wser)     -0.207497   0.290436 -0.7144  0.477065
log(crime_df$prbarr)   -0.394513   0.110162 -3.5812  0.000589 ***
log(crime_df$prbconv)  -0.288761   0.098720 -2.9251  0.004493 **
log(crime_df$taxpc)     0.272711   0.307538  0.8868  0.377902
log(crime_df$pctmin80)  0.234004   0.043330  5.4006 6.800e-07 ***
log(crime_df$pctymle)   0.142471   0.139462  1.0216  0.310101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> se.modelD = sqrt(diag(vcovHC(modelD)))
> modelD$AIC <- AIC(modelD)
> cat("Final Model D AIC is: ", modelD$AIC)
```

```
Final Model D AIC is:  26.79992
```

```
> waldtest(modelA,modelD, vcov=vcovHC)
```

```
Wald test

Model 1: log(crime_df$crmrte) ~ log(crime_df$density) + log(crime_df$taxpc) +
    log(crime_df$prbconv) + log(crime_df$prbarr) + log(crime_df$wser)
Model 2: log(crime_df$crmrte) ~ log(crime_df$density) + log(crime_df$wser) +
    log(crime_df$prbarr) + log(crime_df$prbconv) + log(crime_df$taxpc) +
    log(crime_df$pctmin80) + log(crime_df$pctymle)
  Res.Df Df      F    Pr(>F)
1     81
2     79  2 17.319 5.797e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing this model against model A, we see that the effect of adding in `pctymle` and `pctmin80` is significant; the p-value for the Wald test is < 0.05. The AIC is significantly lower and the adjusted R-squared is higher than model A, showing a large increase in ability to explain variation in crime rate with few added variables. This is an excellent example of variable parsimony.

The adjusted R-squared value for the Final model is 0.73: approximately 73% of the variation in crime rate can be explained by population density, tax revenue per capita, the 'probabilities' of conviction and arrest, the service wage (as a proxy for minimum wage), the percent minority and the percent young male.

Interpreting the coefficients:

- For a 1% population density increase, crime rate increases by 0.3%.
- For a 1% tax revenue per capita increase, crime rate increases by 0.3%.
- For a 1% conviction to arrest ratio increase, crime rate decreases by 0.3%.
- For a 1% arrests to offenses ratio increase, crime rate decreases by 0.4%.
- For a 1% service wage increase, crime rate decreases by 0.2%.
- For a 1% fraction of minorities increase, crime rate increases by 0.2%.
- For a 1% fraction of young males increase, crime rate increases by 0.1%.

The detailed analysis of the MLR assumptions and Standard Error analysis for Model D will be carried out in Section 6 below. The conclusion from that analysis is that the estimators from this model can be used for causal inference with confidence.

4.6 Comparison of Models

See Table 3 for a comparison of the 4 models. Note that the Final Model has the lowest (best) AIC score compared with the Base Model, with an increased Adjusted R-squared and appears to be the best trade-off between model complexity and variable parsimony. This table begins to answer our

initial research question, which was to identify the top determinants of crime rate that we will be addressed in our campaign. Based on the results from our model, we plan to address population density, conviction rate, and minimum wage with direct policy recommendations, and discuss the implications of tax revenue and demographics.

```
> stargazer(modelA, modelB, modelC, modelD,
+           column.sep.width="1pt",
+           single.row=TRUE,
+           float.env = "sidewaystable",
+           type = "latex",
+           align=TRUE,
+           title = "Linear Models of Effects on North Carolina Crime Rate",
+           dep.var.caption = "Multivariate Models",
+           dep.var.labels = "log(Crimes Per Person)",
+           column.labels = c("Base","Revised","Inclusive", "Final"),
+           se = list(se.modelA, se.modelB, se.modelC, se.modelD),
+           keep.stat = c("aic","rsq","adj.rsq","n","f"),
+           model.numbers = FALSE,
+           star.cutoffs = c(0.05, 0.01, 0.001),
+           no.space=TRUE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Apr 17, 2019 - 20:22:08 % Requires LaTeX packages: dcolumn rotating

Table 4: Linear Models of Effects on North Carolina Crime Rate

	Multivariate Models			
	log(Crimes Per Person)			
	Base	Revised	Inclusive	Final
density)	0.404*** (0.086)	0.297** (0.104)	0.282** (0.090)	0.347*** (0.071)
taxpc)	0.395* (0.199)	0.394 (0.239)	0.276 (0.332)	0.273 (0.308)
wcon)		0.014 (0.326)	0.083 (0.253)	
wtuc)		−0.069 (0.379)	0.110 (0.351)	
wtrd)		−0.120 (0.470)	−0.115 (0.350)	
wfir)		−0.361 (0.439)	−0.227 (0.416)	
wmfg)		0.109 (0.206)	0.138 (0.215)	
wfed)		1.207* (0.574)	0.782* (0.386)	
wsta)		−0.170 (0.323)	−0.343 (0.268)	
wloc)		0.630 (0.703)	0.249 (0.656)	
pctmin80)			0.217*** (0.049)	0.234*** (0.043)
pctymle)			0.286 (0.222)	0.142 (0.139)
prbconv)	−0.211 (0.135)	−0.270 (0.144)	−0.317** (0.113)	−0.289** (0.099)
prbarr)	−0.324* (0.139)	−0.367* (0.155)	−0.407** (0.127)	−0.395*** (0.110)
wser)	−0.463 (0.303)	−0.729* (0.364)	−0.506 (0.379)	−0.207 (0.290)
Constant	−2.986 (1.701)	−9.146* (4.309)	−6.617 (4.037)	−4.436** (1.436)
Observations	87	87	87	87
R^2	0.577	0.636	0.781	0.749
Adjusted R^2	0.551	0.571	0.735	0.727
Akaike Inf. Crit.	68.177	71.084	30.687	26.800
F Statistic	22.087*** (df = 5; 81)	9.811*** (df = 13; 73)	16.928*** (df = 15; 71)	33.649*** (df = 7; 79)
Note:				*p<0.05; **p<0.01; ***p<0.001

5. Omitted Variables Discussion

While our Final model explains approximately 73% of the crime rate in North Carolina, this leaves 27% remaining that should be explained by external factors not included in our data set. While some nominal amount of the crime rate will never be explained and can be attributed to randomness, it is crucial to explore some of the factors that were omitted from our analysis that could be contributing factors to the crime rate in North Carolina.

5.1 Police Per Capita

While we actually found a positive correlation between crime rate and police per capita (higher police per capita is correlated with higher crime rates), we expect that the data set may not capture all individuals with arrest powers, including sheriffs, county police, and federal law enforcement. A proper analysis on the relationship between police per capita and crime rate cannot be conducted without confirming inclusion of all individuals that can be considered law enforcement. Additionally, a better understanding of the relationship between crime rate and police presence should be understood. For example, is there more crime reported because police are more visible, or because there is actually more crime? Further, we imagine that the relationship may be causal in the reverse, meaning that in areas of high crime, governments invest in adding police forces with the intent of reducing crime. However, this may have the effect of causing a high correlation between police presence and high crime rate.

If the current data set does not include all members of law enforcement, and we were to add them to the analysis in a second pass, we expect it would further increase the direction of bias with this variable, meaning it would be even more correlated with high rates of crime. However again we do not expect this relationship to be causal in the sense that removing law enforcement would reduce crime. Thus, we predict the beta value of this omitted variable in the model to be large despite the high positive correlation. A time series analysis may actually reveal that the beta value for this variable is indeed negative, meaning more police presence over time has a negative effect on crime rate, despite the positive correlation. Further information on the inclusivity of the definition of 'police' is needed in order to test these hypotheses in a follow-up analysis.

5.2 Weather

Another factor for consideration is weather. Extreme conditions and errant weather patterns are known to have effects on crime rates and have been widely studied. North Carolina is located in an area that experiences frequent hurricanes. Looking at historical data, one major hurricane occurred in 1987 with a single death, which may have had an impact on some county crime rates. We would expect a spike in crime before and after extreme weather events in a county as opportunistic offenders take advantage of abandoned property.

This omitted variable may cause some of the effects defined in this paper to be slightly reduced, or it may be able to explain some of the 27% of unexplained variation. We expect that the beta coefficient of this omitted variable would be positive in the non-zero direction, having the aforementioned effect of reducing the effect of the independent variables in this analysis. The effects of climate change are reducing stability in weather conditions globally, so if we were to further investigate and

discover a link to crime rate, our platform could include policies aimed at mitigating the effects of climate change.

5.3 Geography and Climate

Geography and climate could be another omitted variable. It is hypothesized that counties may experience higher crime rates due to warmer weather as a result of their geographic location. Areas with milder weather may also be more densely populated. The inclusion of this relationship would mean that a positive correlation between mild weather and density and crime would move the beta value for density in a positive direction, away from zero, increasing its statistical significance.

5.4 Minorities

While the data set did include a breakdown of the minority representation in each county in North Carolina, perhaps more granular data and/or more recent data (rather than 7 years out of date with the panel data) would allow us to establish a correlation or causal relationship. Furthermore, the data element included does not discuss the relative numbers of minority groups and the breakdown of each group included in each county which certainly could be relevant.

Increased percentage of minorities is expected to be positively correlated with crime rate, thus our current minority slope factor may be too low. Another factor that would assist the analysis is further detail on the type of minorities present in the area as the presence of some groups may increase crime, indicating that different minority groups may have higher or lower (even negative or positive) beta values on their respective biases, and these values would be exacerbated based on the minority mix of a particular region However, ethically we must understand that this may be a proxy for another causal factor, such as poverty or inter-generational trauma which are expected to increase crime rate.

5.5 Poverty and Unemployment

Some wage and tax revenue data was provided in the data set, however this fails to account for the unemployed and other factors contributing to the rate of poverty in any given county. While the rate of poverty will likely have high collinearity with the wage data provided, this cannot yet be determined. Further, even if highly correlated, poverty is likely to be a more relevant causal factor in the rate of crime, especially youth poverty.

This is an indication that the beta coefficient for the omitted variable of poverty rate in a particular region is likely to be positive, and may have a high value (and large influence on the alpha) thus possibly reducing the wage variable significance by moving it towards zero (i.e positively for the service wage which has a negative beta). It is expected that employment rate will be negatively correlated with the crime rate, our analysis uses wages as a proxy but this could miss people on social security and other government programs and thus bias our model to attributing higher crime rates to wages instead of other factors.

5.6 Time

One important point of observation with this data set is that we only were able to look at a single cross-section of a single year of data. If we were able to investigate the same variables over time, we would likely be able to prove stronger correlation and causality of different variables. We imagine that crime rates are decreasing in the state over time. Thus, an analysis of counties in which crime rates are decreasing vs. those in which crime is increasing could provide insight into some policy recommendations that could be effective in reducing crime.

5.7 Politics, Culture and Society

In the US, the political party in power and its influence on wages, tax revenue, and policies can be highly influential on crime, incorporating this data could also reveal which policies have historically been beneficial or detrimental to crime in North Carolina. This is another instance where time series data would be useful to look at. Related but not exactly the same is the cultural and social climate of the time. In the wake of the 2016 general election in the United States, there was a widespread increase in hate crimes in the country.

This is just one example of the social and cultural impact on crime rate, which we did not get a chance to study in this analysis. We expect that times of high conflict and turmoil either politically or socially would increase crime, meaning the beta value of this omitted variable is positive, but we cannot determine to what degree without first quantifying and further studying the variable.

5.8 Public Works and Facilities

Another factor excluded from this data is public works and facilities. For example, there is a widely-held belief that mass transit systems attract more crime, but several studies in major population centers have mostly dispelled these ideas. So while we might find a high correlation between public works and crime, we expect this to be more so attributed to the population density that was already identified in our model as highly linked with high rates of crime.

Another aspect of the discussion around public works is crime reporting. If there are easier methods of crime reporting, we expect this would eventually decrease crime rate as offenders worry about being reported, but we would need to investigate whether it would only affect the rate of arrest and conviction rather than crime rate overall. If these hypotheses were to stand, the beta values of variables indicating increased public facilities and crime reporting solutions would be negative in direction on their effect on the alpha, crime rate. Determining this factor would likely require time series data in order to draw any causal inferences.

We believe some of these factors could contribute to the 27% of the crime rate that remains unexplained by our models. A follow-up analysis would incorporate some or all of these factors into our models in order to get closer to an adjusted R-squared value of 1.

6. Standard Error and MLR assumptions:

6.1 Multiple Linear Regression Assumptions for ModelD Specification

The model with the best balance of complexity and parsimony and most actionable items is Model D, our Final model specification. The following is a detailed analysis of the Multiple Linear Regression assumptions with regard to the specification of Model D

MLR 1: Linearity:

The model is specified to be linear in the parameters, thus this assumption must be true.

$$log(crmrte_D) = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(taxpc) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(prbarr) + \beta_5 \cdot log(wser) + \beta_6 \cdot log(pctmin80) + \beta_7 \cdot log(pctymle) + u$$

MLR 2: Random Sampling

Referring to Table 4 above, we can see that 87 of 100 North Carolina counties are represented in the final model. The data was originally sourced as a panel; that is, the original data was taken as a time series as well as across all the counties in North Carolina. This data, however, is a snapshot from the panel for the year 1987. Each record in the data set pertains to a single county. This introduces inherent clustering in the data because the measurements are taken as a statistic for the entire county. This may mean that some variables are not measured with sufficient granularity to be truly random. Additionally, gerrymandering may have affected the placement of county lines, which could affect the randomness of the data.

For the purposes of this study, however, we will assume that the large sample size of 87 counties is sufficient to ensure that any one sample picked from all the counties will be random. Finally, as discussed in the demographic study in section 3.4, the fact that rural counties are weighted the same way as urban counties may cause the randomness assumption to be violated, although we attempt to avoid this by using population density as a proxy for location. Demographics are often related to affluence/poverty and may be highly variable within the county's suburban, rural and urban areas. This study is limited to county level data set which may cause some of the variation in the model.

MLR 3: No Perfect Collinearity

```
> mlr <- crime_df[,c(1,2)]
> mlr$logcrmrte <- log(crime_df$crmrte)
> mlr$logprbarr <- log(crime_df$prbarr)
> mlr$logprbconv <- log(crime_df$prbconv)
> mlr$logdensity <- log(crime_df$density)
> mlr$logtaxpc <- log(crime_df$taxpc)
> mlr$logwser <- log(crime_df$wser)
> mlr$logpctmin80 <- log(crime_df$pctmin80)
> mlr$logpctymle <- log(crime_df$pctymle)
>
> pairs.panels(mlr[,4:10], scale=FALSE, density=TRUE, digits=2, method="pearson",
+              hist.col='blue',rug=FALSE,breaks=20, cex.labels=7)
```

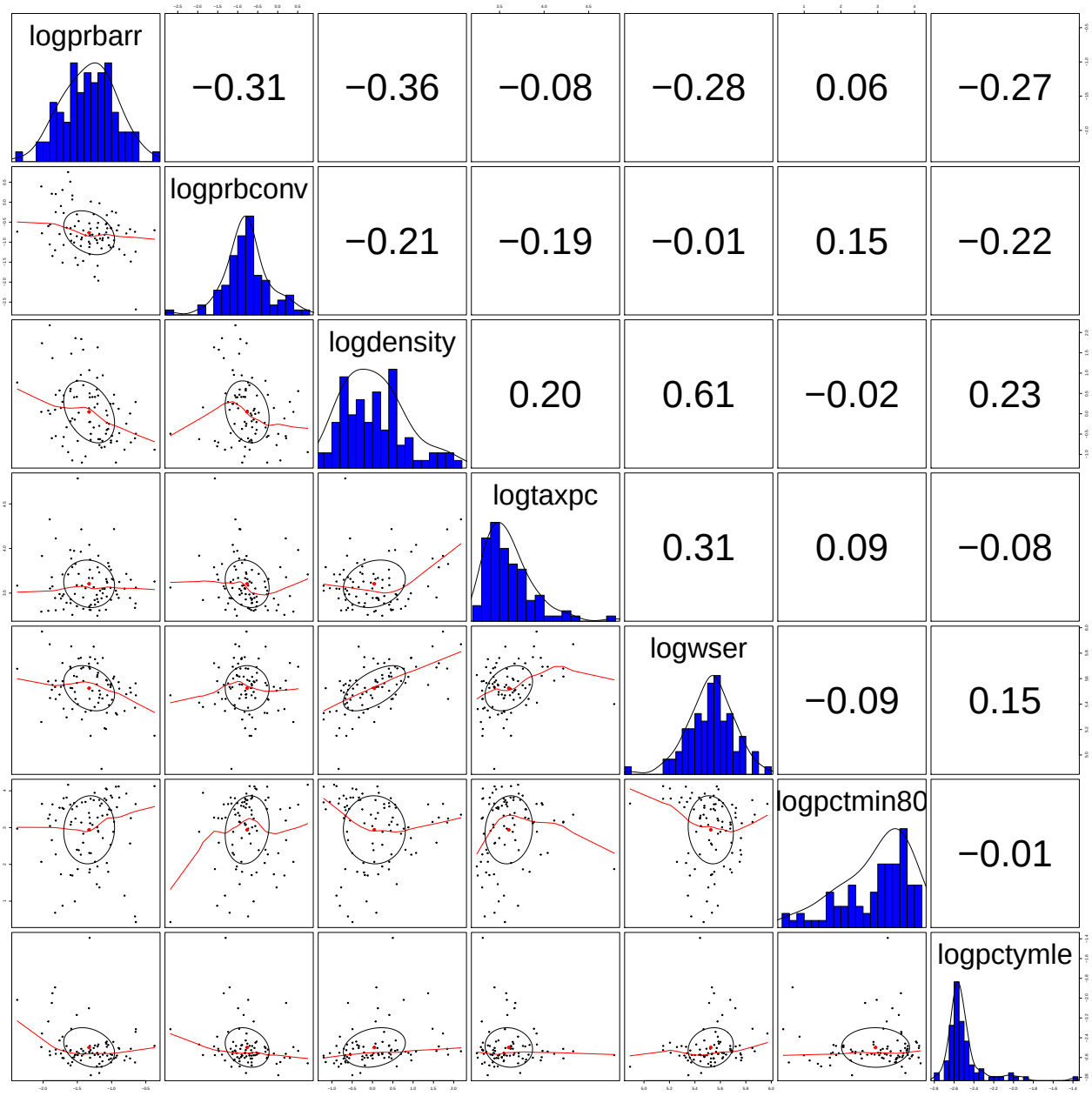

Figure 17: MLR Assumptions Analysis of Collinearity Model D

Referring to Figure 17 above, we can see that there are no correlation coefficients of 1 or -1 amongst our independent variables, indicating perfect collinearity. In fact, there are no large collinearities found even amongst variables that one might expect, such as `prbconv` and `prbarr`. The highest collinearity pair of independent variables is population density and service wage, however at 0.61 (R^2 of 0.37), this should not be a large source of error.

MLR 4: Zero Conditional Mean

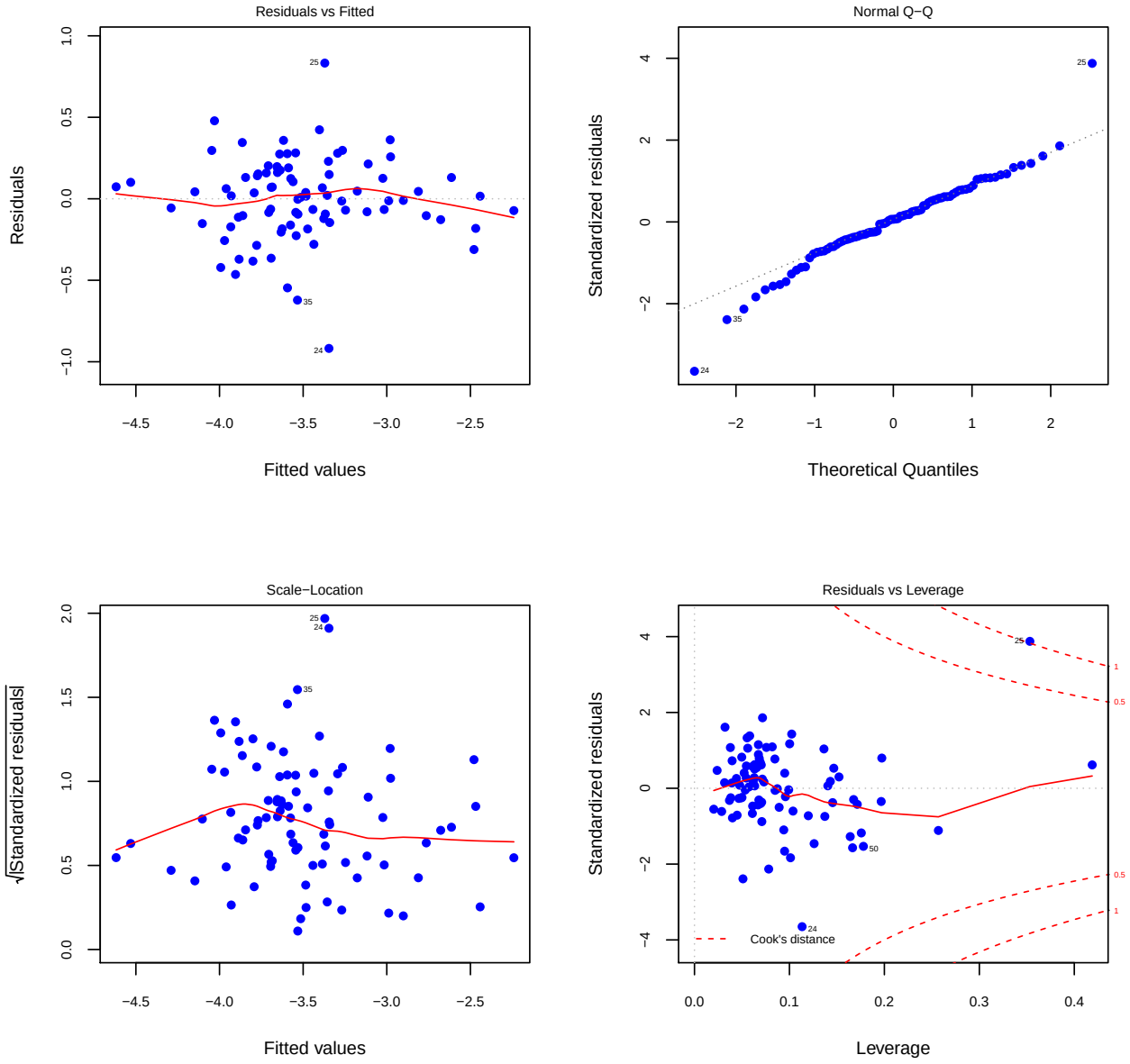

Figure 18: MLR Analysis of Residuals for Model D

Reviewing Figure 18 (Residuals vs. Fitted values), we can see that the assumption of zero conditional mean is reasonably met, and that the errors are thus uncorrelated with the independent variables. As a result of this, we can say that the estimators ($\hat{\beta}$s) we have calculated in Model D are unbiased.

MLR 5: Homoskedascity

Reviewing Figure 18 (Scale-Location), we can see that the variance of the errors *does* vary with the fitted values (and thus with the independent variables). This means that our model exhibits heteroskedasticity. Additionally, we can use the Breusch-Pagan test to see if we reject the null hypothesis that the model is homoskedastic as follows:

```
> bptest(modelD)
```

```
    studentized Breusch-Pagan test

data:  modelD
BP = 25.13, df = 7, p-value = 0.0007195
```

The Breusch-Pagan test has a p-value of 0.0007 which is much less than a significance level of 0.05, thus we reject the null hypothesis that the errors in the model are homoskedastic. This is in agreement with the plot. In order to counteract this, we used covariance matrices in the model that were heteroskedastic-robust, thus our conclusions should not be affected by this.

Because we have accounted for heteroskedasticity and we have unbiased estimators, at this point we can declare that we have the BLUE (Best Linear Unbiased Estimators) for the $\hat{\beta}$s.

MLR 6: Normality

Reviewing Figure 18 (Q-Q plot), we can see that the errors are reasonably normal, as they mostly line up on the 45 degree line. Additionally, we look at the histogram of residuals in Figure 19:

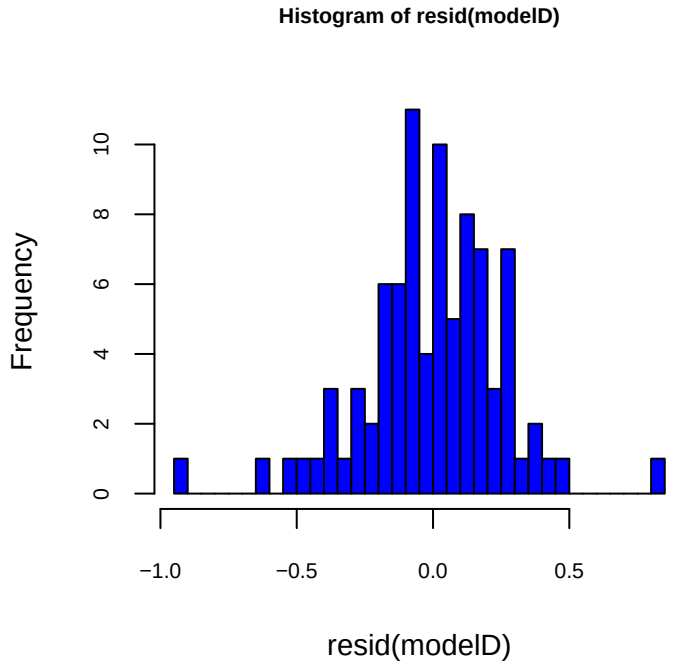

Figure 19: Histogram of Residuals for Model D

From the histogram, we can see that the distribution of errors (residuals) looks mostly normal, possibly slightly negatively skewed. Finally, we will do a Shapiro-Wilk test for normality:

```
> shapiro.test(resid(modelD))

    Shapiro-Wilk normality test

data:  resid(modelD)
W = 0.97097, p-value = 0.04749
```

The null hypothesis in the Shapiro-Wilk test is that the values are normally distributed. As we can see, with a p-value of 0.05, we cannot reject the null at a 5% rejection criterion. Thus, the QQ-plot, histogram and Shapiro-Wilk test together prove that the residuals are normally distributed.

Finally, this means that our ModelD meets all the requirements of the Classical Linear Model Assumptions once Heteroskedascticity is taken into account. We can say that our Final Model estimators are the Best Linear Unbiased Estimators for the model and due to meeting normality of errors, we can use these estimators for causal inference.

Outlier

In figure 18d, the Leverage plot, we see that there is one county with a Cook's distance of 1 - county 25 has both high influence and high leverage which are due to the combination of high tax revenue and high crime rate in that county. Upon review, we do not believe that the data is in error, in fact tax revenue is a variable that often has outliers and we do not feel that this variable should be removed, thus it remains in our analysis.

6.2 Standard Error Analysis

Referring to the standard errors reported in Table 4 for the model we conduct a standard error analysis. By ensuring that our table reports the standard error using the HC covariance matrix, we know that the value in brackets after the coefficient is the correct standard error. Additionally, we know that the *** level next to the coefficient is representative of the p-values for the t-test that we set up in Table 4 (* p<0.05; ** p<0.01; *** p<0.001).

- β_1: Population Density: Coefficient value is 0.347. The standard error is 0.071 (20% of the value) and this coefficient is independently significant at the 0.1% level, or very significant.
- β_2: Tax Revenue per Capita: Coefficient value is 0.273. The standard error is 0.308 (112% of the value) and this coefficient is not independently significant.
- β_3: 'Probability' of Conviction: Coefficient value is -0.289. The standard error is 0.099 (7% of the value) and this coefficient is independently significant at the 1% level, or moderately significant.
- β_4: 'Probability' of Arrest: Coefficient value is -0.395. The standard error is 0.110 (28% of the value) and this coefficient is independently significant at the 0.1% level, or very significant.
- β_5: Service Wage: Coefficient value is -0.207. The standard error is 0.290 (140% of the value) and this coefficient is not independently significant.
- β_6: Percent Minority in 1980: Coefficient value is 0.234. The standard error is 0.043 (18% of the value) and this coefficient is independently significant at the 0.1% level, or very significant.
- β_7: Percent Young Male: Coefficient value is 0.142. The standard error is 0.139 (98% of the value) and this coefficient is not independently significant.

From this we can conclude that a further study, looking more explicitly at excluding tax revenue, service wage and percent young males would be needed. Those variables showed significance in bivariate studies and we will not remove them at this point in the study. Remember that a variable can be independently not significant in this type of analysis within the model, yet the model may be more jointly significant due to the presence of these variables. Additionally, the larger variance (standard error) for the tax revenue variable may partly be due to the outlier. Again, further study is warranted.

7. Recommendations

From the previous analysis, we have drawn several conclusions that will inform our crime reduction policy recommendations for the campaign we have been hired to advise.

Addressing Population Density

Throughout our study, it became clear that population density is a major factor contributing to crime in North Carolina. Consequently, we plan to address population density and our urban centers as one of the hallmarks of our campaign. We plan to invest heavily in urban population centers, through funding for education, community revitalization efforts, poverty reduction, and housing. While these efforts will not redistribute populations or decrease population density, it will ensure that dense urban areas are well supported and are given the same opportunities to develop as the rest of the state.

Increasing Rate of Conviction

Another finding that we will incorporate into our policy recommendation is the strong link between increased rates of conviction and decreased crime. In order to do this we are recommending a development program within the District Attorney's office aimed at limiting arrests only to those cases for which conviction is likely. Further, we plan to implement methods that will make crime reporting easier in order to strengthen cases for conviction.

Closing the Wage Gap

An additional finding that arose from our second model was the correlation between wages and crime rates. In particular, we found that increases in blue collar wages, such as in the service sector are associated with lower crime rate. While on the other hand, an increase in white collar wages, such as federal, state, and local government employees, were more likely to be correlated with an higher crime rate. To begin to close the wage gap between blue and white collar workers, we recommend increasing the minimum wage across the state.

Police Presence in High-Crime Communities

One particularly interesting finding from our study was the positive correlation between police per capita and increased crime rate. In addressing this, we do not recommend divesting in law enforcement in high-crime neighborhoods, but rather acknowledge the probable cause of the relationship, which is that communities are likely already investing in increased police forces in high-crime areas. Therefore, we recommend studying the relationship further before making a recommendation to address police forces in our communities.

Relation of Tax Revenue to Crime Rate

While our model did point to a positive correlation between tax revenue and crime rate, most of our other policy recommendations will not be able to be implemented with decreased funds. Therefore, while we acknowledge the relationship, this is one factor for which we do not recommend taking action and incorporating in our policy platform.

Demographic Relationships to Crime

One of our final discoveries from our Final Model (D) was the positive correlation between both percent minority and percent young males in the population and crime rates. However, there are several factors that prevent us from making policy recommendations specifically

related to these findings. Namely, we cannot change the demographics of our state or what are constituents look like. Further, we cannot craft policies to target certain demographics and certain individuals as this would infringe on constitutional rights and freedoms. For these reasons we plan on focusing our efforts on the aforementioned policy recommendations.

Recommendations for Further Analysis

In a second pass at this analysis, we would first append more recent years of the same data to the existing data set. As discussed previously, a longitudinal view of crime rates over time will provide valuable insights into which changes in policy have been most effective on reducing crime.

We would also like to further investigate the causality of some of our variables, such as police presence and the minority mix of the population in order to better understand their relationships with crime and inform policy.

Finally, the last item we would prioritize is an investigation into public facilities and their relationship to crime rates and crime reporting. While this data may be hard to quantify, it is far removed from the variables we studied in this analysis, so we expect it will have a heavy impact on an updated adjusted R-squared value.

8. Conclusions

We believe that our hallmark policy recommendations, including an investment in urban development in densely populated communities, an increased rate of conviction, and increasing the minimum wage will have a considerable effect on reducing crime in our communities. Further, we have identified several areas for further research in order to focus our efforts and continue the important work of crime reduction even after our original recommendations are implemented. For these reasons, we are confident that our crime reduction policy will help guide this campaign to victory in the upcoming election.