Policy Analysis and Recommendations to Reduce Crime

Craig Fleischman

Introduction

The focus of our study is to generate effective and cost minimizing policy recommendations that will help politicians of North Carolina reduce crime rate. Numerous studies demonstrate that decreasing crime rate results in increased business success and increases probability of reelection due to improved economic conditions. This study will identify efficient ways to decrease crime rate and serves as a starting point for future studies. We will use ordinary least squares regression to examine our dataset to make policy suggestions and will recommend opportunities for additional study.

Research Question: *How can North Carolina affect the crime rate by influencing police per capita and the likelihood of conviction?*

Impact: Being able to answer the question will allow actionable and impactful campaign policies that deter crime in North Carolina.

This question promotes examination and discussion on variables in the dataset that politicians can influence and how they will pay for it. Police per capita is easy to influence by moving police officers to different counties to reach desired output based on the proposed models.

Initial hypothesis: Police per capita and probability of conviction will have statistically significant effects on crime rate. To test this, we start with the following formal hypothesis.

Note, we conduct numerous hypothesis tests throughout our analysis. These tests are all oriented towards analyzing the below hypothesis in depth and our models for robustness.

Null Hypothesis:

1) 'Probability' of conviction (prbconv) has no effect on crime rate.
2) Police per capita (polpc) has no effect on crime rate.

Alternate Hypothesis:

1) 'Probability' of conviction (prbconv) has a statistically significant effect on crime rate.
2) Police per capita (polpc) has a statistically significant effect on crime rate.

Classical Hypothesis: (The following hypotheses will be tested for each parameter)

$$H_0^{(1)} : \frac{dcrmrte}{dprbconv} = 0 \Rightarrow \beta_{prbconv} \Rightarrow 0$$
$$H_A^{(1)} : \beta_{prbconv} \neq 0$$

$$H_0^{(2)} : \frac{dcrmrte}{dpolpc} = 0 \Rightarrow \beta_{polpc} \Rightarrow 0$$
$$H_A^{(2)} : \beta_{polpc} \neq 0$$

Table of Contents

1.0 Exploratory Data Analysis

1.1 Data Background

The data was first used by researchers from the University of Georgia and West Virginia University (C. Cornwell and W. Trumball (1994), “Estimating the Economic Model of Crime with Panel Data,” Review of Economics and Statistics 76, 360-366.) The original study was based on a multi-year panel in North Carolina for 1987. In addition, our data comes from four other sources: census data, the FBI's Uniform Crime Report, the North Carolina Department of Correction, and the North Carolina Employment Security Commission. All of the variables were from 1987 except for two variables which were from the 1980 census; (1) percent young male between the ages of 18 to 24 and (2) percent minority. The data is grouped by counties and regions, but does not include all counties with many variables containing ratios and averages versus raw data. We use a single year from a multi-year study that went from 1980-1987. Refer to the Additional Resources section in the Appendix for more details.

The exploration of the data revealed many notable items that needed to be resolved. The raw data contained 25 variables and 97 rows. Below are the discoveries and cleanup:

Summary of Fixes Made

Item	Fix	Description
Data Set	NA Values	Remove the 6 rows of NA's
County 193	Duplicate Row	Removed duplicate Row
prbconv	Set to numeric	prbconv was converted from char to numeric
Year	Set to Null	No variation in variable, all listed as 1987
County	Set to Null	Counties were not used in this analysis and were incomplete. The data set included 90 counties out of the 100 counties in North Carolina
Density Variable (density)	Divided by 100	Verified with the North Carolina State Records from the 1980s (https://files.nc.gov/ncosbm/demog/dens7095.html)
Percent Minority Variable (pctmin80)	Converted from percentage to decimal	The variable pctmin80, percent minority from 1980, was entered as a percentage (ex. 25.46) while pctymle, percent male between ages 15 and 24, was entered as decimal (ex. .2546). Since other variables were in decimal form and to be consistent with pctymle, pctmin80 was converted to decimal.
Service Industry Wages (wser) Warren County 185	Reduced value by one order of magnitude (2177.06 to 217.71)	Made conservative adjustment wser due to influential outlier based on census and legal data concerning Warren County (county 185) and data format. County is low income with low population, previously high value did not make sense and was not supported by additional reliable data sources. https://www.energy.gov/lm/services/environmental-justice/environmental-justice-history, https://www.bls.gov/news.release/history/annpay_091693.txt

(See Section 0.2 in the Appendix for Region & Urban details.)

Once the data was cleaned, there were 23 variables and 90 rows in the dataset. Each of the variables were analyzed for oddities and outliers that would require further investigation.

1.2 Variable Oddities

Region & Urban Variables: The Region and Urban variables could be helpful to pinpoint where in North Carolina would policies have the larger effect based on specific policies targeted at that area for reducing crime. There are 22 'west' regions, 34 'central' regions, and the regions not west or central were labeled 'other'. Urban regions as defined by the Standard Metropolitan Statistical Area (SMSA) included 1 urban area in the 'west' region and 5 in the 'central' region, and 2 in the 'other' region for a total of 8 urban areas. (See Section 0.1: Data Load, Analysis, and Cleaning.)

Probability Variables (prbarr, probconv, prbpris, & mix) These variables were presented as probabilities and 3 variables were labeled as: 'probability of arrest', 'probability of conviction', and 'probability of prison sentence'. A probability must between 0 and 1, but prbarr had a max of 1.091 and prbconv had a max of 2.121. See table below.

Probability Variable Summary

Variables	Max	Description	Definition
prbarr	1.091	probability of arrest	ratio of arrests to offenses
prbconv	2.121	probability of conviction	ratio of convictions to arrests
prbpris	0.600	probability of prison sentence	ratio of prison sentence to total convictions
mix	0.465	offense mix: face-to-face/other	n/a

Given that these variables are ratios it is possible for the number to be > 1, where arrests could be greater than offenses and convictions greater than arrests. This can be explained by yearly rollovers where an action occurs in one year and the complement happens in the next year. Also, the greater the backlog of processing also increases the likelihood of numbers that exceed 1. It is plausible that there were more arrests than convictions and more convictions than arrests in a given year, so we did not modify this data. The residual vs leverage plots also confirm that these outliers are not considerably influential to be removed from our data set. The regression lines also confirm this as well. (See Reference Section 0.4)

The 'probability' variables were not altered.

The 'probability' variables were not altered.

1.3 Outlier Examination

The outliers were identified by examining the mean, the standard deviation, the maximum and minimum values as well as examining boxplots and Cook's Distance plots. The variables that were of interest or had an extreme potential outlier (polpc and wser) were examined closer. (Reference details: Section 0.1 - 0.6)

- Police per Capita (polpc):
  The variable for police per capita (polpc) has a notable outlier (.009054) in the west region, specifically county 115. The interpretation is that there are 90.54 police for every 10,000 residents. It is plausible that there are 90.54 police officers in that county, so we did not modify this observation. In 2016, Atlantic City, New Jersey had 70.9 officers for every 10,000 residents (Source: FBI's Uniform Crime reporting (UCR) data.) In addition, county 115 had similar potential extremes for prbarr (1.09091) and close to the maximum for prbconv (1.5). The residual vs leverage plots confirm that these outliers are considerably influential and have high leverage (See reference Section 0.5)
- Service Industry Wages (wser):
  In the weekly wages variable for the service industry there is a county that has an exceptionally higher value ($2,177.068) relative to other data points as highlighted in the boxplot. This variable was 10x larger than the average. Since this number is an average data point instead of an individual data point the Cook's number is > 3, this could be an input error. After more research into county 185, Warren County, this observation was deemed an input error. Warren County is a low income and low population county. If the number was divided by 10 (2,177.068 -> 217.71), the new value looks to be consistent with the other county wser observations. This was confirmed using additional resources found at https://www.energy.gov/lm/services/environmental-justice/environmental-justice-history (https://www.energy.gov/lm/services/environmental-justice/environmental-justice-history) and https://www.bls.gov/news.release/history/annpay_091693.txt (https://www.bls.gov/news.release/history/annpay_091693.txt). Refer to the Additional Resources Section of the Appendix for more information.

In general, there is not a good argument to adjust any other observations without a more thorough understanding of the data. Removing the outliers to improve the significance is closer to data-mining; therefore, the additional outliers that were identified were included in the analysis.

More research should be conducted on policy implications of these outlier variables. However, census data from 1980 supported the demographic and density information included https://www2.census.gov/library/publications/decennial/1990/cp-1/cp-1-35.pdf (https://www2.census.gov/library/publications/decennial/1990/cp-1/cp-1-35.pdf). Refer to the Additional Resources Section of the Appendix for more information.

1.4 Variable Correlation

The correlations analysis (reference Section 0.11: Correlation Analysis in the Appendix) identified important variables to examine due to their correlation to the dependent variable, crmrte. In the table below, the variables with a coefficient strength ≥abs(.34) were as follows: prbarr, prbcon, density, taxpc, west (binary variable), urban (binary variable), wcon, wtrd, wfir, wmfg, wfed, and wloc. The table also indicates the direction (negative = red or positive = blue).

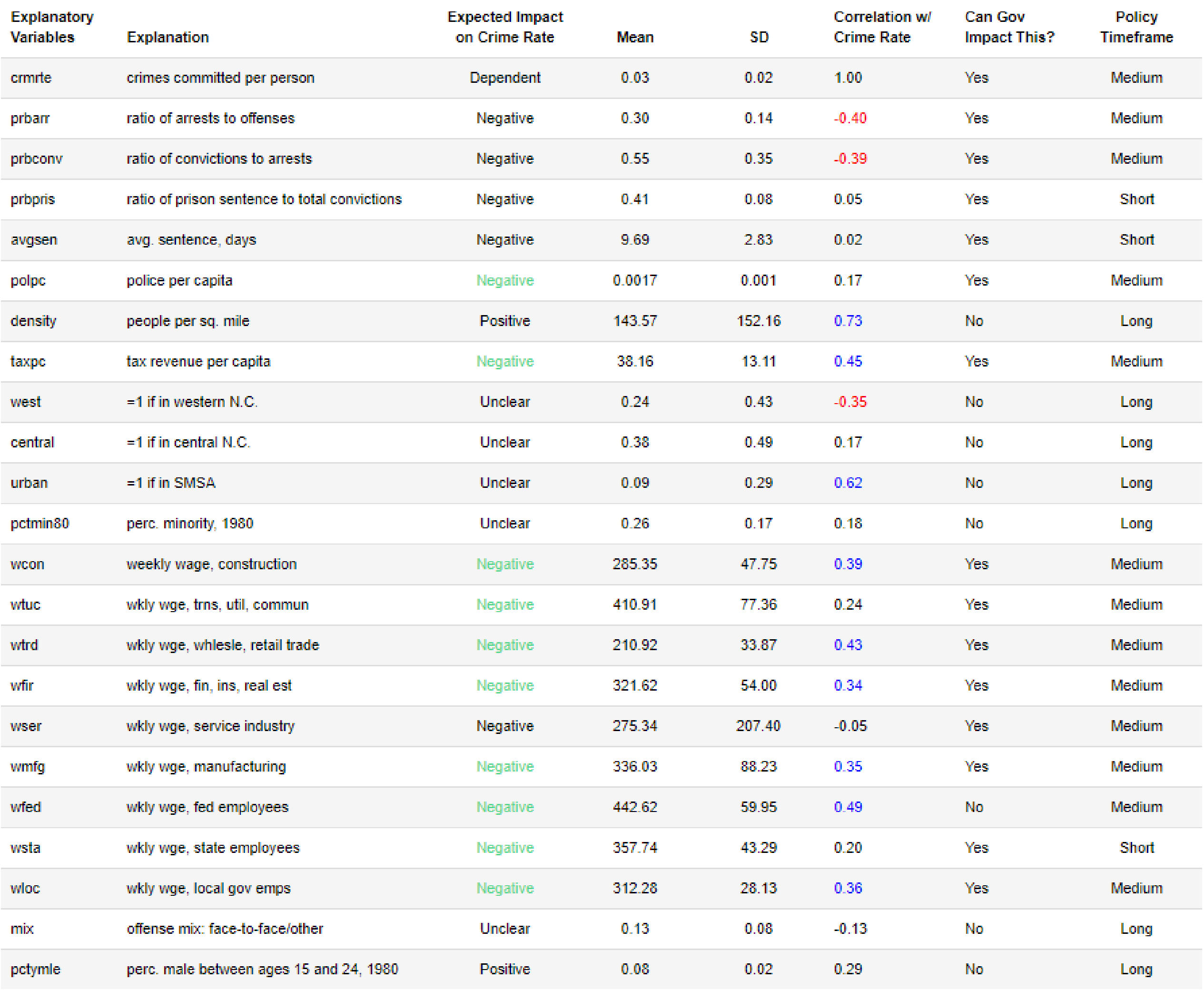
The expected impact on crime rate for each variable is identified in the table (Negative or Positive), it is color coded in green to indicate a mismatch with the variable coefficient direction and relative strength. Interest results would include: both police per capita and tax revenue per capita have a positive correlation. As the number of police officers increase the crime rate increases, this is true for tax revenue as well. This is the opposite of what was expected. For additional research it would be ideal to examine the variable 'mix', which is face-to-face crime, as well as omitted variable bias, such as providing more details on the type of crime.

Another interesting result was that wages were positively correlated with crime rate. As wages increased so did crime. Additional analysis of potential reasons can be found in the simple model and the extended model. More information would have to be obtained from the data set to dive deeper.

[https://www.continentaltelegraph.com/economy/increasing-the-minimum-wage-increases-crime-obviously-enough/] (https://www.continentaltelegraph.com/economy/increasing-the-minimum-wage-increases-crime-obviously-enough/])

There were two groups of variables which are significantly correlated with each other and that could be important covariates to examine. They are: density:urban and pctmin80:west. These interactions will be discussed in the Optimization Model.

Policy Levers That Could Impact Crime Rate

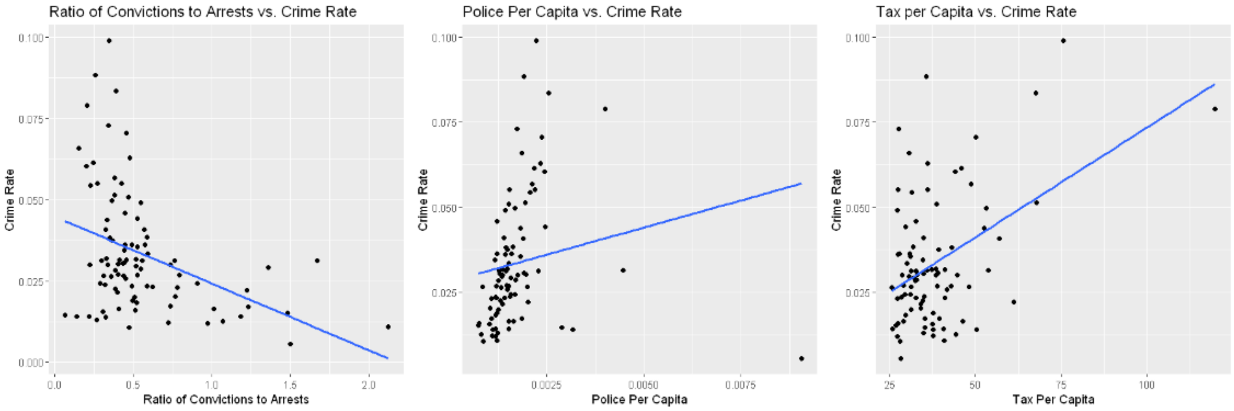


Explanatory Variables	Explanation	Expected Impact on Crime Rate	Mean	SD	Correlation w/ Crime Rate	Can Gov Impact This?	Policy Timeframe
crmrte	crimes committed per person	Dependent	0.03	0.02	1.00	Yes	Medium
prbarr	ratio of arrests to offenses	Negative	0.30	0.14	-0.40	Yes	Medium
prbconv	ratio of convictions to arrests	Negative	0.55	0.35	-0.39	Yes	Medium
prbpris	ratio of prison sentence to total convictions	Negative	0.41	0.08	0.05	Yes	Short
avgsen	avg. sentence, days	Negative	9.69	2.83	0.02	Yes	Short
polpc	police per capita	Negative	0.0017	0.001	0.17	Yes	Medium
density	people per sq. mile	Positive	143.57	152.16	0.73	No	Long
taxpc	tax revenue per capita	Negative	38.16	13.11	0.45	Yes	Medium
west	=1 if in western N.C.	Unclear	0.24	0.43	-0.35	No	Long
central	=1 if in central N.C.	Unclear	0.38	0.49	0.17	No	Long
urban	=1 if in SMSA	Unclear	0.09	0.29	0.62	No	Long
pctmin80	perc. minority, 1980	Unclear	0.26	0.17	0.18	No	Long
wcon	weekly wage, construction	Negative	285.35	47.75	0.39	Yes	Medium
wtuc	wkly wge, trns, util, commun	Negative	410.91	77.36	0.24	Yes	Medium
wtrd	wkly wge, whlesle, retail trade	Negative	210.92	33.87	0.43	Yes	Medium
wfir	wkly wge, fin, ins, real est	Negative	321.62	54.00	0.34	Yes	Medium
wser	wkly wge, service industry	Negative	275.34	207.40	-0.05	Yes	Medium
wmfg	wkly wge, manufacturing	Negative	336.03	88.23	0.35	Yes	Medium
wfed	wkly wge, fed employees	Negative	442.62	59.95	0.49	No	Medium
wsta	wkly wge, state employees	Negative	357.74	43.29	0.20	Yes	Short
wloc	wkly wge, local gov emps	Negative	312.28	28.13	0.36	Yes	Medium
mix	offense mix: face-to-face/other	Unclear	0.13	0.08	-0.13	No	Long
pctymle	perc. male between ages 15 and 24, 1980	Positive	0.08	0.02	0.29	No	Long

The table outlines all of the variables post data cleaning and their correlation to the Explanatory Variable (crmrte). This includes the expected impact the variable will have crime, the correlation to the crime rate, how effectively can government impact this variable, and if so what would the time frame be for a policy directed at this variable to have an impact. Understanding the impact and time frame will provide additional feedback on which variable or combination of variables to focus in on. This table also functions as a key to identify each variable throughout the study.

Key Variable Correlation with Crime Rate Plots

Three key parameters; taxpc, polpc, and prbconv were identified after completing our EDA and were incorporated into the research question.



Note that tax per capita correlates postively with crime rate, as does police per capita. This is interesting, and we look forward to seeing whether increasing taxes and increasing police may increase crime rate. However, the ratio of convictions to arrests has a negative correlation with crime rate.

1.5 Data Transformation

To determine which variables to examine for potential transformations the below principles were followed:

- **Variable understandable post transform** - Variables must be understandable and meaningful after a transformation so the variable interpretations can be clear and actionable.
- **Reducing skewness** - A transformation may be used to reduce skewness. A distribution that is symmetrically distributed around the mean or nearly so is often easier to incorporate and interpret.
- **Reducing systematic change in the spread** - This assists with achieving approximate homoscedasticity (reducing "heteroscedasticity").
- **Linear relationships** - When looking at relationships between variables, it is often far easier to think about patterns that are approximately linear than patterns that are highly curved. [Source: http://fmwww.bc.edu/repec/bocode/t/transint.html] (http://fmwww.bc.edu/repec/bocode/t/transint.html])

2.0 Main Analysis

2.1 Key Parameters

The variables chosen to be analyzed in the base model were the following:

- $\log(crmrte)$ - The log of 'crmrte' was chosen because the distribution is approximately symmetric (skewness = -.1017) and histogram verifies this as well compared to 'crmrte' which was highly skewed at 1.2817.
- $polpc$ - Police per capita could be directly influenced by government policy; hence, it is informative to establish whether there is statistical justification for changing existing policing level in North Carolina as a vehicle for reducing crime rate.
- $prbconv$ -There is a negative correlation between crime rate and likelihood of conviction. The likelihood of crime conviction (proxied by 'probability of conviction' in the dataset) can be influenced by government policies and could be an effective crime deterrent.
- $taxpc$ - There is a fairly strong linear positive correlation between crime rate and tax per capita. Although fiscal initiatives can be directly influenced by government policies, the impact of tax policies on crime is uncertain.

2.2 Simple Regression Model

The simple regression model (Model 1) explores the relationship between crime rate as a dependent variable on tax per capita, likelihood of conviction and police per capita. Since these variables can be directly influenced by government policy, our model results provide a technical basis for a political campaign focused on crime rate reduction in North Carolina. These variables were chosen from a political and policy standpoint. Reduced crime rate makes North Carolina appealing to residents and businesses; dependent variables of fiscal nature and deterrents of crime are more easily influenced by government policies and thus are highly important as input for political campaign policies.

Log of crime rate is dependent on police per capita, likelihood of conviction and tax per capita.

$$\widehat{logcrmrte} = \hat{\beta}_0 + \hat{\beta}_1 polpc + \hat{\beta}_2 prbconv + \hat{\beta}_3 taxpc + \hat{u}$$

The following hypotheses will be tested for each parameter:

- $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$
- $H_0 : \beta_2 = 0$ and $H_A : \beta_2 \neq 0$
- $H_0 : \beta_3 = 0$ and $H_A : \beta_3 \neq 0$

Simple Regression Model Interpretation

Model 1 suggests that policing, tax per capita and likelihood of crime conviction accounts for about 30% in crime rate variation. Likelihood of conviction and tax per capita are statistically significant while police per capita is not statistically significant. Police per capita has a homoskedastic standard error one order of magnitude greater than the value of the coefficient, reducing its practical significance in this model.

$$\frac{\partial crmrte}{\partial polpc} \Rightarrow -3.449$$

Model1 suggests that we should reject the $H_0^{(1)}$.

$$\frac{\partial crmrte}{\partial prbconv(***)} \Rightarrow -0.630$$

Model1 suggests that we should reject $H_0^{(2)}$.

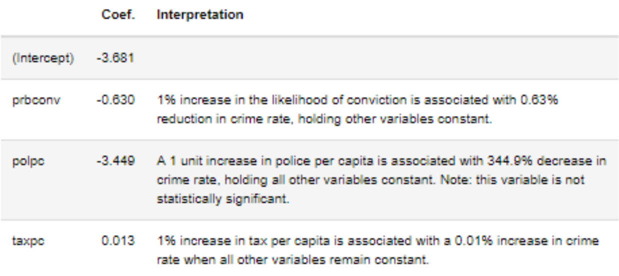
$$\frac{\partial log(crmrte)}{\partial taxpc(**)} \Rightarrow 0.013$$

Model1 suggests that we should reject $H_0^{(3)}$.

$$\widehat{logcrmrte} = -3.681 - 0.630(prbconv) - 3.449(polpc) + 0.013(taxpc)$$

$$n = 90 \qquad Adj. R^2 = 26.7$$

Simple Regression Model Variable Summary



	Coef.	Interpretation
(Intercept)	-3.681	
prbconv	-0.630	1% increase in the likelihood of conviction is associated with 0.63% reduction in crime rate, holding other variables constant.
polpc	-3.449	A 1 unit increase in police per capita is associated with 344.9% decrease in crime rate, holding all other variables constant. Note: this variable is not statistically significant.
taxpc	0.013	1% increase in tax per capita is associated with a 0.01% increase in crime rate when all other variables remain constant.

Simple Regression Model Summary of coefficients table

2.3 Simple Regression Model Assumptions

Our Base Model is a linear model based on Ordinary Least Squares Regression. Due to the potential of non random sampling in our data, this model is not necessarily the best linear unbiased estimator. The density of the 10 missing counties are lower than the average (1990 Census Data); we do not know of any other defining characteristics. Additionally, this model is homoskedastic and meets the zero conditional mean assumption with no perfect multicollinearity. See Appendix for more details.

2.4 Optimization Model

The Optimization Model (Model 2) will continue to build off of our base model using the dependent variable of $\log(crmrte)$ and the independent (or explanatory) variables. The independent variables chosen were based on their uniform distribution, their homoscedasticity, and variables that would be understood and actionable to develop policy to reduce crime.

The below variables were considered for analysis in the preliminary Optimization model:

- $\log(crmrte)$ - See simple regression model.
- $wklywage$ - A newly created variable, 'wage' was used as the average wage for all 9 sectors. Without detailed weighting information or an understanding of the population for each county or more details below county, it was difficult to use each of the 9 mean wages sectors separately. In addition, the implementation and understanding of using each of the 9 variables might reduce the chances of implementation and reduce broad support across constituents. Therefore, the mean for all the wage sector variables were used. The histogram of the new wage variable was mostly normally distributed with a low skewness score of 1.0 and with a high coefficient with $\log(crmrte)$ of 0.46.
- $density$ - Density has the highest correlation with $\log(crmrte)$ at 0.63.
- $\log(pctymle)$ - The log of percentage of males between ages 15 and 24 (as defined in 1980) in variable 'pctymle' had a moderate correlation at 0.31 and the transformation reduced the skewness by almost twice from 4.56 to 2.89. The histogram visual verification also demonstrates an improved distribution.
- $\log(polpc)$ - The log of police per capita, 'polpc' had a 0.28 correlation with $\log(crmrte)$ vs 0.01 with 'crmrte'. In addition the skewness reduced from 4.9834 (polpc) to 1.4083 $\log(polpc)$. The histogram visual verification also demonstrates an improved distribution.
- $prbarr$ - The ratio of arrests to offenses (prbarr) was used due to a very strong correlation with $\log(crmrte)$ of -0.47. The log of 'prbarr had a correlation of -0.44 and was more symmetrical, but the variable without the transformation made it easier to understand and thus was chosen.
- $prbconv$ - The ratio of convictions to arrests (prbconv) was used do to a strong negative correlation with $\log(crmrte)$ of -0.47. Similar to 'prbarr' the $\log(prbconv)$ was more symmetrical, but the ease of understanding the variable without the transformation made it easier to understand and thus was chosen.
- $pctmin80$ - The percent minority (1980) has a positive correlation of 0.23 and the skewness is -0.4422. Both the log and cube root of the variable has a worse skewness score, so the variable was not transformed.
- $taxpc$ - The tax revenue per capita variable had a positive coefficient of 0.36 and had a skewness score that was twice the size of $\log(taxpc)$. This variable was left as is since it was also easier to explain. When both variables (transformed and non) were run in a model, neither of them were significant; thus, were left in the model.

All the variables that were linear or mostly linear, individually statistically significant, and helped to answer the research question were placed into the Optimization model using robust standard errors.

Interactions

Four interaction variables were explored based on their high correlation values with themselves (pctmin80:west, density:urban, density:logpctymle) or of their interest to our main research question (prbarr:prbconv). When adding these interaction variables to the model, none of them were statistically significant. (See Reference Section 2.135: Model2 (original, step-wise, interaction))

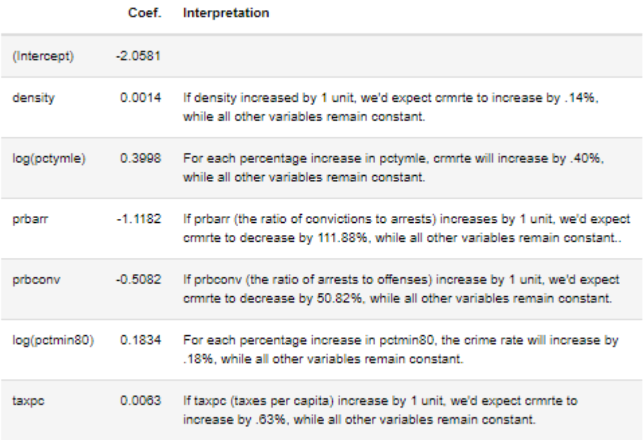
2.5 Optimization Model Summary & Formula

After combining the 6 variables and building a model, two of the variables become not statistically significant (prbarr and taxpc), while the rest of the variables remained statistically significant and used in the model below. From a practical standpoint prbarr, prbconv, and taxpc are the most actionable and practical in the formula. Density, percentage of males between 18 to 24, and percentage of minority are less practical, but still could be interesting to include for targeted programs aimed at those specific groups.

$$\widehat{logcrmrte} = -2.0581 + 0.0014(density) + 0.3998(logpctymle) - 1.1188(prbarr) - 0.5082(prbconv) \\ +0.1834(logpctmin80) + 0.0063(taxpc)$$

$$n = 90 \qquad Adj. R^2 = 72.9$$

Optimization Model Variable Summary



	Coef.	Interpretation
(Intercept)	-2.0581	
density	0.0014	If density increased by 1 unit, we'd expect crmrte to increase by .14%, while all other variables remain constant.
log(pctymle)	0.3998	For each percentage increase in pctymle, crmrte will increase by .40%, while all other variables remain constant.
prbarr	-1.1182	If prbarr (the ratio of convictions to arrests) increases by 1 unit, we'd expect crmrte to decrease by 111.88%, while all other variables remain constant..
prbconv	-0.5082	If prbconv (the ratio of arrests to offenses) increase by 1 unit, we'd expect crmrte to decrease by 50.82%, while all other variables remain constant.
log(pctmin80)	0.1834	For each percentage increase in pctmin80, the crime rate will increase by .18%, while all other variables remain constant.
taxpc	0.0063	If taxpc (taxes per capita) increase by 1 unit, we'd expect crmrte to increase by .63%, while all other variables remain constant.

Optimization Summary of coefficients table

Optimization Model Stepwise Analysis

Using the step-wise command the highest correlated variables mentioned above plus variables prbpris, avgsen, and polpc will be used to compare and contrast the differences between the optimization model and the step-wise model.

Optimization Model STEP formula

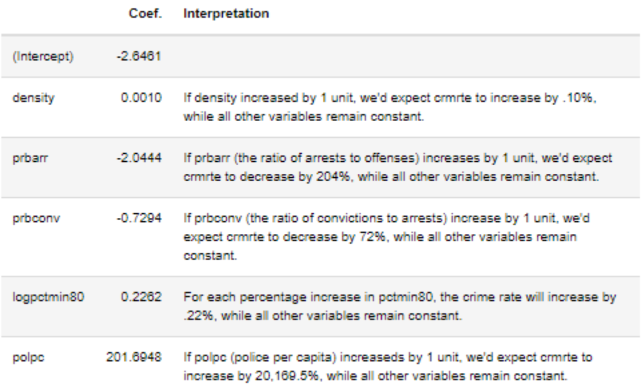
$$\widehat{logcrmrte} = -2.6461 + 0.00103(density) - 2.0444(prbarr) - 0.7294(prbconv) \\ +0.2262(logpctmin80) + 201.695(polpc)$$

$$n = 90 \qquad Adj. R^2 = 78.2$$

Items of Interest:

1) All of the variables are significant, with $polpc$ having the least statistically significance at p < 0.01, while the rest of the variables are at p < .001
2) The $polpc$ variables, or police per capita, was the new variable that was added that was not in the original optimization model. Individually this variable was not statistically significant and not linear. As noted earlier, there was an extreme potential outlier present.
3) The $logpctymle$ and $taxpc$ variables were removed and not in the original optimization model. $taxpc$ had little practical significance with a coefficient of .0069.

2.6 Optimization Model Stepwise Summary



	Coef.	Interpretation
(Intercept)	-2.6461	
density	0.0010	If density increased by 1 unit, we'd expect crmrte to increase by .10%, while all other variables remain constant.
prbarr	-2.0444	If prbarr (the ratio of arrests to offenses) increases by 1 unit, we'd expect crmrte to decrease by 204%, while all other variables remain constant.
prbconv	-0.7294	If prbconv (the ratio of convictions to arrests) increase by 1 unit, we'd expect crmrte to decrease by 72%, while all other variables remain constant.
logpctmin80	0.2262	For each percentage increase in pctmin80, the crime rate will increase by .22%, while all other variables remain constant.
polpc	201.6948	If polpc (police per capita) increaseds by 1 unit, we'd expect crmrte to increase by 20,169.5%, while all other variables remain constant.

Optimization Step-wise Summary of coefficients table

All of the assumptions were achieved accept for random sampling, which is true for all models. This model is not necessarily the best linear unbiased estimator.

CLM 1.0: Linearity of Parameters:
The model is linear in parameters: log of crime rate is a linear function. As observed on the diagnostic regression plot of residuals against fitted values, a random pattern indicates an appropriate model has been fit to the data.

CLM 2.0: Random Sampling and Independence:
The data does not appear to be random. 10 counties are missing from the dataset with no explanation. Those counties notably include some of the most sparsely populated counties, each having a population of less that 68,000 people and are not densely populated. The average population per county in North Carolina is just over 105,068. Additionally, our data is from two different databases. One is the North Carolina Department of Correction and the other is the FBI's Uniform Crime Reports. Each of these purports that it is comprehensive, but we don't necessarily have enough information to be sure that combining these datasets creates a representative sample, or if it biases the data towards crimes more likely to be investigated by the FBI to include many felonies, kidnappings, and cross-state line offenses that wind up or start in North Carolina. It is reasonable to assume independence.

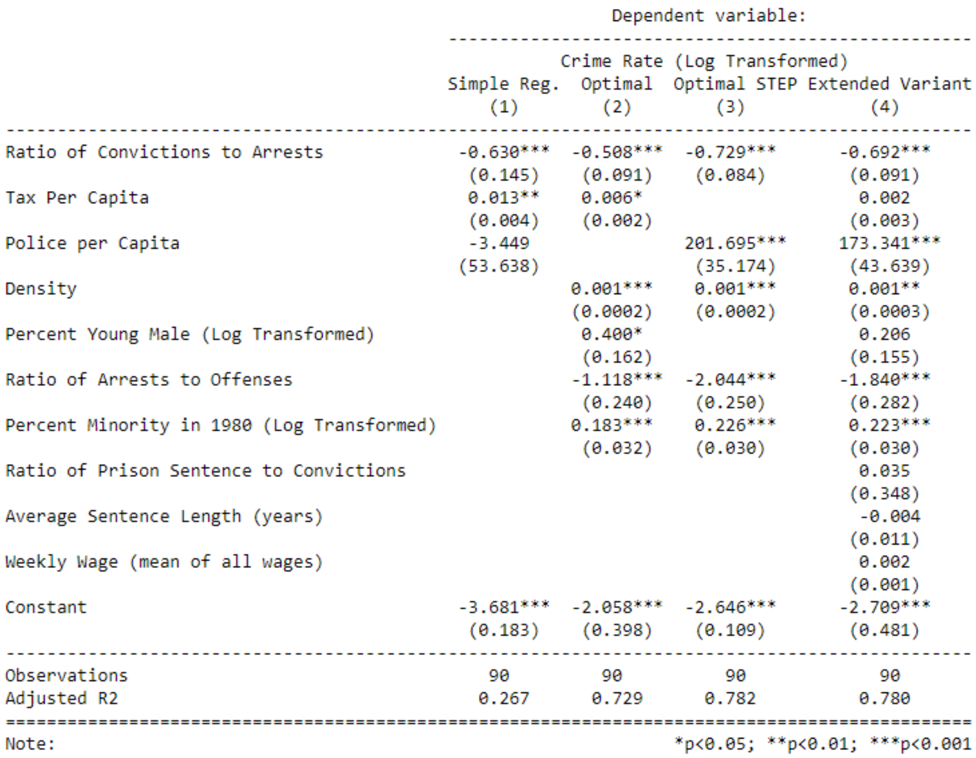
CLM 3.0 Zero Conditional Mean:
The expected value of residuals is approximately zero. (Reference Section 2.14)

CLM 4.0': Normality of residuals:
The qqPlot suggests slight deviation from normality of residuals. This was addressed by investigation of outliers and their leverage on the linear model.

CLM 5.0 Homoscedasticity:
The scale-location plot of the Linear Model diagnostic plots graphically evaluates the assumption that errors are identically- distributed and have homogeneous variance. The plot shows the square root of the absolute residuals against the fitted values, along with a smooth red line. Significant departures from a horizontal line typically suggest heteroscedasticity. The Model plot shows approximately horizontal line which implies homogeneous variance. Quantitatively, assumption of constant error variance can be tested using the ncvTest function. With a p-value of (Optimization model was 0.79 and step-wise model was 0.88), we fail to reject the null hypothesis (that variance of residuals is constant) and therefore infer that the residuals are homoscedastic. (Reference Section 2.14)

CLM 6.0 Multicollinearity:
The Variance Inflation Factor (VIF) can detect and represent collinearity. If any of the variance inflation factors exceed 5 (the cut-off typically used), the regression coefficients are poorly estimated due to multicollinearity. All of the variables in the model are below 1.4. This suggests that the Multicollinearity assumption is not violated. (Reference Section 2.14)

2.7 Main Regression Table



```
                                                     Dependent variable:
                                          ---------------------------------------------------
                                                 Crime Rate (Log Transformed)
                                          Simple Reg.  Optimal  Optimal STEP Extended Variant
                                              (1)        (2)        (3)            (4)
-----------------------------------------------------------------------------------------------
Ratio of Convictions to Arrests            -0.630***  -0.508***  -0.729***      -0.692***
                                            (0.145)    (0.091)    (0.084)        (0.091)
Tax Per Capita                              0.013**    0.006*                     0.002
                                            (0.004)    (0.002)                   (0.003)
Police per Capita                           -3.449                201.695***     173.341***
                                           (53.638)               (35.174)       (43.639)
Density                                                0.001***   0.001***       0.001**
                                                      (0.0002)   (0.0002)       (0.0003)
Percent Young Male (Log Transformed)                   0.400*                     0.206
                                                       (0.162)                   (0.155)
Ratio of Arrests to Offenses                          -1.118***  -2.044***      -1.840***
                                                       (0.240)    (0.250)        (0.282)
Percent Minority in 1980 (Log Transformed)             0.183***   0.226***       0.223***
                                                       (0.032)    (0.030)        (0.030)
Ratio of Prison Sentence to Convictions                                           0.035
                                                                                 (0.348)
Average Sentence Length (years)                                                  -0.004
                                                                                 (0.011)
Weekly Wage (mean of all wages)                                                   0.002
                                                                                 (0.001)
Constant                                   -3.681***  -2.058***  -2.646***      -2.709***
                                            (0.183)    (0.398)    (0.109)        (0.481)
-----------------------------------------------------------------------------------------------
Observations                                  90         90         90             90
Adjusted R2                                  0.267      0.729      0.782          0.780
===============================================================================================
Note:                                                       *p<0.05; **p<0.01; ***p<0.001
```

This regression table includes the base model (1), the optimal model (2), the step model (3), and a variation on the extended model (4). Note the inclusion of three additional variables on model (3) only increases our explanation of variance in crime rate by about 5% compared to the optimal model and reduces explanation of variance compared to the step model. We sacrifice parsimony for marginal additional explanation. We also know due to the joint statistical testing that our wages (which are incorporated into the mean of all wages variable), the average sentence length, and the ratio of prison sentences to convictions are not statistically or practically significant due to the standard error and their lack of influence on the model. We also note that the ratio of convictions to arrests is inversely proportional to crime rate when we control for tax per capita, police per capita, density, percent young male, ratio of arrests to offenses, wages, sentence length, ratio of prison sentence to convictions and percent minorities in 1980. Additionally, police per capita and tax per capita have a direct relationship, so increases in both variables correspond to an increase in crime rate when controlling for the variables discussed above. An increase in the ratio of arrests to convictions also results in a decrease in crime rate. For more detailed discussion, to include unit analysis, interactions, and indicators, refer to each model section of the report.

2.8 Hypothesis Evaluation

Initial hypothesis: Police per capita and probability of conviction will have statistically significant effects on crime rate. To test this, we start with the following formal hypothesis.

Note, we conduct numerous hypothesis tests throughout our analysis. These tests are all oriented towards analyzing the below hypothesis in depth and our models for robustness.

Null Hypothesis:

1) 'Probability' of conviction (prbconv) has no effect on crime rate.
2) Police per capita (polpc) has no effect on crime rate.

Alternate Hypothesis:

1) 'Probability' of conviction (prbconv) has a statistically significant effect on crime rate.
2) Police per capita (polpc) has a statistically significant effect on crime rate.

Classical Hypothesis: (The following hypotheses will be tested for each parameter)

$$H_0^{(1)} : \frac{dcrmrte}{dprbconv} = 0 \Rightarrow \beta_{prbconv} \Rightarrow 0$$
$$H_A^{(1)} : \beta_{prbconv} \neq 0$$

$$H_0^{(2)} : \frac{dcrmrte}{dpolpc} = 0 \Rightarrow \beta_{polpc} \Rightarrow 0$$
$$H_A^{(2)} : \beta_{polpc} \neq 0$$

With p<.05, we rejected our null hypothesis for prbconv in all three models (simple regression, optimization and step, extended). With a p<.05, we rejected our null hypothesis for polpc in the optimization STEP and extended models.

In every model, our ratio of convictions to arrests (prbconv) was inversely proportional to crime rate. In each model, crime rate is reduced by about 50% to 70% with a one unit increase in prbconv.

Police per capita was also inversely proportional to crime rate for the simple regression model. However, this relationship was of low practical significance, since the standard error for this term was one order of magnitude larger than the term itself and polpc was not statistically significant in this model. When we controlled for additional variables, polpc had a direct relationship with crime rate.

The optimization STEP model provides the best balance between parsimony and explanation of key parameters allowing us to control for other factors that influence crime rate.

3.0 Extended Model

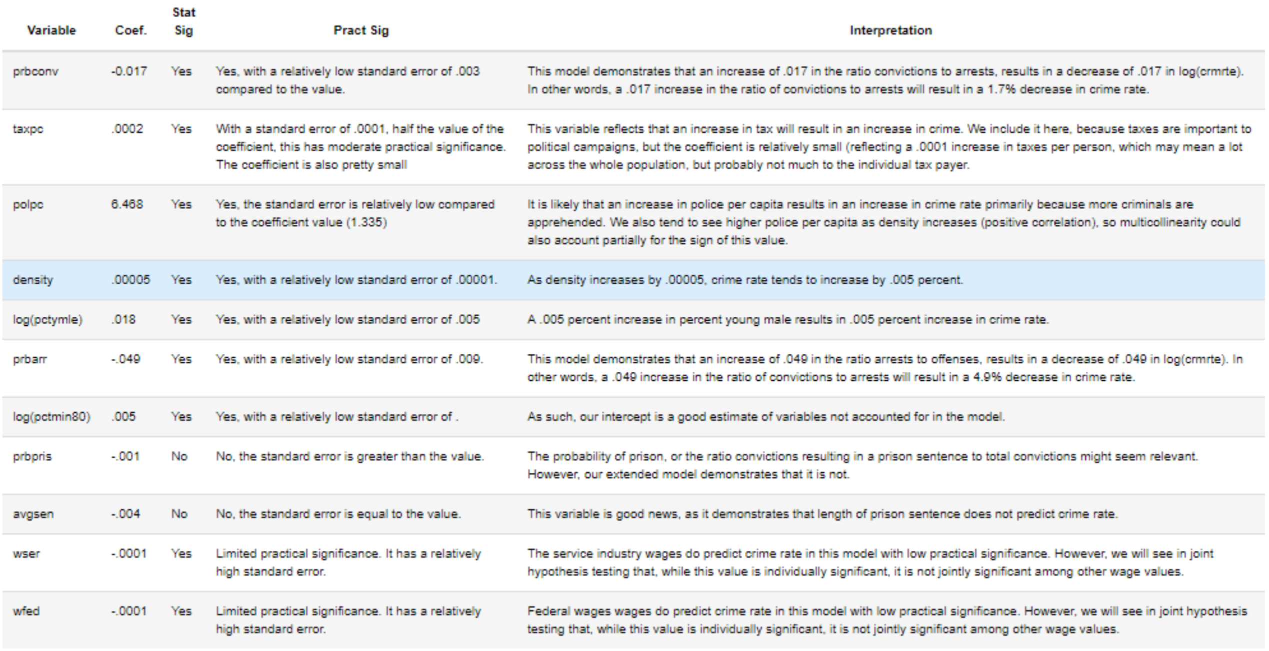
3.1 Extended Model Formula

$$\begin{aligned}\log(crmrte) &= .082 - 0.017prbconv + .0002taxpc + 6.468polpc + .00005density + .018log(pctymle)\\ &-.049prbarr + 0.005log(pctmin80) - .001prbpris - .004avgsen + .00002wcon + .00001wtuc\\ &+.00002wtrd - .00002wfir - .0001wfed - .00003wsta + .00004wloc\end{aligned}$$

$$n = 90 \qquad Adj.R^2 = 81.3$$

81.3% of the variance of Crime Rate is explained by our model. However, this model sacrifices parsimony for accuracy. Our Joint Statistical Testing discussed in Section 3.2 Extended Model Interpretation demonstrates the relatively low practical value of including a number of variables in this model.

Extended Model Variable Summary



Variable	Coef.	Stat Sig	Pract Sig	Interpretation
prbconv	-0.017	Yes	Yes, with a relatively low standard error of .003 compared to the value.	This model demonstrates that an increase of .017 in the ratio convictions to arrests, results in a decrease of .017 in log(crmrte). In other words, a .017 increase in the ratio of convictions to arrests will result in a 1.7% decrease in crime rate.
taxpc	.0002	Yes	With a standard error of .0001, half the value of the coefficient, this has moderate practical significance. The coefficient is also pretty small	This variable reflects that an increase in tax will result in an increase in crime. We include it here, because taxes are important to political campaigns, but the coefficient is relatively small (reflecting a .0001 increase in taxes per person, which may mean a lot across the whole population, but probably not much to the individual tax payer.
polpc	6.468	Yes	Yes, the standard error is relatively low compared to the coefficient value (1.335)	It is likely that an increase in police per capita results in an increase in crime rate primarily because more criminals are apprehended. We also tend to see higher police per capita as density increases (positive correlation), so multicollinearity could also account partially for the sign of this value.
density	.00005	Yes	Yes, with a relatively low standard error of .00001.	As density increases by .00005, crime rate tends to increase by .005 percent.
log(pctymle)	.018	Yes	Yes, with a relatively low standard error of .005	A .005 percent increase in percent young male results in .005 percent increase in crime rate.
prbarr	-.049	Yes	Yes, with a relatively low standard error of .009.	This model demonstrates that an increase of .049 in the ratio arrests to offenses, results in a decrease of .049 in log(crmrte). In other words, a .049 increase in the ratio of convictions to arrests will result in a 4.9% decrease in crime rate.
log(pctmin80)	.005	Yes	Yes, with a relatively low standard error of .	As such, our intercept is a good estimate of variables not accounted for in the model.
prbpris	-.001	No	No, the standard error is greater than the value.	The probability of prison, or the ratio convictions resulting in a prison sentence to total convictions might seem relevant. However, our extended model demonstrates that it is not.
avgsen	-.004	No	No, the standard error is equal to the value.	This variable is good news, as it demonstrates that length of prison sentence does not predict crime rate.
wser	-.0001	Yes	Limited practical significance. It has a relatively high standard error.	The service industry wages do predict crime rate in this model with low practical significance. However, we will see in joint hypothesis testing that, while this value is individually significant, it is not jointly significant among other wage values.
wfed	-.0001	Yes	Limited practical significance. It has a relatively high standard error.	Federal wages wages do predict crime rate in this model with low practical significance. However, we will see in joint hypothesis testing that, while this value is individually significant, it is not jointly significant among other wage values.

Table displays in a digestible format the value, statistical, and practical significance for important variables in the extended model. It does not take joint statistical significance, interaction terms or indicators into consideration. We excluded most variables that were neither statistically nor practically significant from this table. Those we included were to generate discussion. However, the regression table in section 3.3 Extended Model Assumptions contains these variables and associated values. Coef. refers to the coefficient, Stat Sig refers to statistical significance with a p value less than .05, practical significance refers to interpretation in terms of standard error. Interpretation refers to the relevance to the political campaign.

3.2 Extended Model Interpretation

Indicators

We examined whether specific regions had an influence on crime rate. We found that the west and central regions have no impact on crime rate. In other words, variance in crime rate does not differ with state region. However, urban areas tended to have higher crime rates. In fact a one unit increase in urban areas results in a 18.0% increase in crime rate. While urban areas tend to attract more crime, we used density in our final models, as this was easier to interpret across counties especially for those who do not have major metropolitan areas. However, for those counties with urban centers, this variable is worth considering. An extended model identical to the one above that excluded density due to multicollinearity concerns but included urban still had an adjusted R^2 value of 78.5.

Interestingly, excluding density and including urban only changed our ratio of convictions to arrests variable by .002.

Interactions

Additional interaction variables were examined throughout the models. In the optimization model pctmin80:west, density:urban, and density:logpctymle were explored due to their high correlation with each other. They were all found to be non statistically significant. In the extended model an additional interaction variable was added between prbconv:prbarr. This was not statistically significant as well and demonstrated that an increase in the ratio of convictions to offenses results in an increase in crime rate of 11.48%. While an increase in convictions compared to arrests (prbconv) results in a decrease in the crime rate, the same is not true when convictions are increased in relation to offenses according to our models. While these variables are individually statistically significant, their interactions are not.

Robustness

For each of these models, prbconv had a negative value. Any increase in prbconv resulted in a decrease in crime rate. An increase in polpc resulted in an increase in crime rate. This is likely due to the fact that areas with higher crime tend to have more police per capita. It likely does not reflect additional police causing higher crime rates.

Joint Statistical Test Results

We used heteroskedastically robust values for all joint statistical tests, as models not discussed or considered for primary analysis, but that were included for joint statistical testing, demonstrated mild heteroskedasticity.

Base Model and Optimization Model

The additional variables included in our Optimization Model relative to our base model are jointly statistically significant. All of these variables, except for taxpc, are also individually statistically significant. The joint significance justifies the inclusion of the variables. We continue to include tax per capita, as it has high practical significance to the political campaign. Taxes influence, votes, the economy, infrastructure and other important aspects to the community. It also allows us to account for income in a variety of different of counties, since taxes paid in North Carolina vary with income.

Optimization Model and Extended Model

The optimization STEP model served as the restricted model for all tests below.

We also ran a test to determine the joint significance of our wage variables. We found that our wage variables were likely jointly statistically insignificant. Though federal wages and service wages appeared to be individually significant, they were not jointly significant. We exclude these variables for the sake of parsimony. However, future study might be worthwhile here, since these wage values can be influenced by state and national laws.

We also determined that we failed to reject the null hypothesis for joint significance of probability of prison and average sentence. These variables are likely individually and jointly insignificant.

3.3 Extended Model Assumptions

Overall, this model does not achieve standards to serve as the best linear unbiased estimator due to violation of CLM 2.0, Random Sampling and independence. Specifically, counties (23, 25 and 82) are outliers. Our data indicates that remaining values for the extended model are potentially predictive, excluding the fact that we do not have random sampling that tends to exclude counties with low populations. We used homoskedastically robust standard error for our model interpretation. For more in depth discussion of the assumptions, refer to the appendix for more in depth discussion. This model fails to achieve standards for best linear unbiased estimator due to lack of random sampling. To see the heteroskedastically robust standard errors, refer to the appendix. These are worth consulting for the model, though the model is homoskedastic according to the near-constant variance test. It's scale-location plot appears mildly heteroskedastic.

4.0 Omitted Variable Bias

Omitting an explanatory variable from linear regression biases the estimation procedure whenever the omitted explanatory variable:

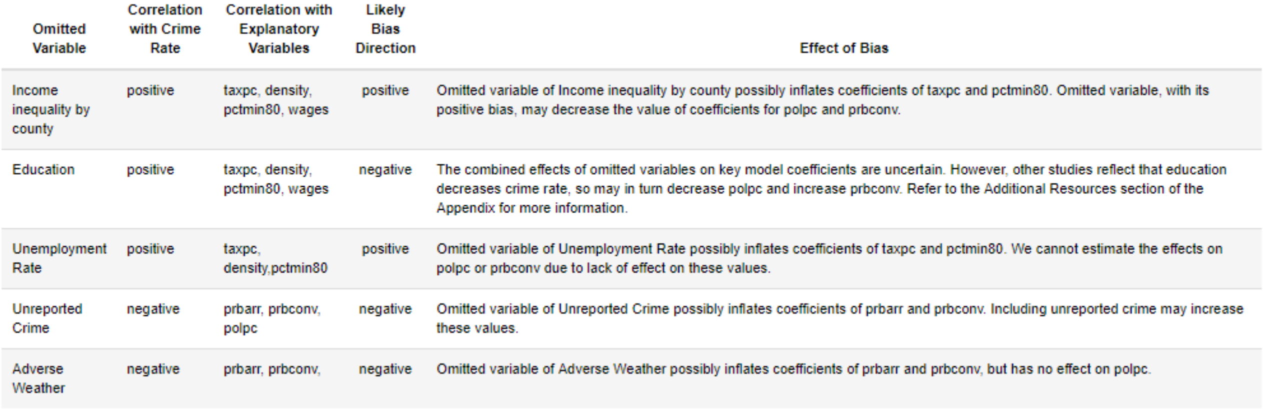
- influences the dependent variable
- is correlated with an included explanatory variable.

When these happen, the coefficient estimate of the included explanatory variable is a composite of two effects:

- the coefficient estimate of the explanatory variables included in the model reflects the influence that the included explanatory variable has on the dependent variable (direct effect)
- omitted explanatory variable has on the dependent variable because the included explanatory variable also acts as a proxy for the omitted explanatory variable (proxy effect).

The analysis for the Omitted Variable section included all the variables since omitted variables could make one (or more) of the variables not included in the model significant and impactful or the opposite, make an included significant variable non significant.

Omitted Variable Bias Summary



Omitted Variable	Correlation with Crime Rate	Correlation with Explanatory Variables	Likely Bias Direction	Effect of Bias
Income inequality by county	positive	taxpc, density, pctmin80, wages	positive	Omitted variable of Income inequality by county possibly inflates coefficients of taxpc and pctmin80. Omitted variable, with its positive bias, may decrease the value of coefficients for polpc and prbconv.
Education	positive	taxpc, density, pctmin80, wages	negative	The combined effects of omitted variables on key model coefficients are uncertain. However, other studies reflect that education decreases crime rate, so may in turn decrease polpc and increase prbconv. Refer to the Additional Resources section of the Appendix for more information.
Unemployment Rate	positive	taxpc, density,pctmin80	positive	Omitted variable of Unemployment Rate possibly inflates coefficients of taxpc and pctmin80. We cannot estimate the effects on polpc or prbconv due to lack of effect on these values.
Unreported Crime	negative	prbarr, prbconv, polpc	negative	Omitted variable of Unreported Crime possibly inflates coefficients of prbarr and prbconv. Including unreported crime may increase these values.
Adverse Weather	negative	prbarr, prbconv,	negative	Omitted variable of Adverse Weather possibly inflates coefficients of prbarr and prbconv, but has no effect on polpc.

The table above summarizes possible omitted variables and their likely impact on variables they most likely correlate with. While we outline the potential effect of each of these variables on polpc and prbconv as per our research question in the interpretation section, we also discuss additional variables that may be of interest to the political campaign and are most influenced by the omitted variables identified above.

Causality Discussion

The findings from our analysis are not causal. The data only covers one year, and it is not randomly selected. Any trends identified in the data are not proven repetitively over time. The model may help drive policy that reduces crime rate, however, additional studies and experimentation must be completed, with modern data focused on specific impacts, to reach causality. This model is not truly predictive either, due to the age of the data and the selection of one year out of a broader multi-year study.

Additional analysis of the omitted variables above is also important to making the model predictive. Examination of endogeneity is important to predicting the effects of including these omitted variables in future studies. We include the analysis here as a starting point for future study.

Endogeneity occurs when we have correlation between our error term u and our independent variables. Each of our models discussed demonstrated exogeneity, which means our independent variables are not correlated with our error term. As such, it is likely that our omitted variables are not accounted for via multicollinearity in the model. However, any potential influence may be small, because the omitted variables are likely not correlated with polpc and prbconv. To observe our plots and zero conditional means tests demonstrating exogeneity, refer to the assumption section for each model in the appendix.

These omitted variables may still have an impact on the amount of variance in crime rate explained by our model, increasing our adjusted R^2 value. This effect is challenging to estimate without additional data.

Additional analysis of the omitted variables would increase model relevance to aid in policy development.

5.0 Conclusion

Research Question: *How can North Carolina affect the crime rate by influencing police per capita and the likelihood of conviction?

Our hypothesis that polpc and prbconv have a statistically significant impact on crime rate hold across the three primary models and the robustness of the key variables of interest.

Initial hypothesis: Police per capita and probability of conviction will have statistically significant effects on crime rate. To test this, we start with the following formal hypothesis.

Note, we conducted numerous hypothesis tests throughout our analysis.

Null Hypothesis:

1) 'Probability' of conviction (prbconv) has no effect on crime rate.
2) Police per capita (polpc) has no effect on crime rate.

Alternate Hypothesis:

1) 'Probability' of conviction (prbconv) has a statistically significant effect on crime rate.
2) Police per capita (polpc) has a statistically significant effect on crime rate.

Actionable Political Implications:

This study suggests that politicians should change policy to increase the ratio of convictions to arrests (prbconv). This variable, across all models, reduced crimes committed per person by approximately 50% to 70% for every unit increase in prbconv. If policy creators can affect a small increase in prbconv, they may see a reduction in crime. This data is not predictive, and merely serves as a descriptive guideline. Further study across time and accounting for additional factors identified in our omitted variable bias section are required.

This study also indicates that politicians consider the fact that an increase in police per capita appears to explain some increase in crime. This is likely due to multicollinearity with density, as denser populations tend to have more police per capita. Politicians can leverage these police to arrest more individuals who have committed crimes with a higher likelihood of conviction, thus potentially reducing crime rate in the most affected areas as identified by our indicator variables. Urban areas experience higher crime rates. While likelihood of arrest was not examined in the simple regression model, the policy makers should focus on increasing convictions instead of reducing arrests. Probability of arrests (prbarr) accounts for reduction in crime rate much like probability of conviction (prbconv), even when both variables are present in the model and both interact.

The models propose that the likelihood of conviction, likelihood of arrest, police per capita and density are the most statistically significant drivers of crime rate in North Carolina. Based on these findings, increased police per capita especially in densely populated counties as well as measures of crime deterrent that aids likelihood of arrest and conviction will likely reduce crime rate in North Carolina.

Further Discussion:

The optimization STEP model best addresses the research question. It balances complexity with explanation of influences on crime rate. The Simple Regression model is a straightforward and useful model, but it excludes important variables in favor of practical significance. The optimization model includes key variables, like tax per capita that are interesting to the campaign even if they are not interesting statistically. Finally, the Extended Model is too complex. It sacrifices interpretation for substantial efforts at including a number of different variables to reach maximum explanation of variance in crime rate. Even so, some variables are included, like separated wage variables, that are interesting to the political campaign but may misrepresent what actually influences crime rate.

The main finding concerning polpc and prbconv is robust to alternative model specifications. Across all three models prbconv was statistically significant and had an inverse relationship with crime rate. Polpc changed signs and significance between the simple regression model and the optimal model. However, when controlling for additional variables and thus accounting for multi-collinearity, polpc was both statistically and practically significant.

All of the models grant opportunities for additional study including how recidivism interacts with probability of arrest, conviction, sentencing to prison, and time in prison. Additionally, 10 counties are missing with no explanation. Finally, there were two dependent variables in the dataset and crmrte and mix. Mix is difficult for politicians to influence and was not examined closer in any of the models.

Appendix

Table of Contents

Libraries Used

In [1]:

```
options(warning=-1)
library(kableExtra)
library(ggplot2)
library(corrplot)
library(dplyr)
library(car)
library(lmtest)
library(sandwich)
library(stargazer)
library(tidyverse)
library(caret)
library(leaps)
library(gridExtra)
library(gvlma)
library(moments)
library(standardize)
```

```
Warning message:
"package 'ggplot2' was built under R version 3.5.3"corrplot 0.84 loaded

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

Loading required package: zoo
Warning message:
"package 'zoo' was built under R version 3.5.2"
Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


Please cite as:

 Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Table
s.
 R package version 5.2.2. https://CRAN.R-project.org/package=stargazer (https://CRAN.R-pr
oject.org/package=stargazer)

-- Attaching packages --------------------------------------- tidyverse 1.2.1 --
v tibble  2.0.1     v purrr   0.2.5
v tidyr   0.8.1     v stringr 1.3.1
v readr   1.3.1     v forcats 0.4.0
Warning message:
"package 'tibble' was built under R version 3.5.2"Warning message:
"package 'readr' was built under R version 3.5.3"Warning message:
"package 'purrr' was built under R version 3.5.2"Warning message:
"package 'forcats' was built under R version 3.5.3"-- Conflicts ------------------------
```

```
----------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x car::recode()   masks dplyr::recode()
x purrr::some()   masks car::some()
Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift

Warning message:
"package 'gridExtra' was built under R version 3.5.3"
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

Exploratory Data Analysis Code and Plots

Section 0.1: Data Load, Analysis, and Cleaning

In [2]:

```
# Load data
crime <- read.csv("crime_v2.csv", stringsAsFactors = FALSE)
dim(crime)
```

97 25

In [3]:

```
# Procedures on data to determine cleaning actions:
str(crime)
summary(crime)
head(crime)
crime[!complete.cases(crime), ]
crime[duplicated(crime$county), ]
```

```
'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 : chr  "0.527595997" "1.481480002" "0.267856985" "0.525424004" ...
 $ 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  453 250 306 281 289
```

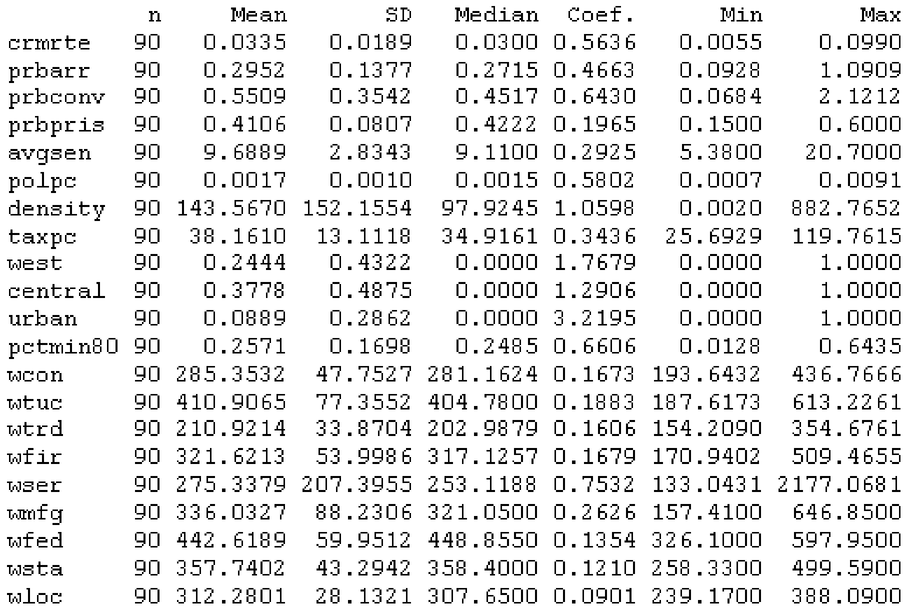
In [4]:

```
# performing cleaning based on data discover
crime <- na.omit(crime)
crime$prbconv <- as.numeric(as.character(crime$prbconv))
crime <- crime[!duplicated(crime$county), ]
crime$year <- NULL
crime$county <- NULL
crime$pctmin80 <- crime$pctmin80/100
crime$density <- crime$density*100   # https://files.nc.gov/ncosbm/demog/dens7095.html
# check results
dim(crime)
```

```
90  23
```

Section 0.2: Data Summary after EDA Cleaning

In [5]:
```
summary_table <- t(sapply(crime, function(x) c(
        "n" = length(x),
        "Mean"= round(mean(x,na.rm=TRUE),4),
        "SD" = round(sd(x),4),
        "Median" = round(median(x),4),
        "Coef." = round(sd(x)/mean(x,na.rm=TRUE),4),
        "Min" = round(min(x),4),
        "Max" = round(max(x),4)
        )
    )
)

options(repr.plot.height = 4.0, repr.plot.width = 6.0, repr.plot.pointsize = 7)
print_output <- function(output, cex = .7) {
  tmp <- capture.output(output)
  plot.new()
  text(0, 1, paste(tmp, collapse='\n'), adj = c(0,1), family = 'mono', cex = cex)
}

print_output(summary_table)
```



```
          n     Mean       SD   Median  Coef.      Min       Max
crmrte   90   0.0335   0.0189   0.0300 0.5636   0.0055    0.0990
prbarr   90   0.2952   0.1377   0.2715 0.4663   0.0928    1.0909
prbconv  90   0.5509   0.3542   0.4517 0.6430   0.0684    2.1212
prbpris  90   0.4106   0.0807   0.4222 0.1965   0.1500    0.6000
avgsen   90   9.6889   2.8343   9.1100 0.2925   5.3800   20.7000
polpc    90   0.0017   0.0010   0.0015 0.5802   0.0007    0.0091
density  90 143.5670 152.1554  97.9245 1.0598   0.0020  882.7652
taxpc    90  38.1610  13.1118  34.9161 0.3436  25.6929  119.7615
west     90   0.2444   0.4322   0.0000 1.7679   0.0000    1.0000
central  90   0.3778   0.4875   0.0000 1.2906   0.0000    1.0000
urban    90   0.0889   0.2862   0.0000 3.2195   0.0000    1.0000
pctmin80 90   0.2571   0.1698   0.2485 0.6606   0.0128    0.6435
wcon     90 285.3532  47.7527 281.1624 0.1673 193.6432  436.7666
wtuc     90 410.9065  77.3552 404.7800 0.1883 187.6173  613.2261
wtrd     90 210.9214  33.8704 202.9879 0.1606 154.2090  354.6761
wfir     90 321.6213  53.9986 317.1257 0.1679 170.9402  509.4655
wser     90 275.3379 207.3955 253.1188 0.7532 133.0431 2177.0681
wmfg     90 336.0327  88.2306 321.0500 0.2626 157.4100  646.8500
wfed     90 442.6189  59.9512 448.8550 0.1354 326.1000  597.9500
wsta     90 357.7402  43.2942 358.4000 0.1210 258.3300  499.5900
wloc     90 312.2801  28.1321 307.6500 0.0901 239.1700  388.0900
```

Preliminary Variable Analysis

Section 0.3: Region & Urban Variable Analysis

In [6]:

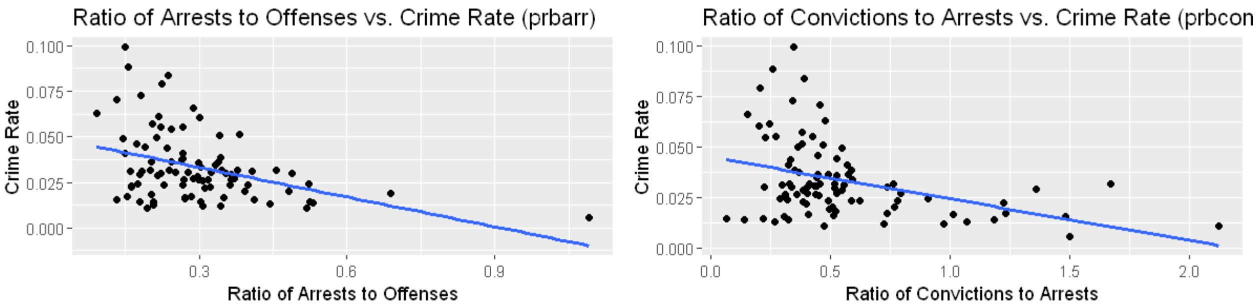
```
# number of western, central, and other counties
c(sum(crime$west == 1),
  sum(crime$central == 1),
  length(crime$west) - sum(crime$west == 1) - sum(crime$central == 1),
  length(crime$west)
 )

# number of urban counties by region
c(sum(crime$west == 1 & crime$urban == 1),
  sum(crime$central == 1 & crime$urban == 1),
  sum(crime$urban == 1) - sum(crime$west == 1 & crime$urban == 1) - sum(crime$central == 1
                                                                        & crime$urban == 1),

  sum(crime$urban == 1)
  )
```

```
22   34   34   90

1   5   2   8
```

Section 0.4: Probability of Arrests (prbarr) & Convictions (prbconv)

In [7]:

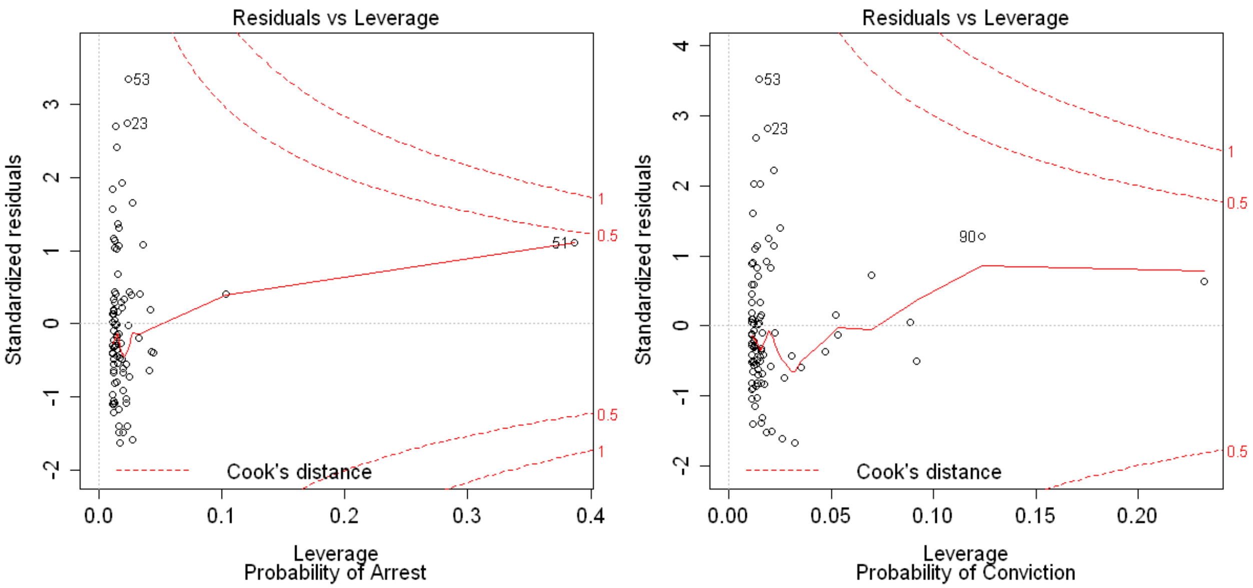
```
options(repr.plot.height = 2.5, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)

plot1 <- ggplot(crime, aes(prbarr, crmrte)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
xlab("Ratio of Arrests to Offenses") +
ylab("Crime Rate") +
ggtitle("Ratio of Arrests to Offenses vs. Crime Rate (prbarr)")

plot2 <- ggplot(crime, aes(prbconv, crmrte)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
xlab("Ratio of Convictions to Arrests") +
ylab("Crime Rate") +
ggtitle("Ratio of Convictions to Arrests vs. Crime Rate (prbconv)")

grid.arrange(plot1, plot2, ncol=2)
```



Negative correlation on both plots indicates that these variables may influence crime rate in a way that is relevant to politicians.

In [8]:

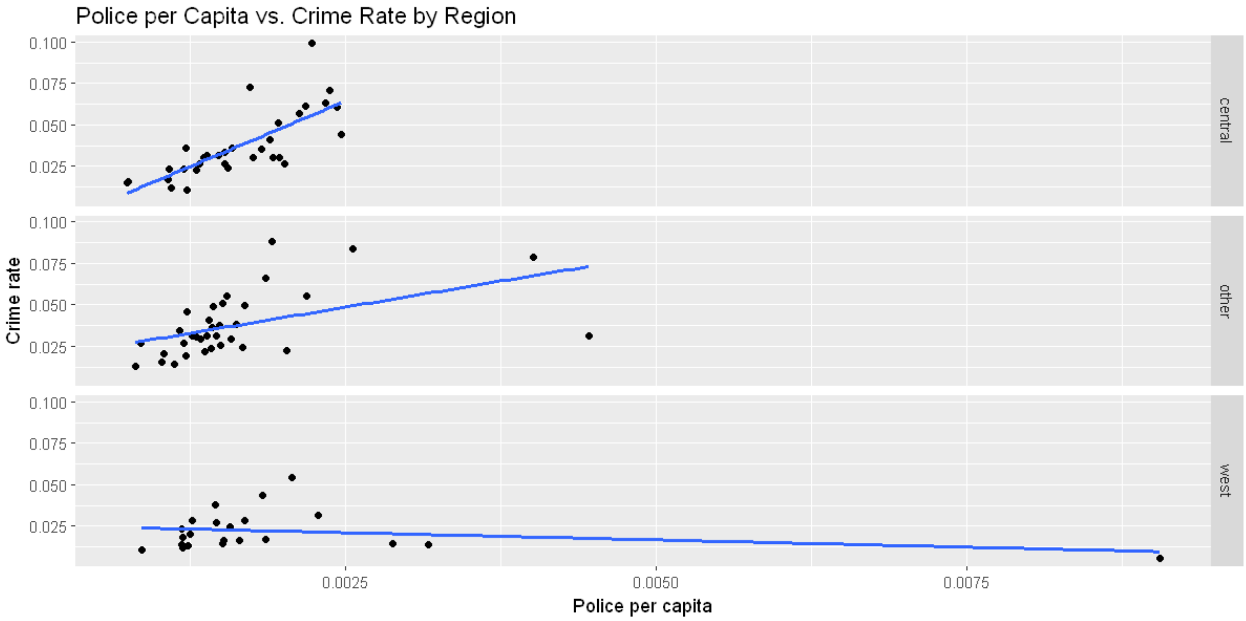
```
modelprbarr <- lm(crmrte ~ prbarr, crime)
modelprbconv <- lm(crmrte ~ prbconv, crime)

options(repr.plot.height = 5, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
par(mfrow = c(1, 2))
plot(modelprbarr, which = 5)
title(sub = "Probability of Arrest")
plot(modelprbconv, which = 5)
title(sub = "Probability of Conviction")
```



No outliers require further examination here. They are all within .5.

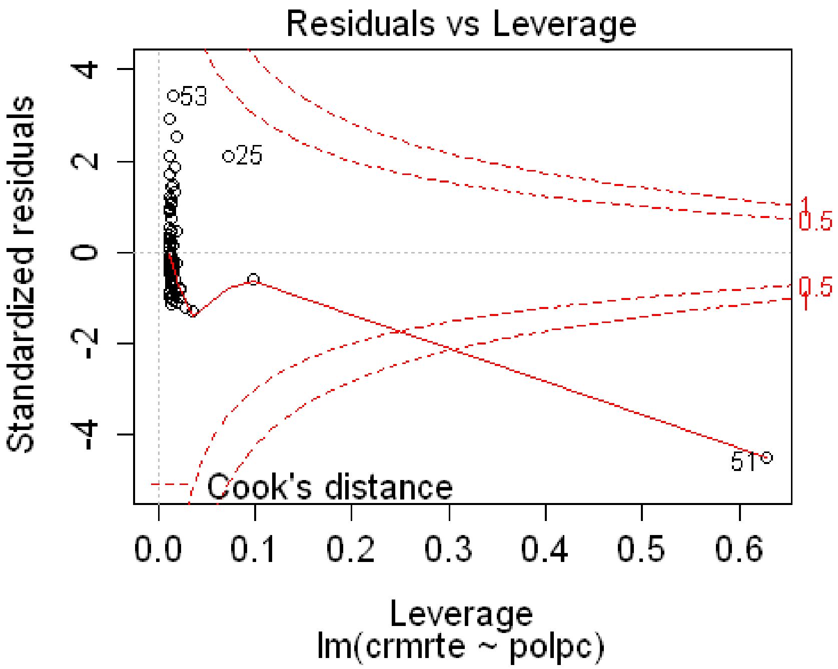
Section 0.5: Police per capita (polpc)

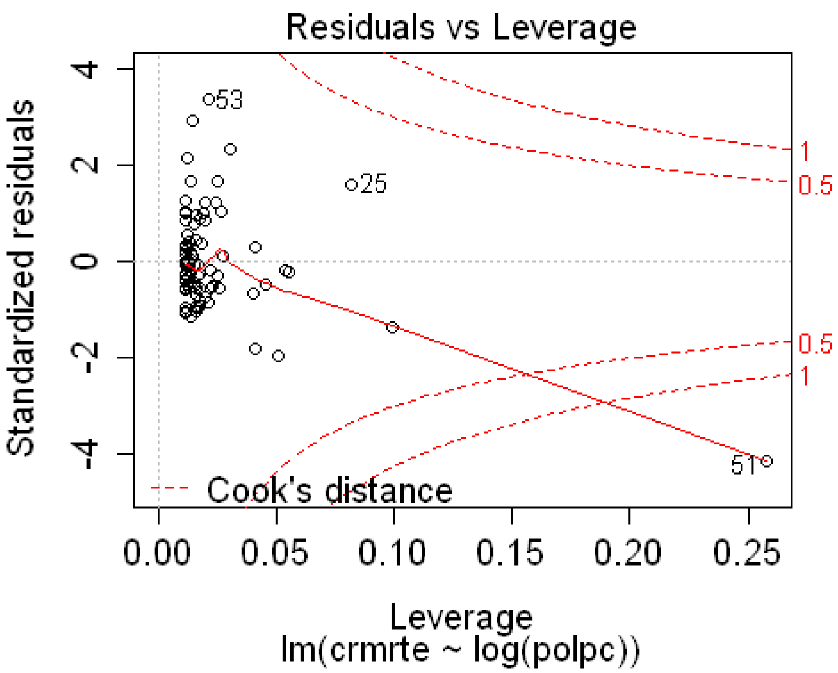
In [9]:
```
# create regions variable
crime$region <- ifelse(crime$west == 1, "west", ifelse(crime$central == 1, "central", "other"
options(repr.plot.height = 5.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
# Police per Capita vs. Crime Rate by Region
ggplot(crime, aes(polpc, crmrte)) + geom_point() + facet_grid(region~.) +
    geom_smooth(method = "lm", se = FALSE) + xlab("Police per capita") + ylab("Crime rate") +
    ggtitle("Police per Capita vs. Crime Rate by Region")
```



Clear correlations here may be worth considering. These are excellent indicator variables and good for regionally focused politicians.

In [10]:
```
options(repr.plot.height = 3.5, repr.plot.width = 4.0, repr.plot.pointsize = 9.0)
modelpolpc <- lm(crmrte ~ polpc, crime)
plot(modelpolpc, which = 5)

modellogpolpc <- lm(crmrte ~ log(polpc), crime)
plot(modellogpolpc, which = 5)
```



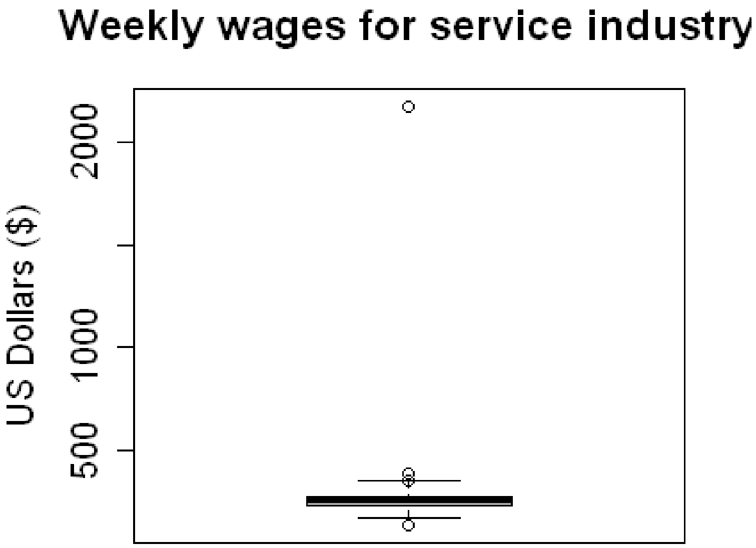


Observation 51 was worth exploring here. It turns out that this value is likely accurate due to density of the county and characteristics.

Section 0.6: Weekly Wage Analysis

In [11]:

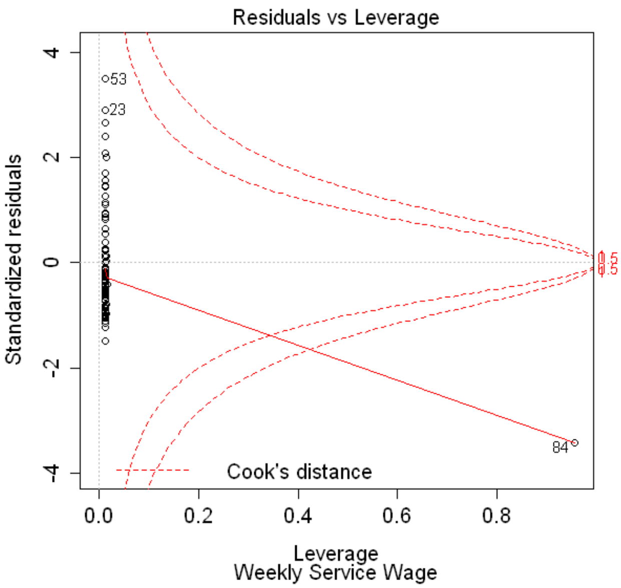
```
#wser variable looks odd, let's do a boxplot and take a look
options(repr.plot.height = 3.5, repr.plot.width = 3.5, repr.plot.pointsize = 9.0)
boxplot(crime$wser, main = "Weekly wages for service industry", ylab = "US Dollars ($)",
        breaks = 10, col = "darkgrey")
```

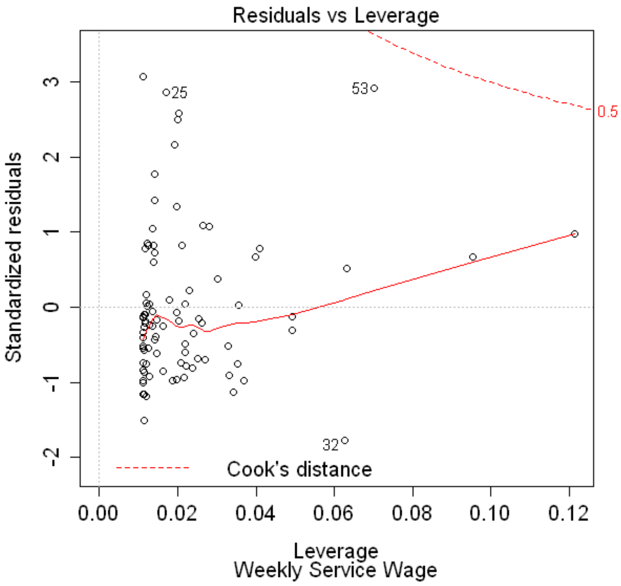


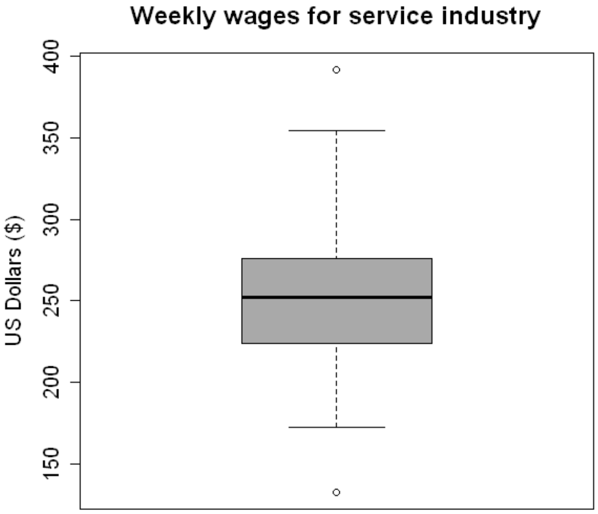
We have a significant outlier here that is not supported by real world data. This county is actually quite poor. It *looks like a decimal was misplaced. Let's move it and see if it reflects this county more accurately.*

In [12]:

```
modelwser <- lm(crmrte ~ wser, crime)

options(repr.plot.height = 5, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
par(mfrow = c(1, 2))
plot(modelwser, which = 5)
title(sub = "Weekly Service Wage")
crime$wser[84] = 217.7068
modelwser <- lm(crmrte ~ wser, crime)

options(repr.plot.height = 5, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
par(mfrow = c(1, 2))
plot(modelwser, which = 5)
title(sub = "Weekly Service Wage")
boxplot(crime$wser, main = "Weekly wages for service industry", ylab = "US Dollars ($)",
        breaks = 10, col = "darkgrey")
```







This plot now better reflects real world verified data with the removal of the variable.

Section 0.7: Summary Table

In [13]:

```r
ind_variables <- c('crmrte', 'prbarr', 'prbconv', 'prbpris', 'avgsen',
'polpc', 'density', 'taxpc', 'west', 'central', 'urban', 'pctmin80', 'wcon',
'wtuc', 'wtrd', 'wfir', 'wser', 'wmfg', 'wfed', 'wsta', 'wloc', 'mix',
'pctymle'
)
var_labels <- c('crimes committed per person',
'ratio of arrests to offenses', 'ratio of convictions to arrests',
'ratio of prison sentence to total convictions', 'avg. sentence, days',
'police per capita', 'people per sq. mile', 'tax revenue per capita',
'=1 if in western N.C.', '=1 if in central N.C.', '=1 if in SMSA',
'perc. minority, 1980', 'weekly wage, construction',
'wkly wge, trns, util, commun', 'wkly wge, whlesle, retail trade',
'wkly wge, fin, ins, real est', 'wkly wge, service industry',
'wkly wge, manufacturing', 'wkly wge, fed employees',
'wkly wge, state employees', 'wkly wge, local gov emps',
'offense mix: face-to-face/other', 'perc. male between ages 15 and 24, 1980'
)
impact <- c("Dependent", "Negative" , "Negative", "Negative", "Negative",
"Negative", "Positive", "Negative", "Unclear", "Unclear", "Unclear", "Unclear",
"Negative","Negative","Negative", "Negative", "Negative", "Negative", "Negative",
"Negative", "Negative", "Unclear","Positive"
)
control <- c("Yes","Yes", "Yes", "Yes", "Yes",
"Yes", "No", "Yes", "No", "No", "No","No",
"Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No",
"Yes", "Yes", "No", "No"
)
last <- c("Medium","Medium", "Medium", "Short", "Short",
"Medium", "Long", "Medium", "Long", "Long", "Long","Long",
"Medium", "Medium", "Medium", "Medium", "Medium", "Medium", "Medium",
"Short", "Medium", "Long", "Long"
)

avg <- round(summary_table[,2],3)
stddev <- round(summary_table[,3],3)
cor_crimerate <- round(cor(crime[,ind_variables])[1,],2)

desc <- data.frame(ind_variables, var_labels, impact, avg, stddev, cor_crimerate, control, la

colnames(desc) <- c("Explanatory Variables",
"Explanation",
"Expected Impact on Crime Rate",
"Mean",
"SD",
"Correlation w/ Crime Rate",
"Can Gov Impact This?",
"Policy Timeframe"
)

kable(desc, booktabs = TRUE, align = c("llccc")) %>%
    kable_styling(latex_options = c("striped", "hover", "condensed"),
    full_width = FALSE) %>%
    row_spec(0, bold = TRUE) %>%
    column_spec(1, width = "8em") %>%
    column_spec(3, width = "10em") %>%
    column_spec(4, width = "8em") %>%
    column_spec(5, width = "9em") %>%
    column_spec(6, width = "9em") %>%
    column_spec(7, width = "8em") %>%
    column_spec(8, width = "9em")
```

```
<table class="table" style="width: auto !important; margin-left: auto; margin-right: aut
o;">
 <thead>
```

```
  <tr>
   <th style="text-align:left;font-weight: bold;"> Explanatory Variables </th>
   <th style="text-align:left;font-weight: bold;"> Explanation </th>
   <th style="text-align:center;font-weight: bold;"> Expected Impact on Crime Rate </th>
   <th style="text-align:center;font-weight: bold;"> Mean </th>
   <th style="text-align:center;font-weight: bold;"> SD </th>
   <th style="text-align:left;font-weight: bold;"> Correlation w/ Crime Rate </th>
   <th style="text-align:left;font-weight: bold;"> Can Gov Impact This? </th>
   <th style="text-align:center;font-weight: bold;"> Policy Timeframe </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;width: 8em; "> crmrte </td>
   <td style="text-align:left;"> crimes committed per person </td>
   <td style="text-align:center;width: 10em; "> Dependent </td>
```

Explanatory Variables	Explanation	Expected Impact on Crime Rate	Mean	SD	Correlation w/ Crime Rate	Can Gov Impact This?	Policy Timeframe
crmrte	crimes committed per person	Dependent	0.034	0.019	1.00	Yes	Medium
prbarr	ratio of arrests to offenses	Negative	0.295	0.138	-0.40	Yes	Medium
prbconv	ratio of convictions to arrests	Negative	0.551	0.354	-0.39	Yes	Medium
prbpris	ratio of prison sentence to total convictions	Negative	0.411	0.081	0.05	Yes	Short
avgsen	avg. sentence, days	Negative	9.689	2.834	0.02	Yes	Short
polpc	police per capita	Negative	0.002	0.001	0.17	Yes	Medium
density	people per sq. mile	Positive	143.567	152.155	0.73	No	Long
taxpc	tax revenue per capita	Negative	38.161	13.112	0.45	Yes	Medium
west	=1 if in western N.C.	Unclear	0.244	0.432	-0.35	No	Long
central	=1 if in central N.C.	Unclear	0.378	0.488	0.17	No	Long
urban	=1 if in SMSA	Unclear	0.089	0.286	0.62	No	Long
pctmin80	perc. minority, 1980	Unclear	0.257	0.170	0.18	No	Long
wcon	weekly wage, construction	Negative	285.353	47.753	0.39	Yes	Medium

Explanatory Variables	Explanation	Expected Impact on Crime Rate	Mean	SD	Correlation w/ Crime Rate	Can Gov Impact This?	Policy Timeframe
wtuc	wkly wge, trns, util, commun	Negative	410.906	77.355	0.24	Yes	Medium
wtrd	wkly wge, whlesle, retail trade	Negative	210.921	33.870	0.43	Yes	Medium
wfir	wkly wge, fin, ins, real est	Negative	321.621	53.999	0.34	Yes	Medium
wser	wkly wge, service industry	Negative	275.338	207.395	-0.05	Yes	Medium
wmfg	wkly wge, manufacturing	Negative	336.033	88.231	0.35	Yes	Medium
wfed	wkly wge, fed employees	Negative	442.619	59.951	0.49	No	Medium
wsta	wkly wge, state employees	Negative	357.740	43.294	0.20	Yes	Short
wloc	wkly wge, local gov emps	Negative	312.280	28.132	0.36	Yes	Medium
mix	offense mix: face-to-face/other	Unclear	0.129	0.082	-0.13	No	Long
pctymle	perc. male between ages 15 and 24, 1980	Positive	0.084	0.023	0.29	No	Long

Data Transformation Analysis

Section 0.8: Data Transformations Preparation

In [14]:

```
# get an average wage
crime$wklywage <- (crime$wcon+crime$wtuc+crime$wtrd+crime$wfir+crime$wser+crime$wmfg+
                   crime$wfed+crime$wsta+crime$wloc)/9
```

In [15]:

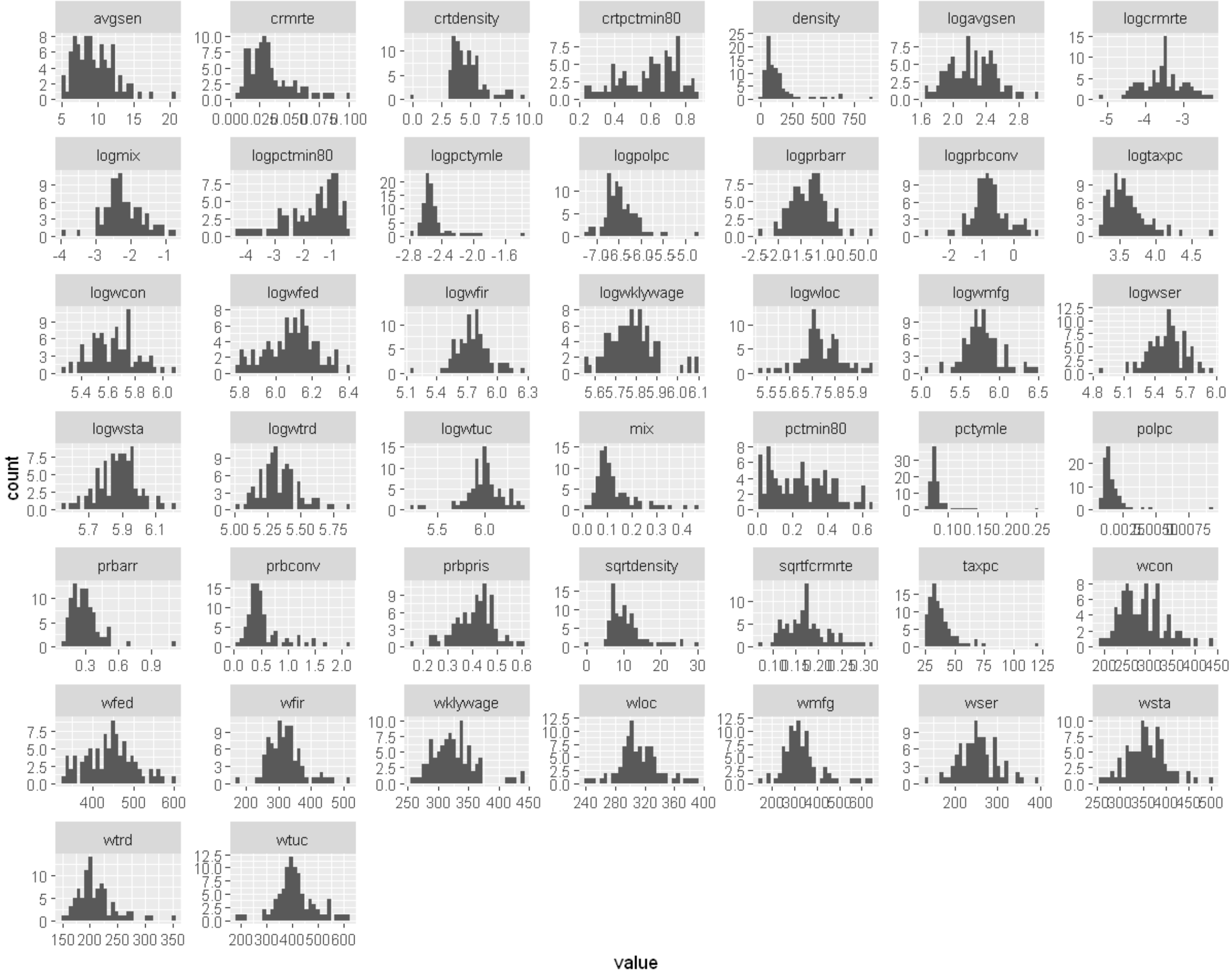
```
# Log transforms and sqrt two of the variables for exploration
crime$sqrtfcrmrte <- sqrt(crime$crmrte)
crime$sqrtdensity <- sqrt(crime$density)
crime$logcrmrte <- log(crime$crmrte)
crime$logpolpc <- log(crime$polpc)
crime$logtaxpc <- log(crime$taxpc)
crime$logprbarr <- log(crime$prbarr)
crime$logprbconv <- log(crime$prbconv)
crime$logpctymle <- log(crime$pctymle)
crime$logmix <- log(crime$mix)
crime$logavgsen <- log(crime$avgsen)
crime$logpctmin80 <- log(crime$pctmin80)
# wages
crime$logwcon <- log(crime$wcon)
crime$logwtuc <- log(crime$wtuc)
crime$logwtrd <- log(crime$wtrd)
crime$logwfir <- log(crime$wfir)
crime$logwser <- log(crime$wser)
crime$logwmfg <- log(crime$wmfg)
crime$logwfed <- log(crime$wfed)
crime$logwsta <- log(crime$wsta)
crime$logwloc <- log(crime$wloc)
crime$logwklywage <- log(crime$wklywage)
# cube squared
crime$crtdensity <- (crime$density)^(1/3)
crime$crtpctmin80 <- (crime$pctmin80)^(1/3)
```

Section 0.9: Histogram Plots

In [16]:

```
# plot every variable except west, central, urban, and region.  year and county were set to N
options(repr.plot.height = 8, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)

plotdata <- crime[!(names(crime) %in% c("west", "central", "urban", "region"))]
ggplot(gather(plotdata), aes(value)) + facet_wrap(~key, scales="free") + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Section 0.10: Skewness Discovery

In [17]:

```
skewness(plotdata)
```

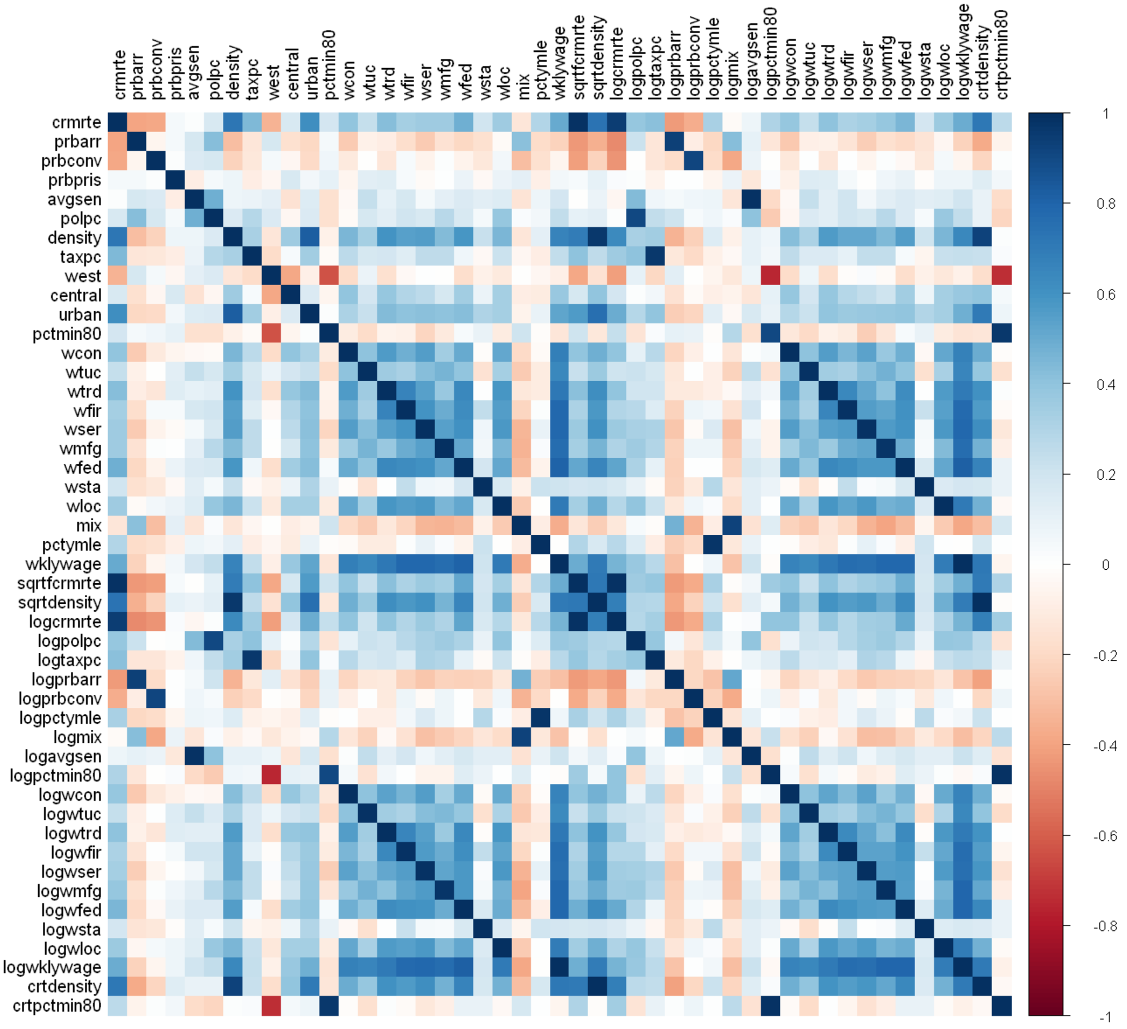
```
     crmrte  1.28174888290774
     prbarr  2.52529595748595
    prbconv  2.03950599070009
    prbpris  -0.452540218833851
     avgsen  1.00116339509738
      polpc  4.9834879501134
    density  2.64350769510431
      taxpc  3.29057446509796
   pctmin80  0.365661694755099
       wcon  0.606802227848182
       wtuc  0.0681976791109087
       wtrd  1.46120657167937
       wfir  0.820631454918991
       wser  0.317109243245966
       wmfg  1.42253165784335
       wfed  0.132237606706429
       wsta  0.362368258673733
       wloc  0.295138077263709
        mix  1.91657046056733
    pctymle  4.56069073587539
   wklywage  0.998814275306913
 sqrtfcrmrte  0.635489953945183
 sqrtdensity  1.50286307506117
   logcrmrte  -0.101791935773808
    logpolpc  1.40836149118853
    logtaxpc  1.54494554081387
   logprbarr  0.299156622783786
  logprbconv  0.0428416270346315
  logpctymle  2.89238604700579
      logmix  0.204657830601069
   logavgsen  0.213025894195614
 logpctmin80  -0.959203693713306
     logwcon  0.187812866591893
     logwtuc  -1.03968405040768
     logwtrd  0.850977095187381
     logwfir  -0.00350869712802304
     logwser  -0.37571182249873
     logwmfg  0.384037225431027
     logwfed  -0.192308425470541
     logwsta  -0.0773450235744132
     logwloc  -0.0757821230224933
 logwklywage  0.604584990243857
  crtdensity  0.832054476774815
 crtpctmin80  -0.442273974821855
```

Correlation Analysis

Section 0.11: Correlation Analysis

In [18]:

```
options(repr.plot.height = 11.0, repr.plot.width = 12.0, repr.plot.pointsize = 9.0)

plotdata <- crime[!(names(crime) %in% c("region"))]
M <- cor(plotdata[, c(1:47)])

corrplot(M, method="color",
         tl.col="black",
         sig.level = 0.01,
         insig = "blank",
         diag=TRUE)
```



Section 1.0: Correlation of crmrte with polpc/logpolpc

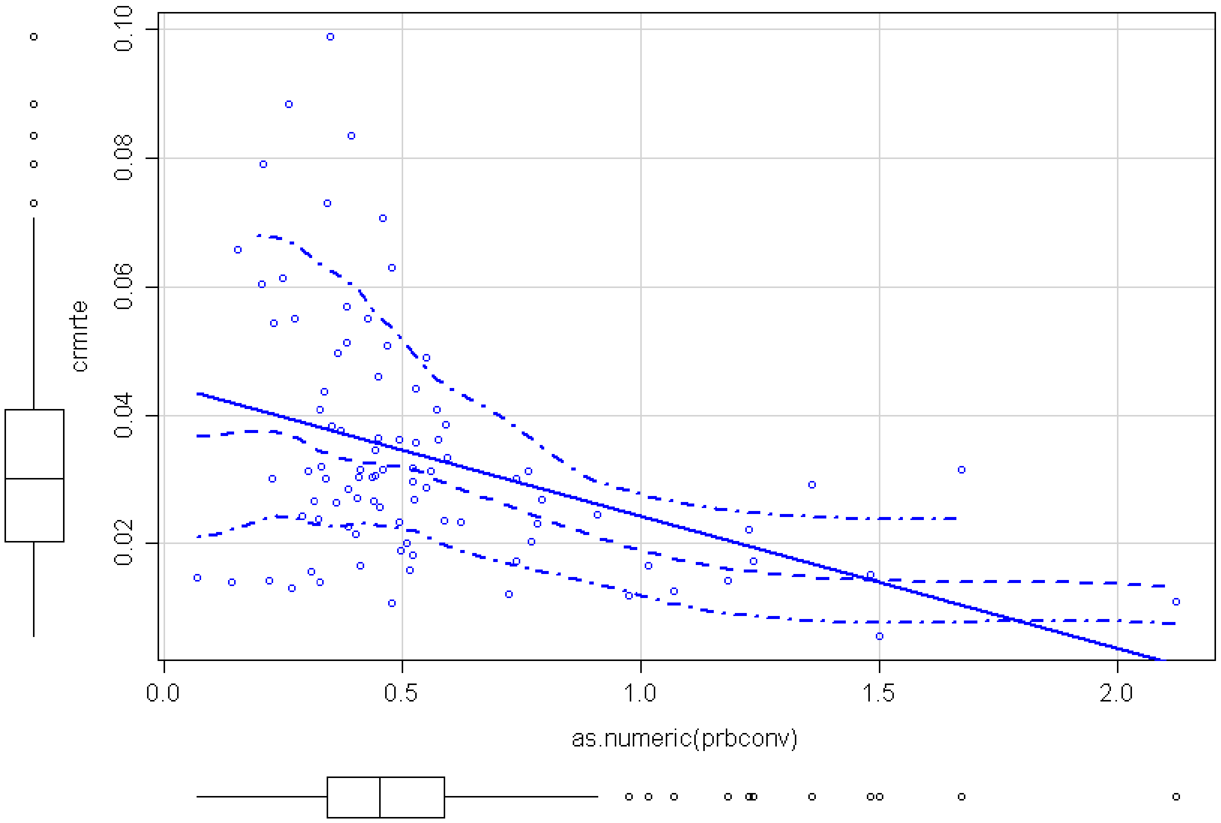
In [19]:

```
round(c(cor(crime$crmrte,crime$polpc),cor(log(crime$crmrte),log(crime$polpc))),4)
```

```
0.1673   0.2845
```

Section 1.1: Plot of crmrte and prbconv

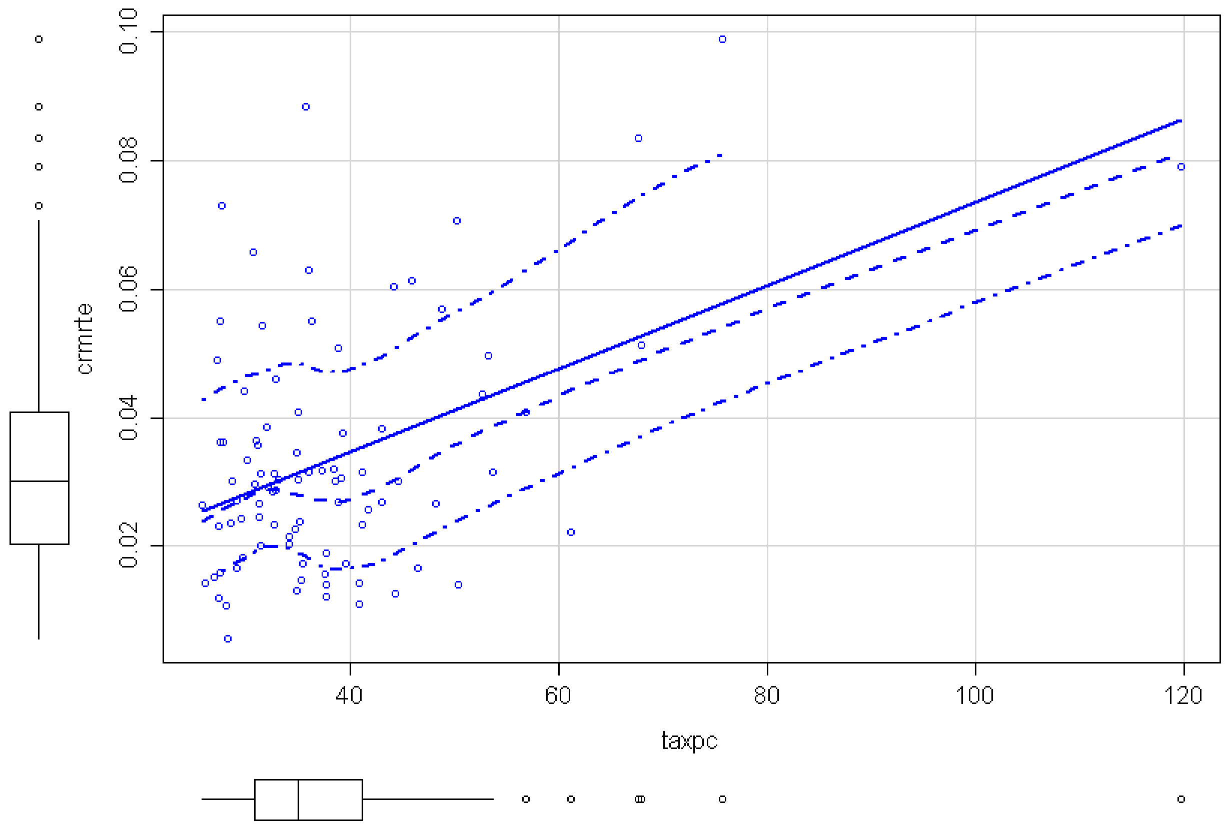
```
In [20]: options(repr.plot.height = 5.0, repr.plot.width = 7, repr.plot.pointsize = 9.0)
         scatterplot(crmrte ~ as.numeric(prbconv), data=crime)
```



The high negative correlation here suggests that prbconv is worth considering further. This variable is cheap and easy to influence.

Section 1.2: Plot of crmrte and taxpc

In [21]:
```
options(repr.plot.height = 5.0, repr.plot.width = 7, repr.plot.pointsize = 9.0)
scatterplot(crmrte ~ taxpc, data=crime)
```



Interesting that paying higher taxes correlates with a higher crime rate. This is worth exploring.

Base Model

Section 1.3: Model 1 Specification

In [22]:

```r
#Model Specification
base_model <- lm(log(crmrte) ~ log(polpc) + as.numeric(prbconv) + taxpc, data = crime)
analysis_model_a <- lm(log(crmrte) ~ log(polpc) + as.numeric(prbconv) + taxpc +
                       as.numeric(prbconv)*log(polpc), data = crime)
analysis_model_b <- lm(log(crmrte) ~ log(polpc) + as.numeric(prbconv) + taxpc +
                       log(polpc)*taxpc, data = crime)
analysis_model_c <- lm(log(crmrte) ~ log(polpc) + as.numeric(prbconv) + taxpc +
                       log(polpc)*taxpc + as.numeric(prbconv)*log(polpc), data = crime)
se.base_model = sqrt(diag(vcovHC(base_model)))
stargazer(base_model, analysis_model_a, analysis_model_b, analysis_model_c,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.base_model),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
===================================================================
                                      Dependent variable:
                               ------------------------------------
                                          log(crmrte)
                                 (1)      (2)       (3)      (4)
-------------------------------------------------------------------
log(polpc)                      0.278   0.859**   -0.068    0.599
                               (0.455)  (0.267)   (0.322)  (0.422)

as.numeric(prbconv)            -0.645** -4.517** -0.653*** -4.308**
                               (0.224)  (1.528)   (0.139)  (1.553)

taxpc                           0.010    0.006     0.069    0.046
                               (0.007)  (0.004)   (0.050)  (0.049)

log(polpc):as.numeric(prbconv)          -0.613*            -0.579*
                                        (0.241)            (0.245)

log(polpc):taxpc                                   0.010    0.007
```

Section 1.4: Model 1 Interpretation

In [23]:

```r
# Table of Coefficient
model1_ind_vars <- c("logpolpc", "prbconv", "taxpc")
model1_formula <- as.formula(paste("log(crmrte) ~ ", paste(model1_ind_vars,
                                                            collapse = " + "), sep = ""))
model1_table <- lm(model1_formula, data = crime)

interpret_model1 <- c(
"",
"The interpretation is more difficult. For each percentage (square root) per square
mile increase in density, the crime rate will increase by 4.84% crimes per person,
while all other variables remain constant.",

"For each percentage increase in pctymle, the crime rate will increase by .34%,
while all other variables remain constant.",

"As the ratio of convictions to arrests increases by 1%, the crime rate will decrease
by 9.4%, while all other variables remain constant."
)

coef1_1 <- data.frame("Coef." = round(base_model$coefficients, 4),
                      "Interpretation" = interpret_model1)

kable(coef1_1, booktabs = TRUE) %>%
kable_styling(font_size = 10, full_width = FALSE) %>%
column_spec(3, width = "35em")
```

```
<table class="table" style="font-size: 10px; width: auto !important; margin-left: auto; m
argin-right: auto;">
 <thead>
  <tr>
   <th style="text-align:left;">   </th>
   <th style="text-align:right;"> Coef. </th>
   <th style="text-align:left;"> Interpretation </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;"> (Intercept) </td>
   <td style="text-align:right;"> -1.7590 </td>
   <td style="text-align:left;width: 35em; ">  </td>
  </tr>
  <tr>
   <td style="text-align:left;"> log(polpc) </td>
   <td style="text-align:right;"> 0.2778 </td>
   <td style="text-align:left;width: 35em; "> The interpretation is more difficult. For e
ach percentage (square root) per square
mile increase in density, the crime rate will increase by 4.84% crimes per person,
while all other variables remain constant. </td>
  </tr>
  <tr>
   <td style="text-align:left;"> as.numeric(prbconv) </td>
   <td style="text-align:right;"> -0.6447 </td>
   <td style="text-align:left;width: 35em; "> For each percentage increase in pctymle, th
e crime rate will increase by .34%,
while all other variables remain constant. </td>
  </tr>
  <tr>
   <td style="text-align:left;"> taxpc </td>
   <td style="text-align:right;"> 0.0096 </td>
   <td style="text-align:left;width: 35em; "> As the ratio of convictions to arrests incr
eases by 1%, the crime rate will decrease
by 9.4%, while all other variables remain constant. </td>
  </tr>
```

```
</tbody>
</table>
```

	Coef.	Interpretation
(Intercept)	-1.7590	
log(polpc)	0.2778	The interpretation is more difficult. For each percentage (square root) per square mile increase in density, the crime rate will increase by 4.84% crimes per person, while all other variables remain constant.
as.numeric(prbconv)	-0.6447	For each percentage increase in pctymle, the crime rate will increase by .34%, while all other variables remain constant.
taxpc	0.0096	As the ratio of convictions to arrests increases by 1%, the crime rate will decrease by 9.4%, while all other variables remain constant.

Base Model Assumptions

Linearity of Parameters - the true relationship between the dependent and explanatory variables is linear. The model is linear in parameters: crime rate is a linear function of tax per capita, likelihood of conviction and police per capita. As observed on the diagnostic regression plot of residuals against fitted values, a random pattern indicates an appropriate model has been fit to the data.

Random Sampling and Independence -for the purpose of the analysis, the data is assumed to have been randomly sampled. In reality, it is challenging to obtain randomly generated and independent geographical survey data

Zero Conditional Mean - the expected value of residuals should be approximately zero. This when calculated as follows confirms zero mean residuals.

No Perfect Multicollinearity- we can detect and represent collinearity using the Variance Inflation Factor (VIF). If any of these variance inflation factors exceed 5 (the cut-off typically used), the associated regression coefficients are poorly estimated due to multicollinearity. This suggests that the Multicollinearity assumption is not violated.

Homoscedasticity- the scale-location plot of the Linear Model diagnostic plots graphically evaluates the assumption that errors are identically- distributed and have homogeneous variance. The plot shows the square root of the absolute residuals against the fitted values, along with a smooth red line. Significant departures from a horizontal line typically suggest heteroscedasticity. Model 1 plot shows non horizontal line which implies heterogeneous variance. Quantitatively, assumption of constant error variance can be tested using the ncvTest function. The score test is not significant (p>.05), we fail to reject our null hypothesis that the model is homoskedastic and therefore use homoskedastically robust standard errors.

Normality of Residuals- the qqPlot suggests slight deviation from normality of residuals; however, our sample size is large enough to handle the skew.

Section 1.5: Zero Conditional Mean

In [24]:

```
round(mean(base_model$residuals), 4)
```

```
0
```

Section 1.6: No Perfect Multicollinearity

```
In [25]: round(vif(base_model), 4)
```

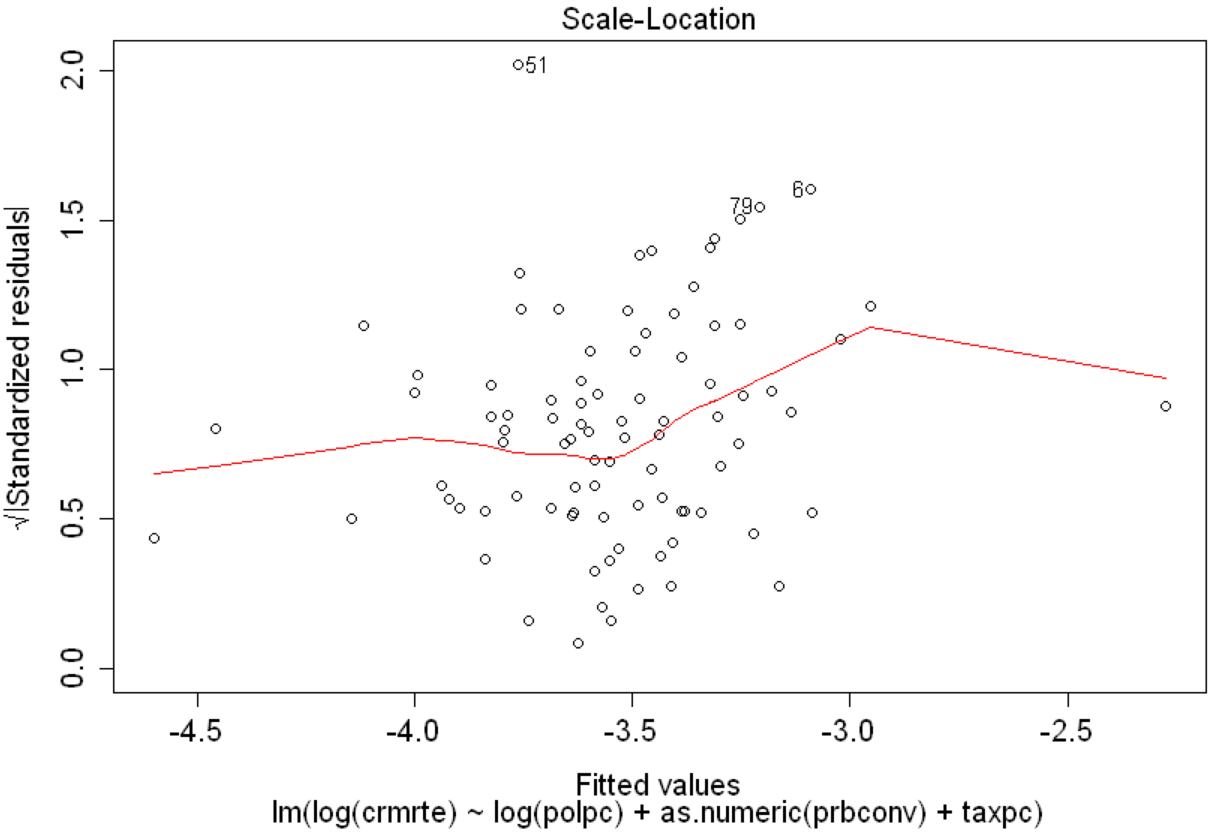
```
          log(polpc)   1.1943
as.numeric(prbconv)    1.0188
               taxpc   1.2139
```

Section 1.7: Homoscedasticity

```
In [26]: ncvTest(base_model)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.213495    Df = 1     p = 0.07303331
```

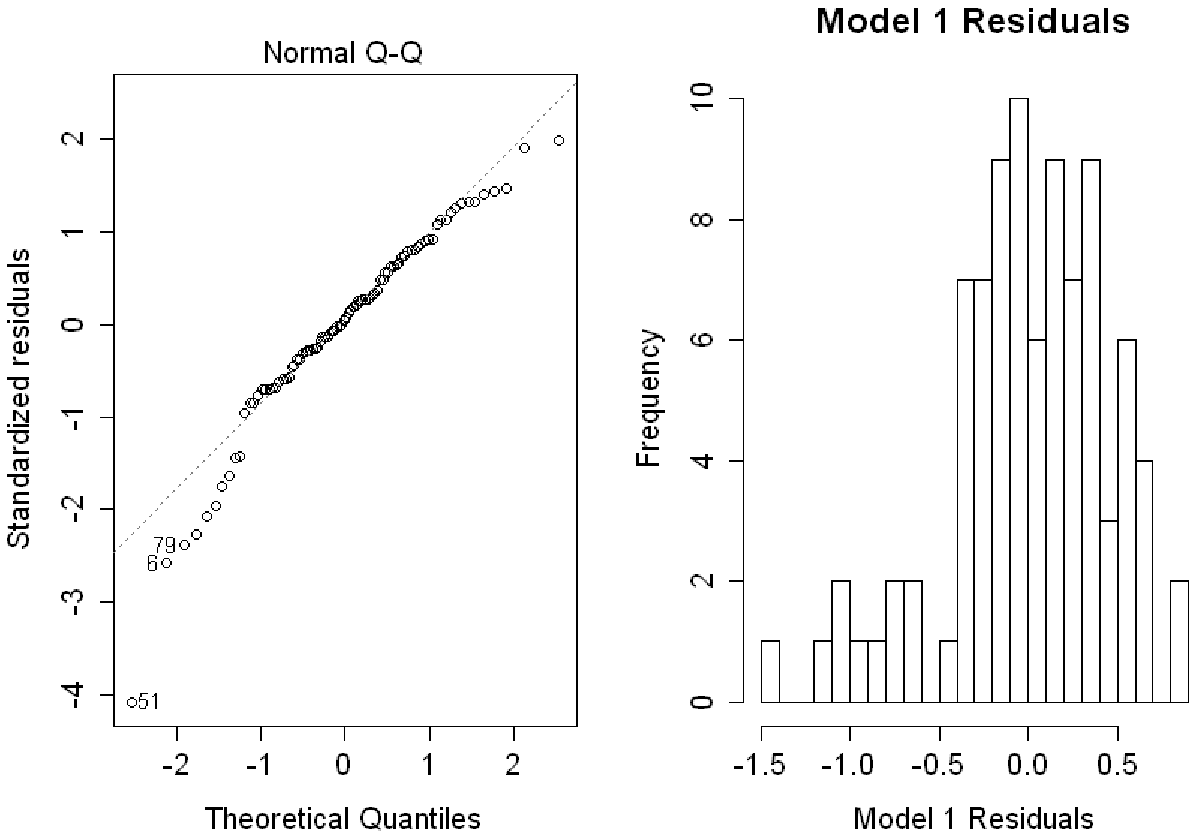
```
In [27]: plot(base_model, which = 3)
```



Section 1.8: Normality of Residuals

In [28]:

```
par(mfrow=c(1,2))
plot(base_model, which = 2)
hist(base_model$residuals, breaks = 20, main = "Model 1 Residuals",
     xlab="Model 1 Residuals")
```



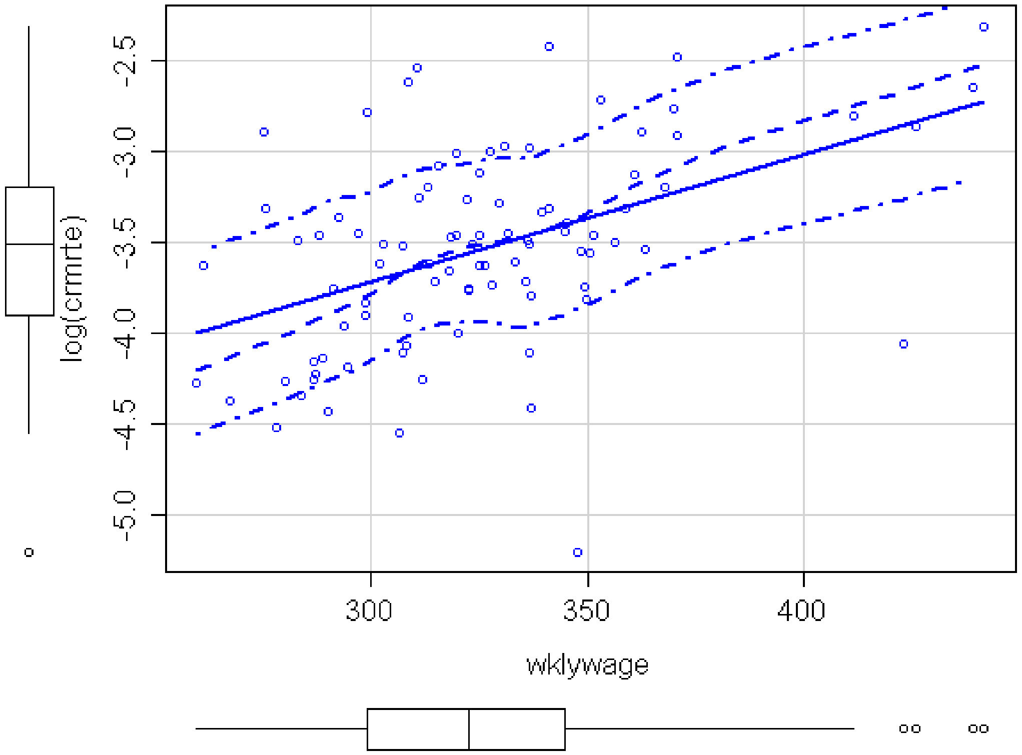
Optimal Model

Section 2.0: $\log(crmrte) - wklywages$ **and all 9 wage variables**

wklywages (average of all the wages) mostly linear, but only variable 'wfed' was significant at .05. 'wfed' is federal wages and mostly not in control of the state of of North Carolina. The rest of the wage variables individually and when the mean is taken together are all insignificant are not practical to be used. So all the wages will be eliminated from the model.

In [29]:

```
options(repr.plot.height = 4.0, repr.plot.width = 5.0, repr.plot.pointsize = 9.0)
scatterplot(log(crmrte) ~ wklywage, data=crime)
```



Section 2.1: wklywage vs wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc

Variable 'wfed' was significant at p<0.05 all other variables were insignificant. The state has little if any control of federal wages.

In [30]:

```
# Define models for Wage.  Is there a difference between the mean of all the
# wages vs. the wages individually?
model2_1 = lm(log(crmrte) ~ wklywage, data = crime)
model2_2 = lm(log(crmrte) ~ wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
              wsta + wloc, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))
se.model2_2 = sqrt(diag(vcovHC(model2_2)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1, model2_2,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1, se.model2_2),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

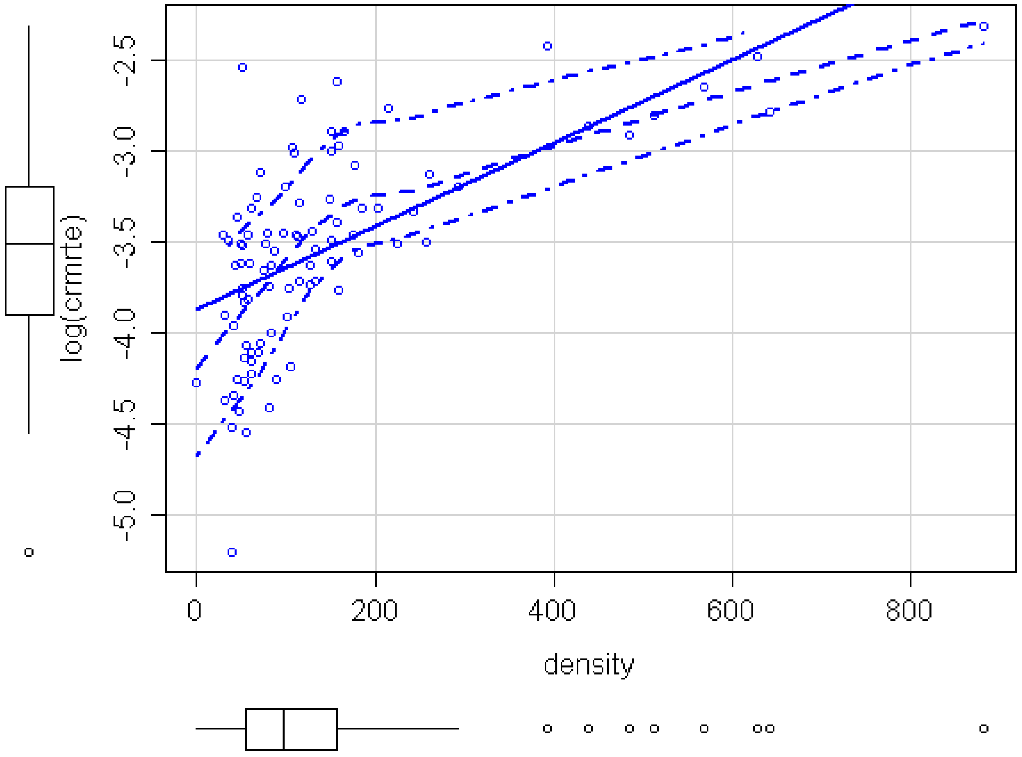
```
===========================================
                  Dependent variable:
             ------------------------------
                      log(crmrte)
                   (1)            (2)
-------------------------------------------
wklywage        0.007***
                 (0.002)

wcon                             0.002
                                (0.002)

wtuc                            -0.0003
                                (0.001)

wtrd                             0.003
                                (0.003)

wfir                             -0.002
                                (0.001)

wser                             0.001
                                (0.002)

wmfg                             0.001
                                (0.001)

wfed                             0.004*
                                (0.002)

wsta                             0.002
                                (0.001)

wloc                             -0.002
                                (0.005)

Constant        -5.828***      -6.089***
                 (0.519)        (0.858)

-------------------------------------------
Observations       90             90
Adjusted R2       0.209          0.258
===========================================
Note:         *p<0.05; **p<0.01; ***p<0.001
```

Section 2.2: $\log(crmrte) - density$

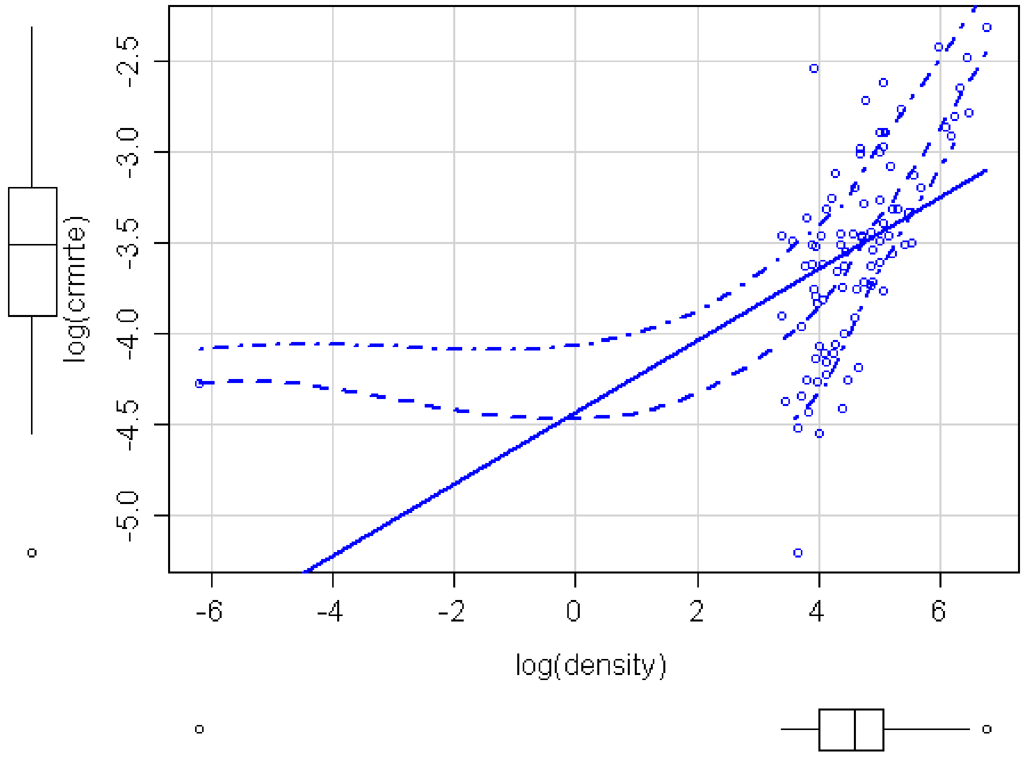
Density is a non-linear relationship. The log transformations made the data even more non-linear. The sqrt transformation on density improved the linear relationship significantly. The sqrt of density was significant at very small levels ($p < 0.001$)

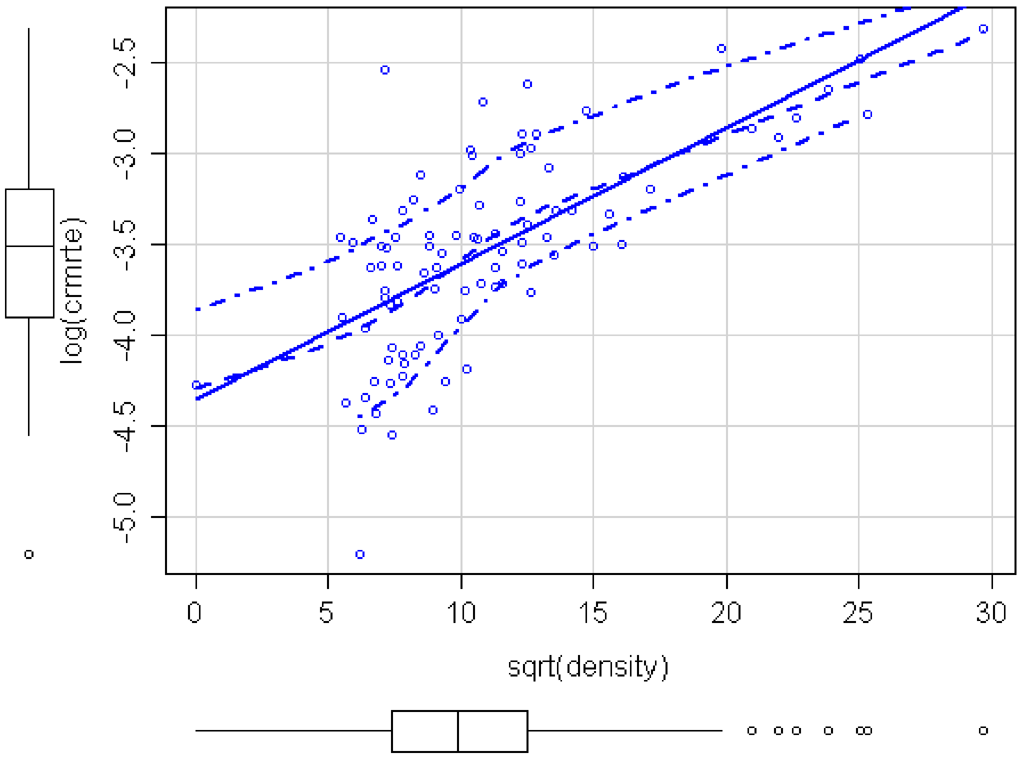
In [31]:

```
options(repr.plot.height = 4.0, repr.plot.width = 5.0, repr.plot.pointsize = 9.0)
scatterplot(log(crmrte) ~ (density), data=crime)
```



In [32]:
```
# Try to deal with non-linearity above, using Log and sqrt.
options(repr.plot.height = 4.0, repr.plot.width = 5.0, repr.plot.pointsize = 9.0)
scatterplot(log(crmrte) ~ log(density), data=crime)
scatterplot(log(crmrte) ~ sqrt(density), data=crime)
```




log(crmte)
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
0
5
10
15
20
25
30
sqrt(density)

In [33]:

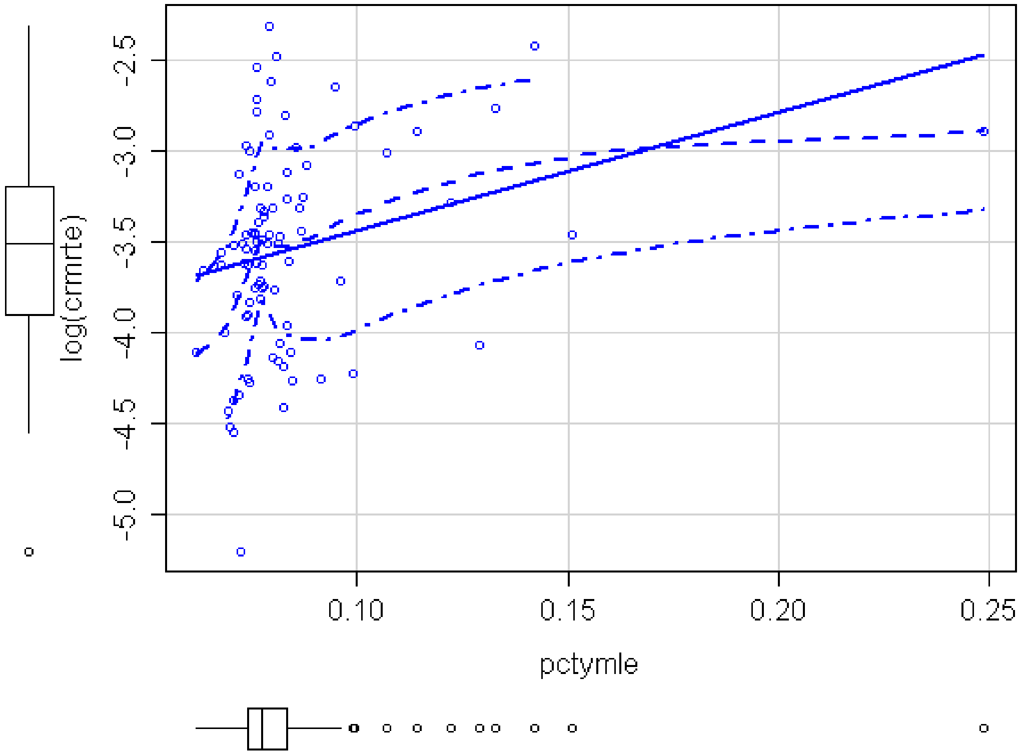
```r
# Define models for density.
model2_1 = lm(log(crmrte) ~ sqrt(density), data = crime)
model2_2 = lm(log(crmrte) ~ density, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))
se.model2_2 = sqrt(diag(vcovHC(model2_2)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1, model2_2,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1, se.model2_2),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

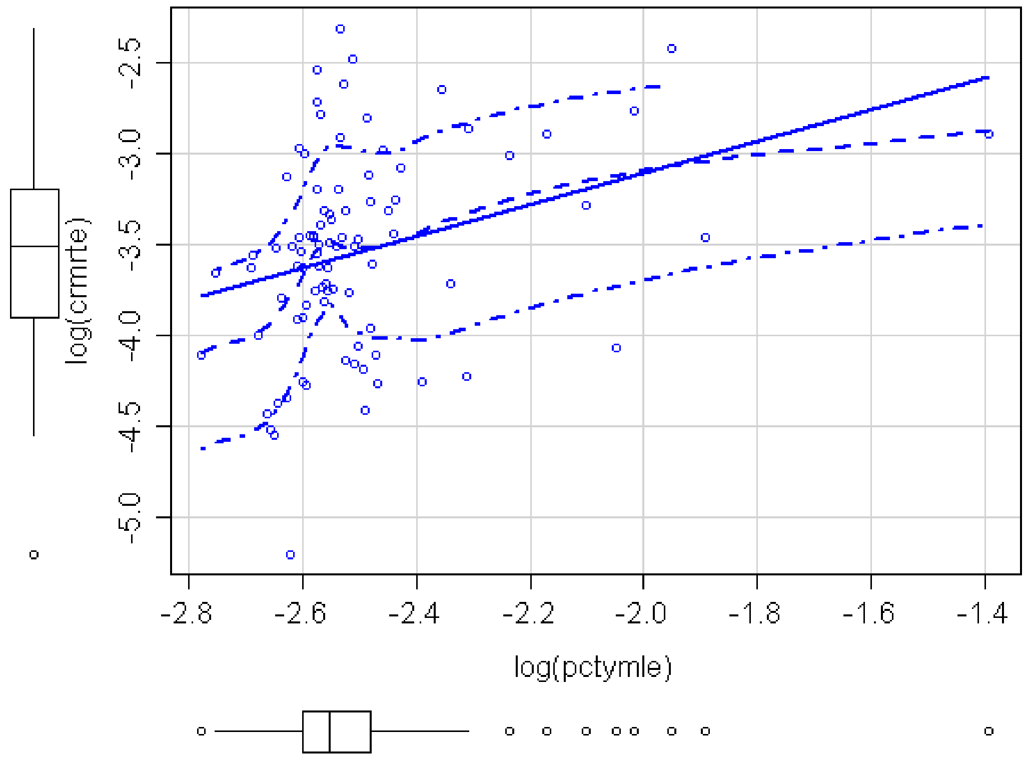
```
============================================
                   Dependent variable:
              ------------------------------
                       log(crmrte)
                    (1)            (2)
--------------------------------------------
sqrt(density)    0.074***
                  (0.007)

density                          0.002***
                                 (0.0003)

Constant         -4.351***      -3.869***
                  (0.100)        (0.069)

--------------------------------------------
Observations        90             90
Adjusted R2       0.457          0.394
============================================
Note:          *p<0.05; **p<0.01; ***p<0.001
```

Section 2.3: $\log(crmrte) - log(pctymle)$

The log transform looks similiar, though log(pctymle) has more of a uniform distribution. However, log(pctymle) is significant where p < 0.01, while pctymle is not significant.

In [34]:
```
options(repr.plot.height = 4.0, repr.plot.width = 5.0, repr.plot.pointsize = 9.0)
scatterplot(log(crmrte) ~ (pctymle), data=crime)
scatterplot(log(crmrte) ~ log(pctymle), data=crime)
```




-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
log(crmrte)
-2.8
-2.6
-2.4
-2.2
-2.0
-1.8
-1.6
-1.4
log(pctymle)

In [35]:

```r
# Define models for log(pctymle).
model2_1 = lm(log(crmrte) ~ log(pctymle), data = crime)
model2_2 = lm(log(crmrte) ~ pctymle, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))
se.model2_2 = sqrt(diag(vcovHC(model2_2)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1, model2_2,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1, se.model2_2),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

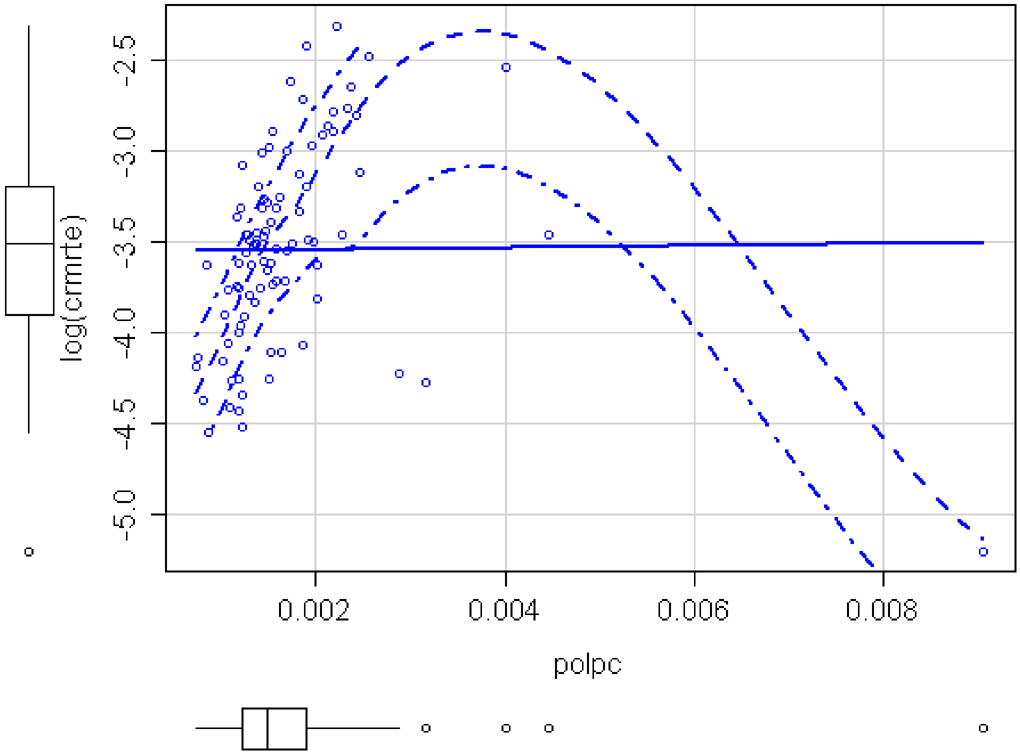
```
===========================================
                  Dependent variable:
             ------------------------------
                      log(crmrte)
                   (1)            (2)
-------------------------------------------
log(pctymle)     0.862**
                 (0.289)

pctymle                         6.509
                               (3.739)

Constant         -1.386        -4.089***
                 (0.729)        (0.311)

-------------------------------------------
Observations       90             90
Adjusted R2       0.087          0.067
===========================================
Note:         *p<0.05; **p<0.01; ***p<0.001
```

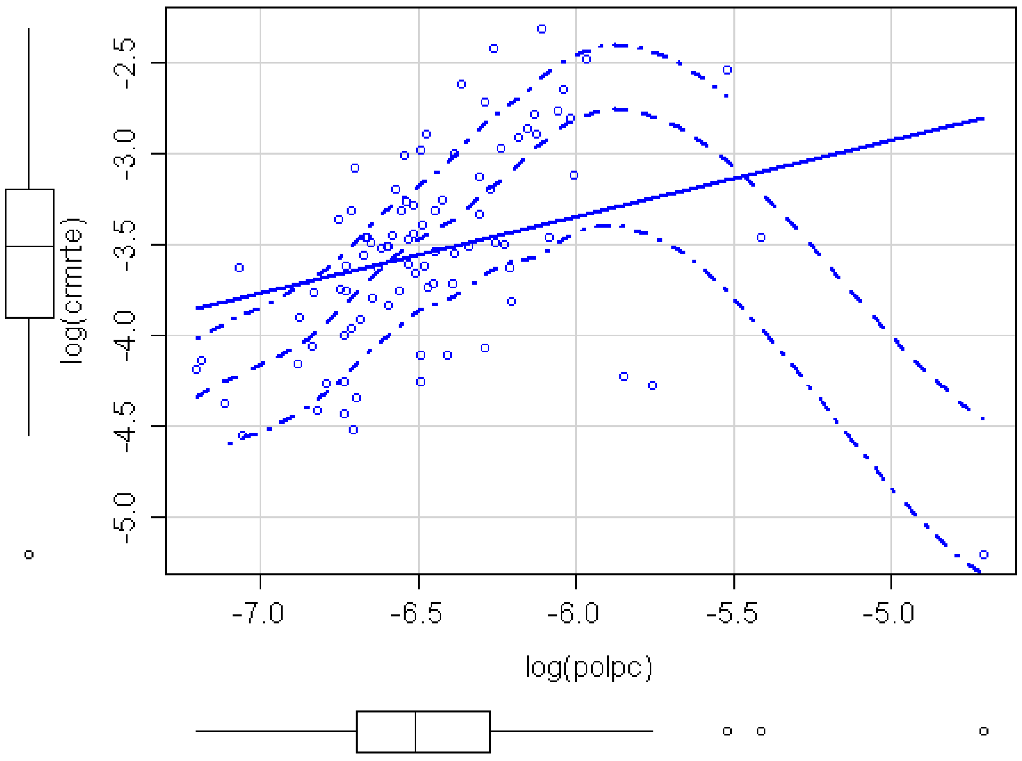
Section 2.4: $\log(crmrte) - polpc$

Polpc is not linear. log(polpc) is more uniformally distributed, has less skewness, and is better with respect to linearality where the majority of observations appear. An outlier was observed, which is part of the reason why the right-hand side of the graph is more skewed. Testing log(polpc) for signficance resulted in none.

The polpc variable was not significant and was unable to improve substantially the linearality. The variable was removed from model.

In [36]:
```
scatterplot(log(crmrte) ~ (polpc), data=crime)
scatterplot(log(crmrte) ~ log(polpc), data=crime)
```





-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
log(crmrte)
-7.0
-6.5
-6.0
-5.5
-5.0
log(polpc)

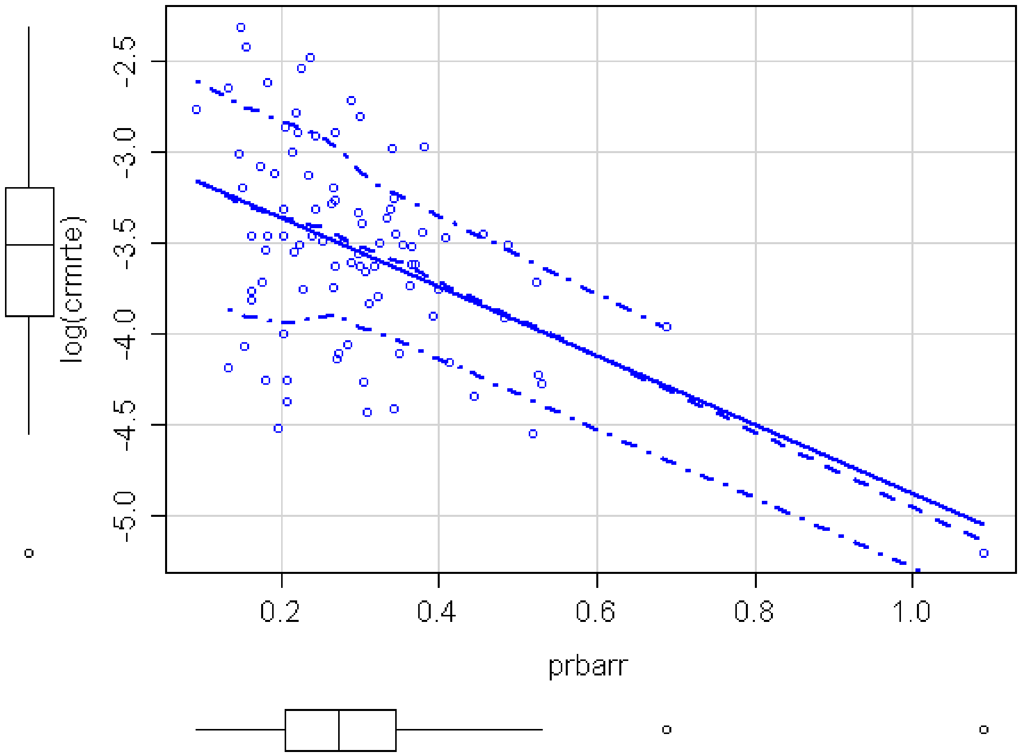
In [37]:
```
# Define models for log(polpc).
model2_1 = lm(log(crmrte) ~ log(polpc), data = crime)
model2_2 = lm(log(crmrte) ~ polpc, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))
se.model2_2 = sqrt(diag(vcovHC(model2_2)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1, model2_2,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1, se.model2_2),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
===========================================
                  Dependent variable:
             ------------------------------
                      log(crmrte)
                   (1)            (2)
-------------------------------------------
log(polpc)        0.417
                 (0.475)

polpc                            5.762
                               (386.802)

Constant         -0.847        -3.552***
                 (3.104)        (0.612)

-------------------------------------------
Observations       90             90
Adjusted R2       0.071         -0.011
===========================================
Note:         *p<0.05; **p<0.01; ***p<0.001
```

Section 2.5: $\log(crmrte) - prbarr$

The variable prbarr was found to be linear and significant where p < 0.001. prbarr will remain in the model.

In [38]:
```
scatterplot(log(crmrte) ~ (prbarr), data=crime)
```



In [39]:

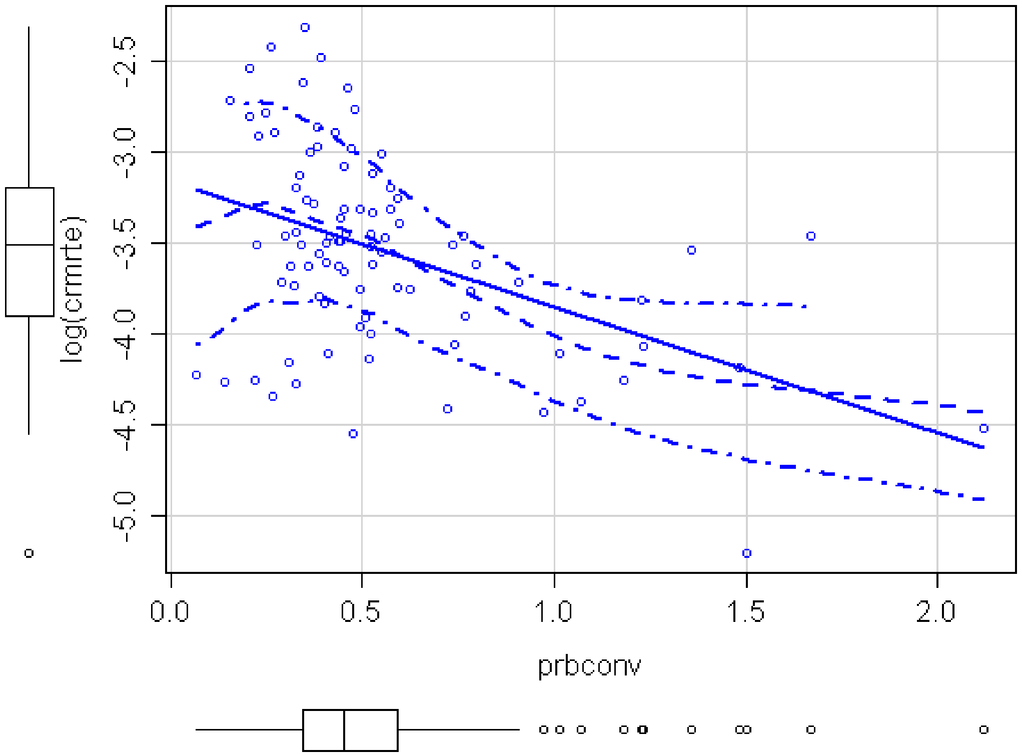
```
# Define models for prbarr.
model2_1 = lm(log(crmrte) ~ prbarr, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
===========================================
                  Dependent variable:
             ------------------------------
                      log(crmrte)
-------------------------------------------
prbarr                 -1.884***
                        (0.321)

Constant               -2.985***
                        (0.124)

-------------------------------------------
Observations              90
Adjusted R2             0.215
===========================================
Note:        *p<0.05; **p<0.01; ***p<0.001
```

Section 2.6: $\log(crmrte) - prbconv$

prbconv is pretty linear, so no transformations will be applied. Testing prbconv for signficance resulted in strong significance finding where p < 0.001.

```
In [40]: scatterplot(log(crmrte) ~ (prbconv), data=crime)
```



In [41]:

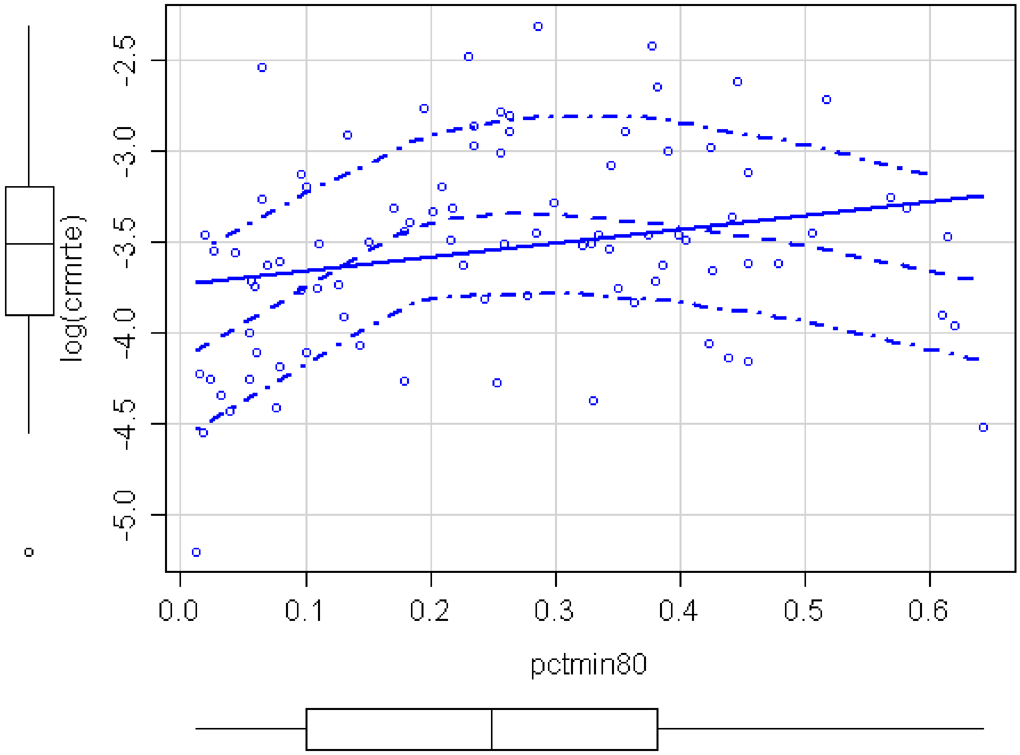
```
# Define models for prbconv.
model2_1 = lm(log(crmrte) ~ prbconv, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

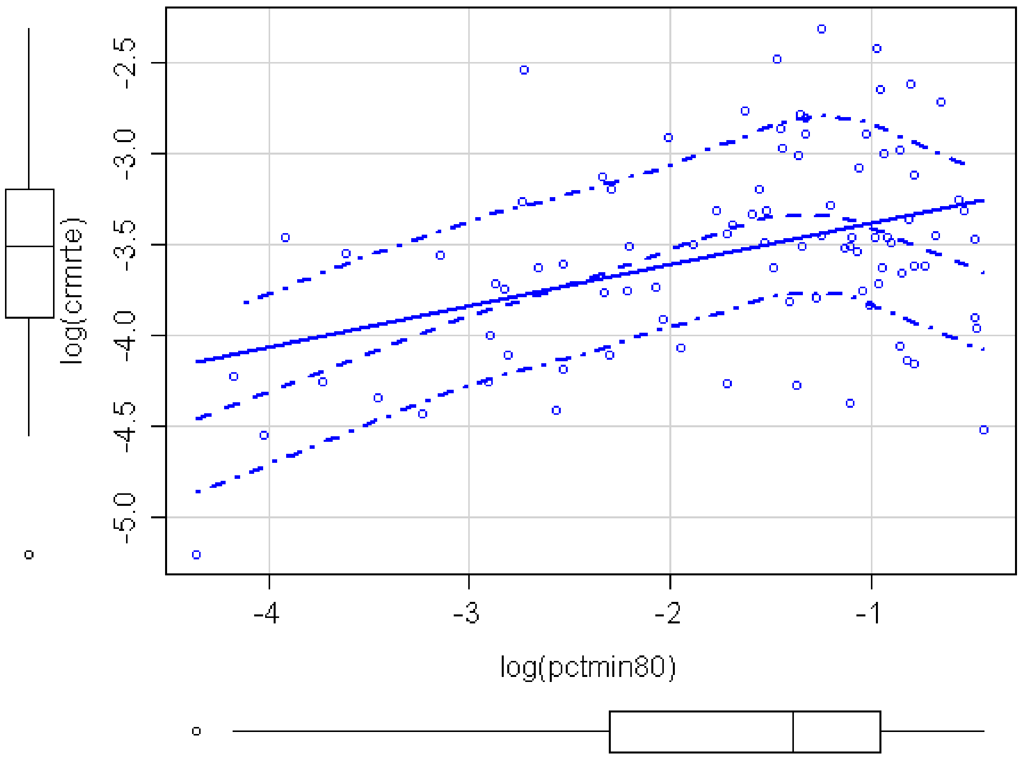
```
==========================================
                  Dependent variable:
             -----------------------------
                      log(crmrte)
------------------------------------------
prbconv                 -0.692***
                         (0.179)

Constant                -3.160***
                         (0.116)

------------------------------------------
Observations               90
Adjusted R2              0.191
==========================================
Note:        *p<0.05; **p<0.01; ***p<0.001
```

Section 2.7: $\log(crmrte) - pctmin80$

The variable pctmin80 is not linear, so a log transformation was applied which improved the linearity of the plot. The log(pctmin80) is significant where p < 0.001.

```
In [42]: scatterplot(log(crmrte) ~ (pctmin80), data=crime)
         scatterplot(log(crmrte) ~ log(pctmin80), data=crime)
```




log(crmte)
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-4
-3
-2
-1
log(pctmin80)

In [43]:

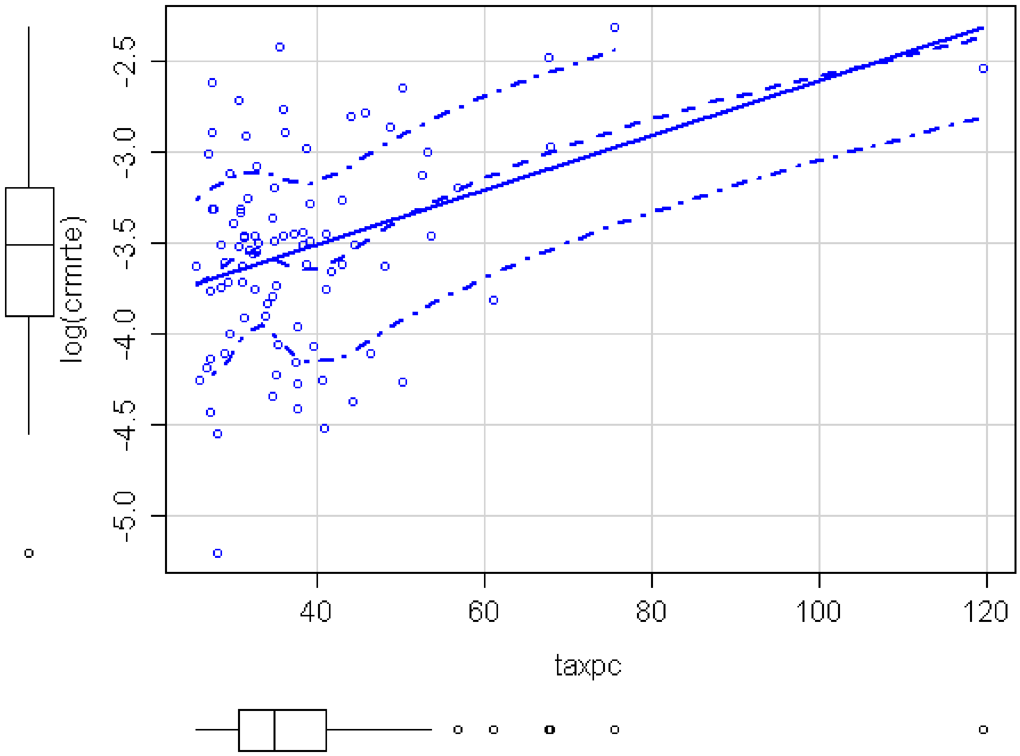
```
# Define models for log(pctmin80).
model2_1 = lm(log(crmrte) ~ log(pctmin80), data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
===========================================
                   Dependent variable:
              -----------------------------
                       log(crmrte)
-------------------------------------------
log(pctmin80)            0.227***
                          (0.065)

Constant                 -3.157***
                          (0.122)

-------------------------------------------
Observations                90
Adjusted R2                0.148
===========================================
Note:         *p<0.05; **p<0.01; ***p<0.001
```

Section 2.8: $\log(crmrte) - taxpc$

The variable taxpc is linear and significant where p < 0.001.

In [44]:
```
scatterplot(log(crmrte) ~ (taxpc), data=crime)
```



In [45]:

```
# Define models for taxpc.
model2_1 = lm(log(crmrte) ~ taxpc, data = crime)

# Compute robust standard errors
se.model2_1 = sqrt(diag(vcovHC(model2_1)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_1,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
==========================================
                  Dependent variable:
             -----------------------------
                      log(crmrte)
------------------------------------------
taxpc                  0.015***
                        (0.004)

Constant               -4.114***
                        (0.165)

------------------------------------------
Observations              90
Adjusted R2              0.118
==========================================
Note:        *p<0.05; **p<0.01; ***p<0.001
```

Section 2.9: Building Model 2

In [46]:

```
model2_1 = lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80)
              + taxpc, data = crime)

se.model2_1 = sqrt(diag(vcovHC(model2_1)))

stargazer(model2_1,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
===========================================
                   Dependent variable:
              -----------------------------
                       log(crmrte)
-------------------------------------------
density                  0.001**
                        (0.0004)

log(pctymle)             0.400**
                         (0.134)

prbarr                   -1.118
                         (0.632)

prbconv                 -0.508**
                         (0.175)

log(pctmin80)           0.183***
                         (0.047)

taxpc                     0.006
                         (0.008)

Constant                -2.058***
                         (0.350)

-------------------------------------------
Observations               90
Adjusted R2               0.729
===========================================
Note:         *p<0.05; **p<0.01; ***p<0.001
```

In [47]:

```
coeftest(model2_1, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
```

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   -2.05806778  0.35034024 -5.8745 8.498e-08 ***
density        0.00137764  0.00044321  3.1083 0.0025771 **
log(pctymle)   0.39983828  0.13366453  2.9914 0.0036558 **
prbarr        -1.11818614  0.63247311 -1.7680 0.0807427 .
prbconv       -0.50824503  0.17533131 -2.8988 0.0047914 **
log(pctmin80)  0.18336927  0.04652371  3.9414 0.0001683 ***
taxpc          0.00625563  0.00849493  0.7364 0.4635663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 2.10: Model 2 formula

$$logcrmrte = -2.0581 + 0.0014(density) + 0.3998(logpctymle) - 1.1182(prbarr) - 0.5082(prbconv)$$

$$+ 0.0063(taxpc)$$

$$n = 90 \qquad Adj.\,R^2 = 72.9$$

Section 2.11: Model 2 Interpretation

In [48]:

```
# Table of Coefficient
model2_ind_vars <- c("density", "logpctymle", "prbarr", "prbconv",
                     "logpctmin80", "taxpc")
model2_formula <- as.formula(paste("log(crmrte) ~ ", paste(model2_ind_vars,
                                                            collapse = " + "), sep = ""))
model2_table <- lm(model2_formula, data = crime)

interpret_model2 <- c(
"",
"If density increased by 1 unit, we'd expect crmrte to increase by .14%,
while all other variables remain constant.",

"For each percentage increase in pctymle, crmrte will increase by .40%,
while all other variables remain constant.",

"If prbarr (the ratio of convictions to arrests) increases by 1 unit,
we'd expect crmrte to decrease by 111.88%, while all other variables remain constant..",

"If prbconv (the ratio of arrests to offenses) increase by 1 unit, we'd expect
crmrte to decrease by 50.82%, while all other variables remain constant.",

"For each percentage increase in pctmin80, the crime rate will increase by .18%,
while all other variables remain constant.",

"If taxpc (taxes per capita) increase by 1 unit, we'd expect crmrte to increase by .63%,
while all other variables remain constant."
)

coef2_1 <- data.frame("Coef." = round(model2_1$coefficients, 4),
                      "Interpretation" = interpret_model2)

kable(coef2_1, booktabs = TRUE) %>%
kable_styling(font_size = 10, full_width = FALSE) %>%
column_spec(3, width = "35em")
```

```
<table class="table" style="font-size: 10px; width: auto !important; margin-left: auto; m
argin-right: auto;">
 <thead>
  <tr>
   <th style="text-align:left;">   </th>
   <th style="text-align:right;"> Coef. </th>
   <th style="text-align:left;"> Interpretation </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;"> (Intercept) </td>
   <td style="text-align:right;"> -2.0581 </td>
   <td style="text-align:left;width: 35em; ">  </td>
  </tr>
  <tr>
   <td style="text-align:left;"> density </td>
   <td style="text-align:right;"> 0.0014 </td>
   <td style="text-align:left;width: 35em; "> If density increased by 1 unit, we'd expect
crmrte to increase by .14%,
while all other variables remain constant. </td>
  </tr>
  <tr>
   <td style="text-align:left;"> log(pctymle) </td>
   <td style="text-align:right;"> 0.3998 </td>
   <td style="text-align:left;width: 35em; "> For each percentage increase in pctymle, cr
mrte will increase by .40%,
while all other variables remain constant. </td>
  </tr>
  <tr>
   <td style="text-align:left;"> prbarr </td>
   <td style="text-align:right;"> -1.1182 </td>
   <td style="text-align:left;width: 35em; "> If prbarr (the ratio of convictions to arre
sts) increases by 1 unit,
we'd expect crmrte to decrease by 111.88%, while all other variables remain constant.. </
td>
  </tr>
  <tr>
   <td style="text-align:left;"> prbconv </td>
   <td style="text-align:right;"> -0.5082 </td>
   <td style="text-align:left;width: 35em; "> If prbconv (the ratio of arrests to offense
s) increase by 1 unit, we'd expect
crmrte to decrease by 50.82%, while all other variables remain constant. </td>
  </tr>
  <tr>
   <td style="text-align:left;"> log(pctmin80) </td>
   <td style="text-align:right;"> 0.1834 </td>
   <td style="text-align:left;width: 35em; "> For each percentage increase in pctmin80, t
he crime rate will increase by .18%,
while all other variables remain constant. </td>
  </tr>
  <tr>
   <td style="text-align:left;"> taxpc </td>
   <td style="text-align:right;"> 0.0063 </td>
   <td style="text-align:left;width: 35em; "> If taxpc (taxes per capita) increase by 1 u
nit, we'd expect crmrte to increase by .63%,
while all other variables remain constant. </td>
  </tr>
</tbody>
</table>
```

	Coef.	Interpretation
(Intercept)	-2.0581	
density	0.0014	If density increased by 1 unit, we'd expect crmrte to increase by .14%, while all other variables remain constant.
log(pctymle)	0.3998	For each percentage increase in pctymle, crmrte will increase by .40%, while all other variables remain constant.
prbarr	-1.1182	If prbarr (the ratio of convictions to arrests) increases by 1 unit, we'd expect crmrte to decrease by 111.88%, while all other variables remain constant..
prbconv	-0.5082	If prbconv (the ratio of arrests to offenses) increase by 1 unit, we'd expect crmrte to decrease by 50.82%, while all other variables remain constant.
log(pctmin80)	0.1834	For each percentage increase in pctmin80, the crime rate will increase by .18%, while all other variables remain constant.
taxpc	0.0063	If taxpc (taxes per capita) increase by 1 unit, we'd expect crmrte to increase by .63%, while all other variables remain constant.

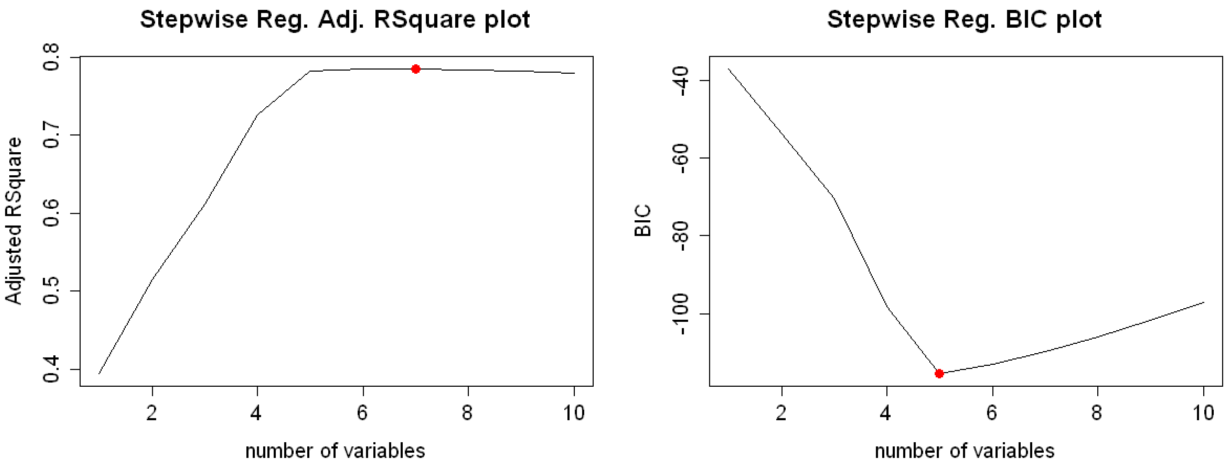
Optimal Model STEP Analysis

Model 2.12: Step-wise Regression Model

In [49]:
```
model2_df <- crime %>% select(logcrmrte, density, logpctymle, prbarr, prbconv,
                              logpctmin80, taxpc, prbpris, avgsen, polpc, west, urban)
```

```
In [50]: # Model 2 using stepwise algorithm
         stepmodel <- regsubsets(logcrmrte ~ ., model2_df, nvmax = 10)
         stepsummary <- summary(stepmodel)
```

```
In [51]: # Find the highest Adjusted RSquare and minimum BIC
         options(repr.plot.height = 4.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
         par(mfrow = c(1, 2))
         plot(stepsummary$adjr2, ylab = "Adjusted RSquare", xlab = "number of variables",
                  main = "Stepwise Reg. Adj. RSquare plot", type = "l")
         points(which.max(stepsummary$adjr2), stepsummary$adjr2[which.max(stepsummary$adjr2)],
                col = "red", cex = 2, pch = 20)

         plot(stepsummary$bic, ylab = "BIC", xlab = "number of variables",
              main = "Stepwise Reg. BIC plot", type = "l")
         points(which.min(stepsummary$bic), stepsummary$bic[which.min(stepsummary$bic)],
                col = "red", cex = 2, pch = 20)
```



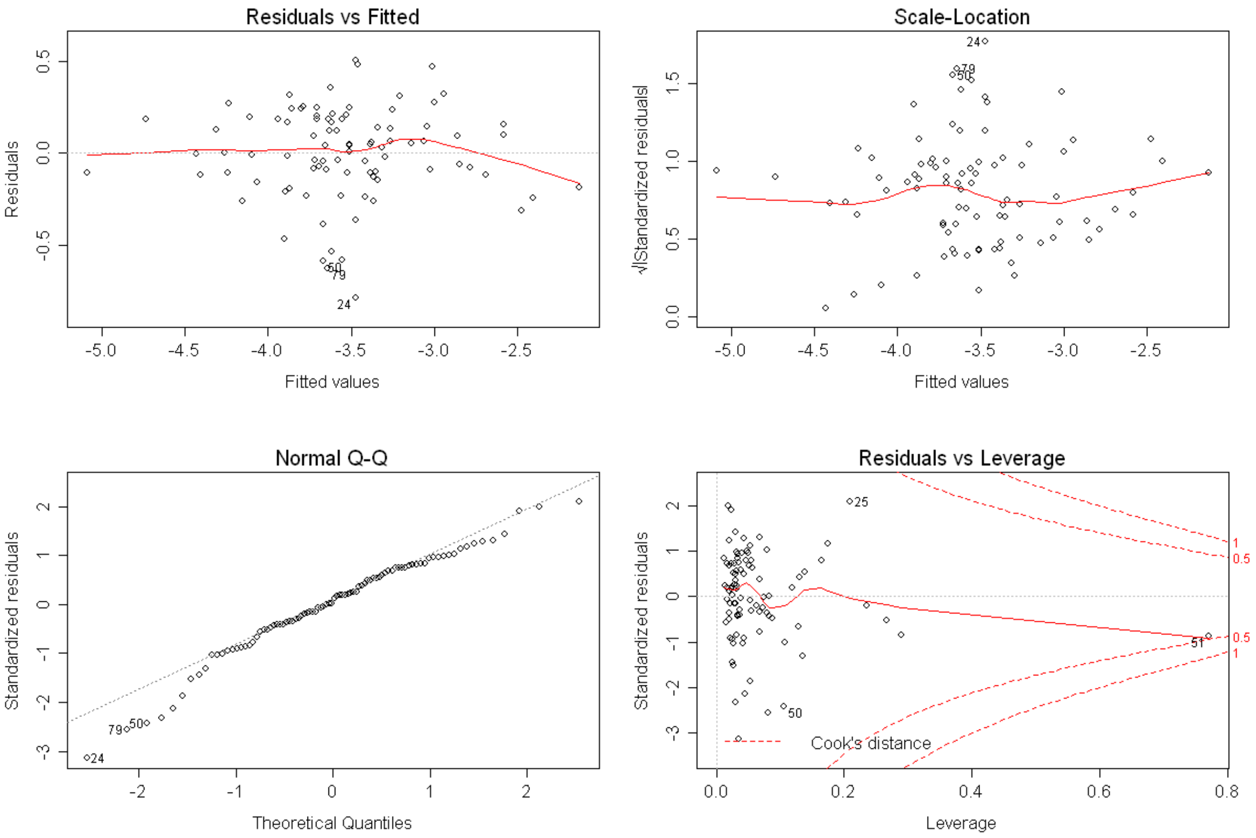
Optimal Model Assumptions

```
In [52]: # Optimal # of variables is 5
         stepsummary$outmat[1:5, ]
```

	density	logpctymle	prbarr	prbconv	logpctmin80	taxpc	prbpris	avgsen	polpc	west	urban
1 (1)	*										
2 (1)	*				*						
3 (1)	*			*	*						
4 (1)			*	*	*				*		
5 (1)	*		*	*	*				*		

In [53]:

```
# 5 variables are determined: sqrtdensity, prbconv, prbarr, logpctmin80, and polpc
model2_step <- lm(formula = logcrmrte ~ density + prbarr + prbconv + logpctmin80 +
                  polpc, data = crime)
options(repr.plot.height = 7.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(model2_step)
```



The Models are Linear Models based on Ordinary Least Squares (OLS) Regression. The validity of the model assumes non-violation of the following assumptions:

CLM 1.0: Linearity of Parameters:
The model is linear in parameters: log of crime rate is a linear function. As observed on the diagnostic regression plot of residuals against fitted values, a random pattern indicates an appropriate model has been fit to the data.

CLM 2.0: Random Sampling and Independence:
The data does not appear to be random. 10 counties are missing from the dataset with no explanation. Those counties notably include some of the most sparsely populated counties, each having a population of less that 68,000 people and are not densely populated. The average population per county in North Carolina is just over 105,068. Additionally, our data is from two different databases. One is the North Carolina Department of Correction

and the other is the FBI's Uniform Crime Reports. Each of these purports that it is comprehensive, but we don't necessarily have enough information to be sure that combining these datasets creates a representative sample, or if it biases the data towards crimes more likely to be investigated by the FBI to include many felonies, kidnappings, and cross-state line offenses that wind up or start in North Carolina. It is reasonable to assume independence.

CLM 3.0 Zero Conditional Mean:
The expected value of residuals should is approximately zero. (Reference Section 3.14)

CLM 4.0': Normality of residuals:
The qqPlot suggests slight deviation from normality of residuals for extreme values across the dataset. Specifically, the model may not be predictive for counties (23, 25, 82). E

CLM 5.0 Homoscedasticity:
The scale-location plot of the Linear Model diagnostic plots graphically evaluates the assumption that errors are identically- distributed and have homogeneous variance. The plot shows the absolute residuals against the fitted values, along with a smooth red line. Significant departures from a horizontal line typically suggest heteroscedasticity. The Model plot shows approximately horizontal line except for the right hand side with observation 51, which implies homogeneous variance for the most part except for the right hand extreme. Quantitatively, assumption of constant error variance can be tested using the ncvTest function. With a p-value of 0.2265, we fail to reject the null hypothesis (that variance of residuals is constant) and therefore infer that the residuals are homoscedastic. (Reference Section 3.14)

CLM 6.0 Multicollinearity:
The Variance Inflation Factor (VIF) can detect and represent collinearity. If any of the variance inflation factors exceed 5 (the cut-off typically used), the regression coefficients are poorly estimated due to multicollinearity. All of the variables in the model are below 1.4. This suggests that the Multicollinearity assumption is not violated. (Reference Section 3.14)

Section 2:13: Model 2 Step-wise Formula

```
In [54]: # Compute robust standard errors
se.model2_step = sqrt(diag(vcovHC(model2_step)))

# We pass the standard errors into stargazer through the se argument.
stargazer(model2_step,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_step),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
==========================================
                 Dependent variable:
             -----------------------------
                      logcrmrte
------------------------------------------
density                0.001***
                       (0.0003)

prbarr                 -2.044***
                        (0.315)

prbconv                -0.729***
                        (0.100)

logpctmin80            0.226***
                        (0.031)

polpc                 201.695***
                       (50.892)

Constant               -2.646***
                        (0.151)

------------------------------------------
Observations               90
Adjusted R2              0.782
==========================================
Note:        *p<0.05; **p<0.01; ***p<0.001
```

```
In [55]: coeftest(model2_step, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
```

```
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -2.6461e+00  1.5109e-01 -17.5128 < 2.2e-16 ***
density      1.0301e-03  2.5536e-04   4.0340 0.0001203 ***
prbarr      -2.0444e+00  3.1527e-01  -6.4845 5.814e-09 ***
prbconv     -7.2937e-01  9.9686e-02  -7.3167 1.388e-10 ***
logpctmin80  2.2618e-01  3.0755e-02   7.3543 1.169e-10 ***
polpc        2.0169e+02  5.0892e+01   3.9632 0.0001546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2 STEP formula

$$\log(crmrte) = -2.6461 + 0.00103(density) - 2.0444(prbarr) - 0.7294(prbconv) + 0.2262(logpctmin$$

$$n = 90 \qquad Adj.\,R^2 = 78.2\%$$

In [56]:

```
# Model2 Stepwise Table of Coefficient
model2_step_ind_vars <- c("density", "prbarr", "prbconv", "logpctmin80", "polpc")
model2_step_formula <- as.formula(paste("log(crmrte) ~ ", paste(model2_step_ind_vars,
                                                                collapse = " + "), sep = ""))
model2_step_table <- lm(model2_step_formula, data = crime)

interpret_model2_step <- c(
"",
"If density increased by 1 unit, we'd expect crmrte to increase by .10%,
while all other variables remain constant.",

"If prbarr (the ratio of arrests to offenses) increases by 1 unit,
we'd expect crmrte to decrease by 204%, while all other variables remain constant.",

"If prbconv (the ratio of convictions to arrests) increase by 1 unit,
we'd expect crmrte to decrease by 72%, while all other variables remain constant.",

"For each percentage increase in pctmin80, the crime rate will increase by .22%,
while all other variables remain constant.",

"If polpc (police per capita) increaseds by 1 unit,
we'd expect crmrte to increase by 20,169.5%, while all other variables remain constant.")

coef2_step <- data.frame("Coef." = round(model2_step$coefficients, 4),
                         "Interpretation" = interpret_model2_step)

kable(coef2_step, booktabs = TRUE) %>%
kable_styling(font_size = 10, full_width = FALSE) %>%
column_spec(3, width = "35em")
```

```
<table class="table" style="font-size: 10px; width: auto !important; margin-left: auto; m
argin-right: auto;">
 <thead>
  <tr>
   <th style="text-align:left;">   </th>
   <th style="text-align:right;"> Coef. </th>
   <th style="text-align:left;"> Interpretation </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;"> (Intercept) </td>
   <td style="text-align:right;"> -2.6461 </td>
   <td style="text-align:left;width: 35em; ">  </td>
  </tr>
  <tr>
   <td style="text-align:left;"> density </td>
   <td style="text-align:right;"> 0.0010 </td>
   <td style="text-align:left;width: 35em; "> If density increased by 1 unit, we'd expect
crmrte to increase by .10%
```

	Coef.	Interpretation
(Intercept)	-2.6461	
density	0.0010	If density increased by 1 unit, we'd expect crmrte to increase by .10%, while all other variables remain constant.
prbarr	-2.0444	If prbarr (the ratio of arrests to offenses) increases by 1 unit, we'd expect crmrte to decrease by 204%, while all other variables remain constant.
prbconv	-0.7294	If prbconv (the ratio of convictions to arrests) increase by 1 unit, we'd expect crmrte to decrease by 72%, while all other variables remain constant.

	Coef.	Interpretation
logpctmin80	0.2262	For each percentage increase in pctmin80, the crime rate will increase by .22%, while all other variables remain constant.
polpc	201.6948	If polpc (police per capita) increaseds by 1 unit, we'd expect crmrte to increase by 20,169.5%, while all other variables remain constant.

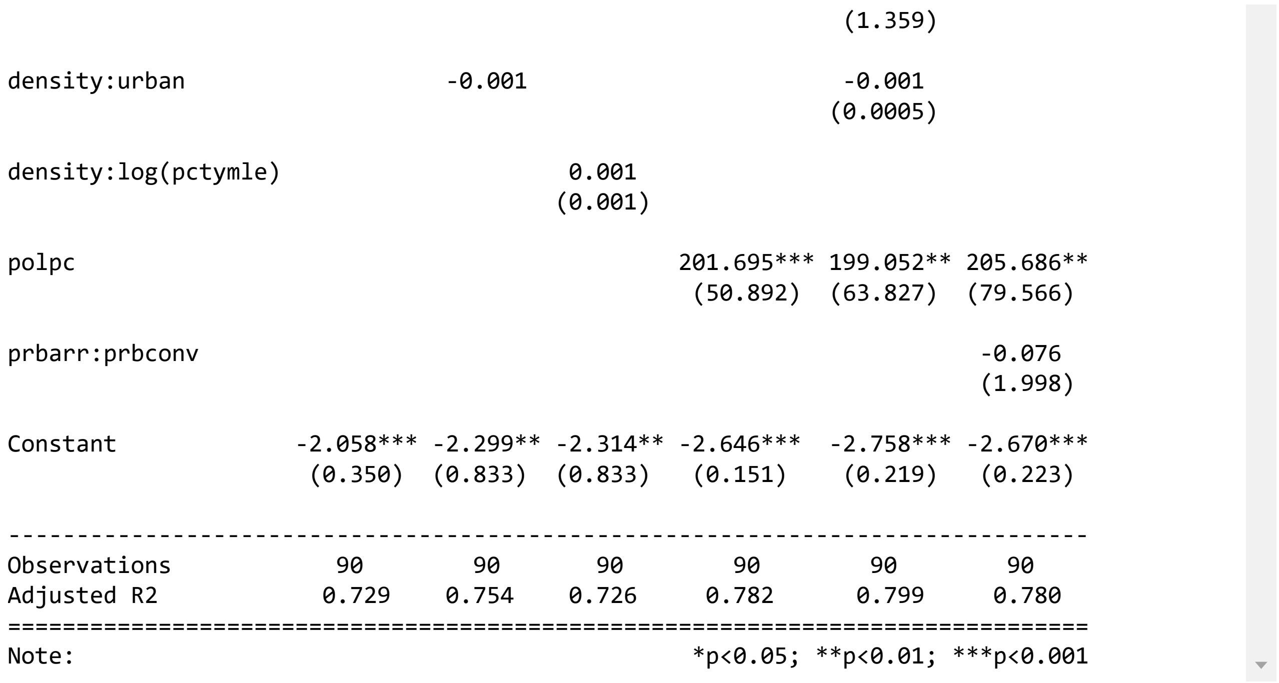
Examination of Interactions Using Optimal Model

Section 2.135: Model2 (original, step-wise, interaction)

In [57]:

```
##### Model 2 base model
model2_1 = lm(log(crmrte) ~ density + prbarr + prbconv + log(pctmin80) + taxpc +
              log(pctymle), data = crime)

# add to Model 2 base --> Interactions pctmin80:west & density:urban
model2_2 = lm(log(crmrte) ~ density + prbarr + prbconv + log(pctmin80) + taxpc +
              log(pctymle) + pctmin80:west + density:urban, data = crime)

# add to Model 2 base --> Interactions density*log(pctymle)
model2_3 = lm(log(crmrte) ~ density + prbarr + prbconv + log(pctmin80) + taxpc +
              log(pctymle) + density*log(pctymle), data = crime)

# Step-wise
model2_4 = lm(log(crmrte) ~ density + prbarr + prbconv + log(pctmin80) + polpc,
              data = crime)

# add to Model 2 step-wise --> Interactions pctmin80:west & density:urban
model2_5 = lm(log(crmrte) ~ density + prbarr + prbconv + log(pctmin80) + polpc +
              pctmin80:west + density:urban, data = crime)

# add to Model 2 step-wise --> Interactions prbconv*prbarr
model2_6 = lm(log(crmrte) ~ density + prbarr + prbconv + log(pctmin80) + polpc +
              prbconv*prbarr , data = crime)

se.model2_1 = sqrt(diag(vcovHC(model2_1)))
se.model2_2 = sqrt(diag(vcovHC(model2_2)))
se.model2_3 = sqrt(diag(vcovHC(model2_3)))
se.model2_4 = sqrt(diag(vcovHC(model2_4)))
se.model2_5 = sqrt(diag(vcovHC(model2_5)))
se.model2_6 = sqrt(diag(vcovHC(model2_6)))

stargazer(model2_1, model2_2, model2_3, model2_4, model2_5, model2_6,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1, se.model2_3, se.model2_3, se.model2_4, se.model2_5,
                    se.model2_6),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```

===============================================================================
                                        Dependent variable:
                     ----------------------------------------------------------
                                            log(crmrte)
                        (1)      (2)      (3)       (4)        (5)       (6)
-------------------------------------------------------------------------------
density               0.001**   0.003    0.003    0.001***   0.002**  0.001***
                     (0.0004)  (0.004)  (0.004)  (0.0003)   (0.001)  (0.0003)

prbarr                -1.118    -0.988   -1.120  -2.044***  -1.935***  -1.989*
                      (0.632)  (0.629)  (0.629)   (0.315)    (0.375)   (0.901)

prbconv              -0.508**  -0.488** -0.510** -0.729***  -0.713***  -0.708
                      (0.175)  (0.175)  (0.175)   (0.100)    (0.107)   (0.409)

log(pctmin80)        0.183***  0.184*** 0.181***  0.226***  0.221***  0.226***
                      (0.047)  (0.048)  (0.048)   (0.031)    (0.039)   (0.033)

taxpc                  0.006    0.008    0.006
                      (0.008)  (0.009)  (0.009)

log(pctymle)          0.400**   0.384    0.300
                      (0.134)  (0.263)  (0.263)

pctmin80:west                   -0.896                       -1.056
```



```
                                                             (1.359)

density:urban                   -0.001                       -0.001
                                                            (0.0005)

density:log(pctymle)                     0.001
                                        (0.001)

polpc                                            201.695*** 199.052** 205.686**
                                                  (50.892)  (63.827)  (79.566)

prbarr:prbconv                                                         -0.076
                                                                       (1.998)

Constant             -2.058*** -2.299** -2.314** -2.646***  -2.758*** -2.670***
                      (0.350)  (0.833)  (0.833)   (0.151)    (0.219)   (0.223)

-------------------------------------------------------------------------------
Observations            90        90       90        90         90        90
Adjusted R2            0.729    0.754    0.726     0.782      0.799     0.780
===============================================================================
Note:                                             *p<0.05; **p<0.01; ***p<0.001
```

In [58]:

```
coeftest(model2_1, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
coeftest(model2_2, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
coeftest(model2_3, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
coeftest(model2_4, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
coeftest(model2_5, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
coeftest(model2_6, vcov=vcovHC, star.cutoffs=c(0.05, 0.01, 0.001))
```

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.05806778  0.35034024 -5.8745 8.498e-08 ***
density        0.00137764  0.00044321  3.1083 0.0025771 ** 
prbarr        -1.11818614  0.63247311 -1.7680 0.0807427 .  
prbconv       -0.50824503  0.17533131 -2.8988 0.0047914 ** 
log(pctmin80)  0.18336927  0.04652371  3.9414 0.0001683 ***
taxpc          0.00625563  0.00849493  0.7364 0.4635663    
log(pctymle)   0.39983828  0.13366453  2.9914 0.0036558 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.29909949  0.37299442 -6.1639 2.622e-08 ***
density        0.00253902  0.00062511  4.0617 0.0001119 ***
prbarr        -0.98821835  0.59863997 -1.6508 0.1026574    
prbconv       -0.48805497  0.15973144 -3.0555 0.0030420 ** 
log(pctmin80)  0.18367862  0.04731329  3.8822 0.0002101 ***
taxpc          0.00785373  0.00766207  1.0250 0.3084083    
log(pctymle)   0.38444582  0.12633494  3.0431 0.0031567 ** 
pctmin80:west -0.89597912  0.82352319 -1.0880 0.2798294    
density:urban -0.00122816  0.00051120 -2.4025 0.0185695 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


t test of coefficients:

                        Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)          -2.31437267  0.83328447 -2.7774 0.0067890 ** 
density               0.00293264  0.00363442  0.8069 0.4220534    
prbarr               -1.12012013  0.62872479 -1.7816 0.0785211 .  
prbconv              -0.50964927  0.17509034 -2.9108 0.0046409 ** 
log(pctmin80)         0.18138610  0.04847860  3.7416 0.0003379 ***
taxpc                 0.00629892  0.00852360  0.7390 0.4620190    
log(pctymle)          0.29974223  0.26272710  1.1409 0.2572376    
density:log(pctymle)  0.00062309  0.00148520  0.4195 0.6759268    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


t test of coefficients:

                 Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept)   -2.6461e+00  1.5109e-01 -17.5128 < 2.2e-16 ***
density        1.0301e-03  2.5536e-04   4.0340 0.0001203 ***
prbarr        -2.0444e+00  3.1527e-01  -6.4845 5.814e-09 ***
prbconv       -7.2937e-01  9.9686e-02  -7.3167 1.388e-10 ***
log(pctmin80)  2.2618e-01  3.0755e-02   7.3543 1.169e-10 ***
polpc          2.0169e+02  5.0892e+01   3.9632 0.0001546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t test of coefficients:

                  Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept)    -2.7582e+00  2.1887e-01 -12.6018 < 2.2e-16 ***
density         1.9265e-03  6.3993e-04   3.0104  0.003466 ** 
prbarr         -1.9345e+00  3.7509e-01  -5.1576 1.705e-06 ***
prbconv        -7.1298e-01  1.0720e-01  -6.6511 3.019e-09 ***
log(pctmin80)   2.2137e-01  3.8801e-02   5.7054 1.787e-07 ***
polpc           1.9905e+02  6.3827e+01   3.1186  0.002507 ** 
pctmin80:west -1.0559e+00  1.3592e+00  -0.7769  0.439477    
density:urban -9.1291e-04  4.8591e-04  -1.8788  0.063831 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t test of coefficients:

                  Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept)    -2.6696e+00  2.2312e-01 -11.9651 < 2.2e-16 ***
density         1.0314e-03  2.5908e-04   3.9809 0.0001464 ***
prbarr         -1.9887e+00  9.0073e-01  -2.2079 0.0300101 *  
prbconv        -7.0848e-01  4.0941e-01  -1.7305 0.0872579 .  
log(pctmin80)   2.2569e-01  3.3494e-02   6.7383 1.966e-09 ***
polpc           2.0569e+02  7.9566e+01   2.5851 0.0114833 *  
prbarr:prbconv -7.6313e-02  1.9979e+00  -0.0382 0.9696227    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 2.14: Model 2 Assumptions

Calculations were performed for CL 3.0, 5.0, and 6.0, which are included below.

In [59]:
```
#CLM 3.0: Zero Conditional Mean
round(mean(model2_1$residuals), 4)
round(mean(model2_step$residuals), 4)
```

0

0

In [60]:
```
#CLM 5.0 Homoscedasticity
ncvTest(model2_1)
ncvTest(model2_step)
```

```
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 0.06929974    Df = 1     p = 0.7923591 

Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 0.0329796    Df = 1     p = 0.8558944 
```

```
In [61]: #CLM 6.0 Multicollinearity
         vif(model2_1)
         vif(model2_step)
```

```
        density   1.28187884639618
         prbarr   1.19288556111998
        prbconv   1.12142945677297
 log(pctmin80)   1.02179920715549
          taxpc   1.15065436606088
  log(pctymle)   1.11959222384417

        density   1.44464565708736
         prbarr   1.61113888556707
        prbconv   1.2135732511958
    logpctmin80   1.0894301595173
          polpc   1.64924225574973
```

Extended Model and Joint Statistical Testing

Section 3.1 Examining Robustness Using Joint Statistical Testing Against Model 1 and Model 2

We find that the additional terms in model 2 (prbarr, pctmin80, pctymle, density) are joint statistically significant, as expected. Next we examine our derivations on model 2 as discussed in section 2 to see which variables are most important.

--need to discuss this in depth--

In [62]:

```
restricted_1 <- lm(log(crmrte) ~ prbconv + taxpc + polpc, data = crime)
coeftest(restricted_1, vcov = vcovHC)
unrestricted_1<- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80) +
              taxpc + polpc, data = crime)
coeftest(unrestricted_1, vcov = vcovHC)
# To test whether the difference in fit is significant, we use the wald test,
# which generalizes the usual F-test of overall significance,
# but allows for a heteroskedasticity-robust covariance matrix
waldtest(restricted_1, unrestricted_1, vcov = vcovHC)

# Another useful command is linearHypothesis, which allows us
# to write the hypothesis we're testing clearly in string form.
#linearHypothesis(unrestricted_1, c("sqrt(density)=0", "log(pctymle) = 0", "prbarr = 0", "prb
#                                   "prbconv*prbarr=0", "log(pctmin80)=0"), vcov = vcovHC)
#this does the same thing as wald
```

```
t test of coefficients:

              Estimate Std. Error  t value  Pr(>|t|)    
(Intercept) -3.681305   0.297749 -12.3638 < 2.2e-16 ***
prbconv     -0.629734   0.175627  -3.5856 0.0005574 ***
taxpc        0.012902   0.009452   1.3650 0.1758015    
polpc       -3.448917 316.067315  -0.0109 0.9913190    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.2883e+00  3.4718e-01 -6.5909 3.937e-09 ***
density        1.0322e-03  3.5978e-04  2.8689 0.0052364 ** 
log(pctymle)   1.7947e-01  1.3324e-01  1.3469 0.1817084    
prbarr        -1.9103e+00  3.1652e-01 -6.0354 4.403e-08 ***
prbconv       -6.9155e-01  1.0617e-01 -6.5134 5.534e-09 ***
log(pctmin80)  2.2254e-01  3.3826e-02  6.5790 4.149e-09 ***
taxpc          1.4847e-03  6.2665e-03  0.2369 0.8133091    
polpc          1.8262e+02  5.2737e+01  3.4629 0.0008519 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
82	4	48.31797	8.131153e-21

Due to the significance of the interaction term prbarr:prbconv, let's examine the effect on model 2 by combining these variables through multiplication. Muyltiplying these variables will give us the ratio of convictions to offenses.

In [63]:

```
crime$convarr <- crime$prbconv*crime$prbarr
crime_f <- select(crime, crmrte, prbarr, prbconv, prbpris, avgsen, polpc, density,
                    taxpc, west, central, urban, pctmin80, wcon, wtuc, wtrd, wfir, wser,
                    wmfg, wfed, wsta, wloc, mix, pctymle, convarr)
#--ask casper, for interaction terms are we actually just multiplying the data?--
```

In [64]:

```
model_2<- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80) +
              taxpc, data = crime)
model_2_a<- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80) +
              taxpc + prbarr*prbconv, data = crime)
model_2_b<- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80) +
              taxpc + convarr, data = crime_f)

stargazer(model_2, model_2_a, model_2_b,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.model2_1, se.model2_3, se.model2_3, se.model2_4, se.model2_5, se.model
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
==============================================
                     Dependent variable:
               -------------------------------
                         log(crmrte)
                  (1)        (2)        (3)
----------------------------------------------
density         0.001**     0.001      0.001
                (0.0004)   (0.004)    (0.004)

log(pctymle)     0.400      0.352      0.352
                (0.134)    (0.263)    (0.263)

prbarr         -1.118***  -2.201***  -2.201***
                (0.632)    (0.629)    (0.629)

prbconv        -0.508***  -0.876***  -0.876***
                (0.175)    (0.175)    (0.175)

log(pctmin80)   0.183***   0.203***  0.203***
                (0.047)    (0.048)    (0.048)

taxpc            0.006      0.006      0.006
                (0.008)    (0.009)    (0.009)

prbarr:prbconv              1.148


convarr                                1.148


Constant       -2.058***   -1.782*    -1.782*
                (0.350)    (0.833)    (0.833)

----------------------------------------------
Observations       90         90        90
Adjusted R2      0.729      0.748      0.748
==============================================
Note:            *p<0.05; **p<0.01; ***p<0.001
```

Section 3.1 Examining Robustness Using Joint Statistical Testing Against Model 2 and Derivations on Model 2 Incorporating Interaction Terms

We will examine three derivations for joint statistical significance based on our robustness examination in section 2 using interaction terms and indicator variables. We find that we could drop taxpc from model 2. We keep it in, since it is an important statistic to the political campaign.

In [65]:

```
model_2<- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80) +
              taxpc, data = crime)
model_2_a <- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80) +
              taxpc + prbarr*prbconv + density*log(pctymle)+
              prbarr*polpc * logpctymle+density*log(pctmin80), data = crime)
model_2_c <- lm(log(crmrte) ~ density + log(pctymle) + prbarr + prbconv + log(pctmin80),
                data = crime)
# To test whether the difference in fit is significant, we use the wald test,
# which generalizes the usual F-test of overall significance,
# but allows for a heteroskedasticity-robust covariance matrix
waldtest(model_2, model_2_a, vcov = vcovHC)
coeftest(model_2, vcov = vcovHC)
coeftest(model_2_a, vcov = vcovHC)
waldtest(model_2_c, model_2_a, vcov = vcovHC)
coeftest(model_2_c, vcov = vcovHC)
#waldtest(restricted_2, unrestricted_2, vcov = vcovHC)
# Another useful command is linearHypothesis, which allows us
# to write the hypothesis we're testing clearly in string form.
#linearHypothesis(unrestricted_1, c("sqrt(density)=0", "log(pctymle) = 0", "prbarr = 0", "prb
#                                   "prbconv*prbarr=0", "log(pctmin80)=0"), vcov = vcovHC)
#this does the same thing as wald
```

Res.Df	Df	F	Pr(>F)
83	NA	NA	NA
75	8	1.197254	0.3122821

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.05806778  0.35034024 -5.8745 8.498e-08 ***
density        0.00137764  0.00044321  3.1083 0.0025771 ** 
log(pctymle)   0.39983828  0.13366453  2.9914 0.0036558 ** 
prbarr        -1.11818614  0.63247311 -1.7680 0.0807427 .  
prbconv       -0.50824503  0.17533131 -2.8988 0.0047914 ** 
log(pctmin80)  0.18336927  0.04652371  3.9414 0.0001683 ***
taxpc          0.00625563  0.00849493  0.7364 0.4635663    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t test of coefficients:

                           Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)              2.9756e+00  5.4003e+00  0.5510   0.58327    
density                  2.7654e-03  3.1575e-03  0.8758   0.38392    
log(pctymle)             2.2272e+00  2.1870e+00  1.0184   0.31177    
prbarr                  -1.8180e+01  1.9375e+01 -0.9383   0.35111    
prbconv                 -1.1619e+00  5.0435e-01 -2.3038   0.02400 *  
log(pctmin80)            2.8678e-01  6.6749e-02  4.2964 5.141e-05 ***
taxpc                    7.6565e-05  5.5952e-03  0.0137   0.98912    
polpc                   -2.6778e+03  2.5685e+03 -1.0426   0.30050    
prbarr:prbconv           2.4047e+00  2.3701e+00  1.0146   0.31357    
density:log(pctymle)     1.2269e-03  1.4454e-03  0.8488   0.39869    
prbarr:polpc             8.3169e+03  1.0291e+04  0.8081   0.42157    
prbarr:logpctymle       -6.1931e+00  7.7450e+00 -0.7996   0.42646    
polpc:logpctymle        -1.1903e+03  1.0521e+03 -1.1313   0.26152    
density:log(pctmin80)   -9.1903e-04  5.3346e-04 -1.7228   0.08905 .  
prbarr:polpc:logpctymle  3.3767e+03  4.0925e+03  0.8251   0.41193    
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Res.Df	Df	F	Pr(>F)
84	NA	NA	NA
75	9	1.087093	0.382455

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -1.97303514  0.28206920 -6.9949 5.958e-10 ***
density        0.00153707  0.00029639  5.1860 1.462e-06 ***
log(pctymle)   0.33567720  0.12894948  2.6032  0.010918 *  
prbarr        -1.16696476  0.64939179 -1.7970  0.075928 .  
prbconv       -0.52989764  0.17870340 -2.9652  0.003936 ** 
log(pctmin80)  0.18532783  0.04393082  4.2186 6.178e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 3.1 Examining Robustness Using Joint Statistical Testing Against Model 2 and Derivations on Model 3

Model 3 incorporates all potential covariants, with the exception of mix (justify based on other y variable and and any other considerations you can think of?). Due to the variety of units in model 3, we will use standardized variables.

In [66]:

```
base_model <- lm(log(crmrte) ~ prbconv + taxpc + polpc, data = crime)
model_2<- lm(log(crmrte) ~ prbconv + taxpc + density + log(pctymle) + prbarr +
             log(pctmin80), data = crime)
model_2_step <- lm(log(crmrte)~prbconv +polpc+ density +prbarr+log(pctmin80),
                   data = crime)
model_3_a <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + density + log(pctymle) +
                prbarr + log(pctmin80) + prbpris + avgsen + taxpc + wcon + wtuc +
                wtrd + wfir + wser + wmfg + wfed + wsta + wloc, data = crime)
model_3 <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + density + log(pctymle) + prbarr +
              log(pctmin80) + prbpris + avgsen + taxpc + wklywage, data = crime)

model_3_wr <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + density + log(pctymle) + prbarr +
                 log(pctmin80) + prbpris + avgsen + taxpc, data = crime)
model_3_wr <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + density + log(pctymle) + prbarr +
                 log(pctmin80) + prbpris + avgsen + taxpc, data = crime)
model_3_wr_dr <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + prbarr + prbpris + avgsen +
                    taxpc, data = crime)
model_3_wr_denr <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + prbarr + prbpris + avgsen + dens
                    taxpc, data = crime)
model_3_wr_pa <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + density + log(pctymle) + prbarr +
                    log(pctmin80)  + taxpc, data = crime)
model_3_optimal <- lm(log(crmrte) ~ prbconv+ taxpc + polpc + density + log(pctymle) +
                      prbarr + log(pctmin80) + taxpc, data = crime)
```

```
In [67]: # All Variables
stargazer(base_model, model_2, model_2_step, model_3,
          type="text", keep.stat=c("n", "adj.rsq"), dep.var.labels=c('Crime Rate (Log Transfo
          star.cutoffs=c(0.05, 0.01, 0.001), covariate.labels=c('Ratio of Convictions to Arre
            'Tax Per Capita', 'Police per Capita', 'Density', 'Percent Young Male (Log Transf
            'Ratio of Arrests to Offenses', 'Percent Minority in 1980 (Log Transformed)',
            'Ratio of Prison Sentence to Convictions', 'Average Sentence Length (years)',
            'Weekly Wage (mean of all wages)'), omit.stat=c("LL","ser","f"), no.space=TRUE,
         column.labels=c('Simple Reg.','Optimal','Optimal STEP','Extended Variant') )

stargazer(base_model, model_2, model_3, model_3_optimal, model_3_a, type='text',
          star.cutoffs=c(.05, .01, .001))
```

```
===========================================================================================
===
                                                           Dependent variable:
                                           ------------------------------------------------
---
                                                      Crime Rate (Log Transformed)
                                           Simple Reg.  Optimal  Optimal STEP Extended Vari
ant
                                               (1)        (2)        (3)            (4)
-------------------------------------------------------------------------------------------
---
Ratio of Convictions to Arrests             -0.630***  -0.508***  -0.729***      -0.692***
                                             (0.145)    (0.091)    (0.084)        (0.091)
Tax Per Capita                               0.013**    0.006*                     0.002
                                             (0.004)    (0.002)                   (0.003)
Police per Capita                            -3.449               201.695***     173.341***
                                            (53.638)               (35.174)       (43.639)
Density                                                0.001***    0.001***       0.001**
                                                       (0.0002)    (0.0002)       (0.0003)
Percent Young Male (Log Transformed)                    0.400*                     0.206
                                                        (0.162)                   (0.155)
Ratio of Arrests to Offenses                           -1.118***  -2.044***      -1.840***
                                                        (0.240)    (0.250)        (0.282)
Percent Minority in 1980 (Log Transformed)             0.183***    0.226***       0.223***
                                                        (0.032)    (0.030)        (0.030)
Ratio of Prison Sentence to Convictions                                            0.035
                                                                                  (0.348)
Average Sentence Length (years)                                                    -0.004
                                                                                  (0.011)
Weekly Wage (mean of all wages)                                                    0.002
                                                                                  (0.001)
Constant                                    -3.681***  -2.058***  -2.646***      -2.709***
                                             (0.183)    (0.398)    (0.109)        (0.481)
-------------------------------------------------------------------------------------------
---
Observations                                   90         90          90             90
Adjusted R2                                   0.267      0.729      0.782          0.780
===========================================================================================
===
Note:                                                             *p<0.05; **p<0.01; ***p<0.
001

===========================================================================================
=============================================
                                                                    Dependent variable:
                    -----------------------------------------------------------------------
---------------------------------------------
                                                                        log(crmrte)
                             (1)                    (2)                     (3)
(4)                     (5)
```

```
--------------------------------------------------------------------------------------------
---------------------------------------------
prbconv                   -0.630***               -0.508***               -0.692***
-0.692***               -0.662***
                           (0.145)                 (0.091)                 (0.091)
(0.091)                 (0.088)

taxpc                      0.013**                  0.006*                  0.002
0.001                   0.004
                           (0.004)                 (0.002)                 (0.003)
(0.002)                 (0.003)

polpc                       -3.449                                       173.341***
182.624***             162.149***
                           (53.638)                                       (43.639)
(40.074)               (43.675)

density                                            0.001***                0.001**
0.001***                0.001**
                                                   (0.0002)               (0.0003)
(0.0002)               (0.0003)

log(pctymle)                                        0.400*                  0.206
0.179                  0.431**
                                                   (0.162)                 (0.155)
(0.153)                 (0.154)

prbarr                                            -1.118***               -1.840***
-1.910***               -1.821***
                                                   (0.240)                 (0.282)
(0.277)                 (0.266)

log(pctmin80)                                      0.183***               0.223***
0.223***               0.194***
                                                   (0.032)                 (0.030)
(0.030)                 (0.030)

prbpris                                                                     0.035
-0.096
                                                                           (0.348)
(0.335)

avgsen                                                                     -0.004
-0.007
                                                                           (0.011)
(0.011)

wklywage                                                                    0.002
                                                                           (0.001)


wcon
0.0004

(0.001)

wtuc
0.0002

(0.0004)

wtrd
0.0002

(0.001)
```

```
wfir
-0.001

(0.001)

wser
-0.002*

(0.001)

wmfg
-0.00005

(0.0004)

wfed
0.003***

(0.001)

wsta
-0.001*

(0.001)

wloc
0.001

(0.001)

Constant                  -3.681***              -2.058***               -2.709***
-2.288***             -2.058***
                             (0.183)                (0.398)                 (0.481)
(0.361)               (0.562)

-------------------------------------------------------------------------------------------
---------------------------------------------
Observations                  90                     90                      90
90                     90
R2                          0.292                  0.747                   0.804
0.798                   0.851
Adjusted R2                 0.267                  0.729                   0.780
0.781                   0.813
Residual Std. Error    0.470 (df = 86)        0.286 (df = 83)         0.258 (df = 79)
0.257 (df = 82)         0.237 (df = 71)
F Statistic         11.824*** (df = 3; 86) 40.842*** (df = 6; 83) 32.463*** (df = 10; 79) 4
6.312*** (df = 7; 82) 22.493*** (df = 18; 71)
===========================================================================================
=============================================
Note:
*p<0.05; **p<0.01; ***p<0.001
```

```
In [68]: coeftest(model_3, vcov = vcovHC) #heteroskadastic SE and values and p values
```

```
t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.7090e+00  6.9434e-01 -3.9015 0.0001997 ***
prbconv       -6.9152e-01  1.0964e-01 -6.3072 1.524e-08 ***
taxpc          1.6516e-03  6.8294e-03  0.2418 0.8095372    
polpc          1.7334e+02  6.8719e+01  2.5225 0.0136622 *  
density        7.9500e-04  3.9931e-04  1.9909 0.0499472 *  
log(pctymle)   2.0628e-01  1.6954e-01  1.2167 0.2273437    
prbarr        -1.8399e+00  3.5385e-01 -5.1995 1.531e-06 ***
log(pctmin80)  2.2289e-01  3.4832e-02  6.3991 1.026e-08 ***
prbpris        3.4835e-02  5.2776e-01  0.0660 0.9475406    
avgsen        -3.5159e-03  1.4197e-02 -0.2476 0.8050480    
wklywage       1.6300e-03  1.7904e-03  0.9104 0.3653830    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
In [69]: coeftest(model_3_optimal, vcov = vcovHC)
         vcovHC(model_3_optimal) #heteroskadastic SE and values and p values
```

```
t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.2883e+00  3.4718e-01 -6.5909 3.937e-09 ***
prbconv       -6.9155e-01  1.0617e-01 -6.5134 5.534e-09 ***
taxpc          1.4847e-03  6.2665e-03  0.2369 0.8133091    
polpc          1.8262e+02  5.2737e+01  3.4629 0.0008519 ***
density        1.0322e-03  3.5978e-04  2.8689 0.0052364 ** 
log(pctymle)   1.7947e-01  1.3324e-01  1.3469 0.1817084    
prbarr        -1.9103e+00  3.1652e-01 -6.0354 4.403e-08 ***
log(pctmin80)  2.2254e-01  3.3826e-02  6.5790 4.149e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	(Intercept)	prbconv	taxpc	polpc	density	log(pctymle)	prbarr
(Intercept)	1.205364e-01	9.577732e-03	-7.893158e-04	-8.591858e+00	3.202033e-05	3.632247e-02	2.605363e-02
prbconv	9.577732e-03	1.127265e-02	-2.126829e-04	-2.521563e+00	2.351726e-05	5.083072e-03	1.936149e-02
taxpc	-7.893158e-04	-2.126829e-04	3.926910e-05	3.781992e-02	-1.639688e-06	1.250480e-04	-5.644318e-04
polpc	-8.591858e+00	-2.521563e+00	3.781992e-02	2.781243e+03	-8.531233e-03	-2.929822e+00	-7.317653e+00
density	3.202033e-05	2.351726e-05	-1.639688e-06	-8.531233e-03	1.294436e-07	1.691234e-06	6.316458e-05
log(pctymle)	3.632247e-02	5.083072e-03	1.250480e-04	-2.929822e+00	1.691234e-06	1.775401e-02	1.495225e-02
prbarr	2.605363e-02	1.936149e-02	-5.644318e-04	-7.317653e+00	6.316458e-05	1.495225e-02	1.001836e-01
log(pctmin80)	2.140989e-03	4.497694e-04	-8.131639e-05	-5.232944e-02	3.408489e-06	-5.315914e-04	2.729333e-03

Robustness model 3 vs model 2 We want to answer three questions 1) Do wages demonstrate joint statistically significance? 2) Does density, pctmale, pctmin80 demonstrate joint statistical significance? -Let's compare this to the equation without any wages, and in this case, they are statistically significant...so let's just readd density and see what happens...ok so let's -So probability of prison, average sentence, and wages do not matter An optimal model might look like

In [70]:

```
waldtest(model_3_wr, model_3, vcov = vcovHC)
coeftest(model_3_wr, vcov = vcovHC)
#wages do not seem to matter#
```

Res.Df	Df	F	Pr(>F)
80	NA	NA	NA
79	1	0.8288212	0.365383

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.2874e+00  4.1323e-01 -5.5353 3.820e-07 ***
prbconv       -6.9015e-01  1.1309e-01 -6.1028 3.526e-08 ***
taxpc          1.5178e-03  6.3994e-03  0.2372  0.813121    
polpc          1.8462e+02  6.3891e+01  2.8897  0.004961 ** 
density        1.0301e-03  3.6512e-04  2.8214  0.006029 ** 
log(pctymle)   1.8245e-01  1.4353e-01  1.2712  0.207354    
prbarr        -1.9108e+00  3.2332e-01 -5.9100 7.994e-08 ***
log(pctmin80)  2.2192e-01  3.5358e-02  6.2764 1.675e-08 ***
prbpris        4.3320e-02  5.6151e-01  0.0771  0.938698    
avgsen        -1.7831e-03  1.5161e-02 -0.1176  0.906668    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In [71]:

```
waldtest(model_3_wr_dr, model_3_wr, vcov = vcovHC)
coeftest(model_3_wr_dr, vcov = vcovHC)
```

Res.Df	Df	F	Pr(>F)
83	NA	NA	NA
80	3	15.24825	6.179305e-08

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)  -2.9352598   0.4190978 -7.0038 6.006e-10 ***
prbconv      -0.8112243   0.1266342 -6.4060 8.527e-09 ***
taxpc         0.0047449   0.0042098  1.1271   0.26294    
polpc       194.4615873 100.5994417  1.9330   0.05664 .  
prbarr       -2.5211265   0.4944821 -5.0985 2.122e-06 ***
prbpris       0.3182615   0.6505608  0.4892   0.62598    
avgsen       -0.0061057   0.0185021 -0.3300   0.74223    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
In [72]: waldtest(model_3_wr_denr, model_3_wr, vcov = vcovHC)
         coeftest(model_3_wr_denr, vcov = vcovHC)
```

Res.Df	Df	F	Pr(>F)
82	NA	NA	NA
80	2	22.40166	1.881713e-08

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) -3.0879e+00  3.9065e-01 -7.9046 1.079e-11 ***
prbconv     -6.5396e-01  1.3688e-01 -4.7777 7.672e-06 ***
taxpc        2.9909e-03  4.6961e-03  0.6369 0.5259686    
polpc        1.2867e+02  1.0329e+02  1.2458 0.2163898    
prbarr      -1.9028e+00  5.1216e-01 -3.7152 0.0003695 ***
prbpris      9.2490e-02  5.6734e-01  0.1630 0.8709008    
avgsen      -7.9560e-03  1.5805e-02 -0.5034 0.6160364    
density      1.2080e-03  3.4103e-04  3.5422 0.0006578 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

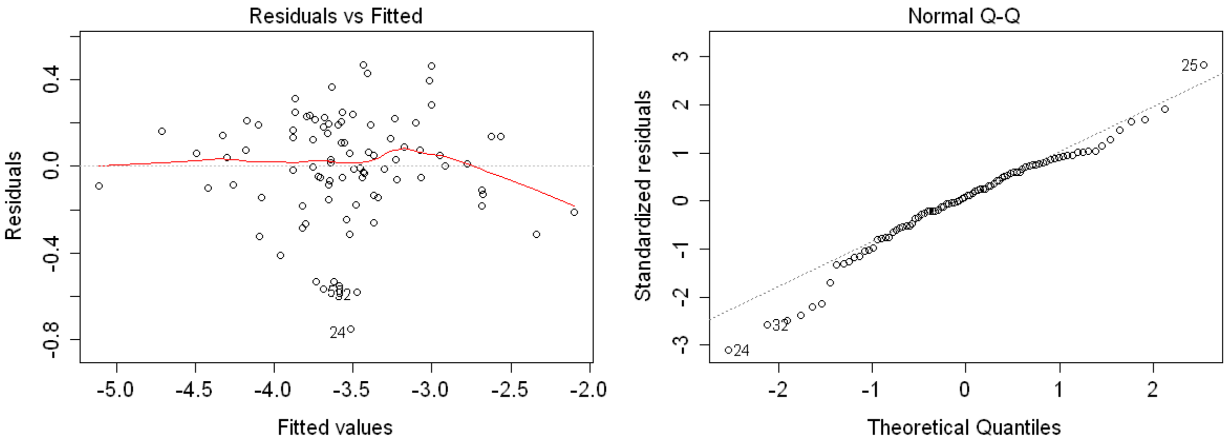
```
In [73]: waldtest(model_3_wr, model_3_wr_pa, vcov = vcovHC)
         coeftest(model_3_wr_pa, vcov = vcovHC)
```

Res.Df	Df	F	Pr(>F)
80	NA	NA	NA
82	-2	0.01536072	0.9847596

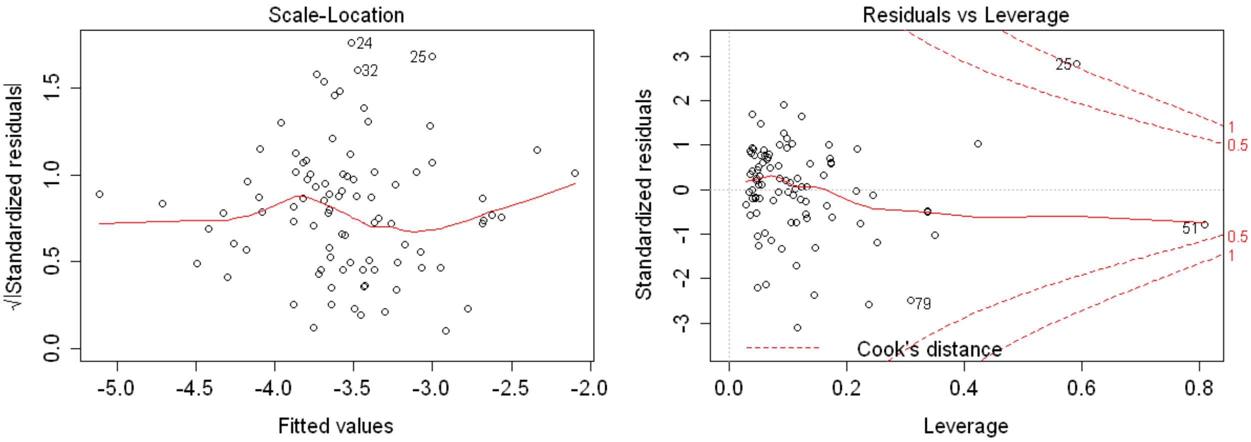
```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)   -2.2883e+00  3.4718e-01 -6.5909 3.937e-09 ***
prbconv       -6.9155e-01  1.0617e-01 -6.5134 5.534e-09 ***
taxpc          1.4847e-03  6.2665e-03  0.2369 0.8133091    
polpc          1.8262e+02  5.2737e+01  3.4629 0.0008519 ***
density        1.0322e-03  3.5978e-04  2.8689 0.0052364 ** 
log(pctymle)   1.7947e-01  1.3324e-01  1.3469 0.1817084    
prbarr        -1.9103e+00  3.1652e-01 -6.0354 4.403e-08 ***
log(pctmin80)  2.2254e-01  3.3826e-02  6.5790 4.149e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

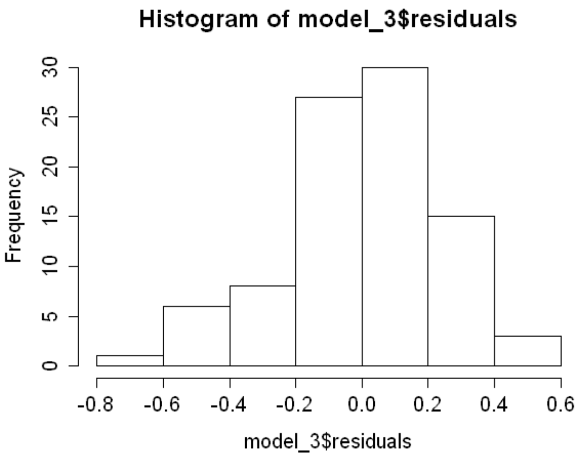
In [74]:

```
par(mfrow=c(1,2))
options(repr.plot.height = 4.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
#plot(m1)
#plot(m2)
plot(model_3)
hist(model_3$residuals)
bptest(model_3)
shapiro.test(model_3$residuals)
par(mfrow=c(1,4))
options(repr.plot.height = 4.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
```



```
	studentized Breusch-Pagan test

data:  model_3
BP = 17.56, df = 10, p-value = 0.06286


	Shapiro-Wilk normality test

data:  model_3$residuals
W = 0.96024, p-value = 0.007581
```





In [ ]:

Extended Model Assumptions

CLM Assumptions

1) **Linear in Parameters** We developed a model with linear parameters.

2) **Random Sampling** The data does not appear to be random. 10 counties are missing from the dataset with no explanation. Those counties notably include some of the most sparsely populated counties, each having a population of less that 68,000 people and are not densely populated. The average population per county in North Carolina is just over 105,068. Additionally, our data is from two different databases. One is the North Carolina Department of Correction and the other is the FBI's Uniform Crime Reports. Each of these purports that it is comprehensive, but we don't necessarily have enough information to be sure that combining these datasets creates a representative sample, or if it biases the data towards crimes more likely to be investigated by the FBI to include many felonies, kidnappings, and cross-state line offenses that wind up or start in North Carolina.

3) **Sample Varation and No Perfect Colinearity** We see from our correlation matrix that we have no perfect collinearity. We can detect and represent collinearity using the Variance Inflation Factor (VIF). If any of these variance inflation factors exceed 5 (the cut-off typically used), the associated regression coefficients are poorly estimated due to multicollinearity. This suggests that the Multicollinearity assumption is not violated.

4) **Zero Conditional Mean/No Omitted Varibles** Our Residuals vs Fitted Plot makes it appear as though we meet the zero conditional mean assumption.

In [75]:

```
round(mean(model_3$residuals), 4)

vif(model_3)
```

```
0

       prbconv   1.40557615762643
         taxpc   1.4405152424213
         polpc   2.5065194586403
       density   2.18404138160075
 log(pctymle)    1.26762381803675
        prbarr   2.01793011265741
log(pctmin80)    1.12446210446578
       prbpris   1.05877251360096
        avgsen   1.38047047835275
      wklywage   1.94648678922126
```

---Our data is linear, but not unbiased as the sampling is not random. All conclusions must be interpreted with the understanding that rural counties are underrepresented.

5) **Homoskedasticity and No Serial Correlation/autocorrelation** Our points also look generally spread among equal width, but they may have a greater massing of points towards the middle, so we will analyze quantitatively. Our Scale-Location plot demonstrates general homeskedasticity, but it does waver toward the center of the plot. Our variance is not completely uniform. We do have outliers according to our Residuals vs Leverage that our outliers still fall within the .5 scale. Quantitatively, assumption of constant error variance can be tested using the ncvTest function.

---Our data is not sufficiently homoskedastic for large and small values. It is also not random, so we have not necessarily achieved best linear unbiased estimator.

6) **Normality** Both our histogram of residuals and Normal Q-Q plot demonstrate that we do have normally distributed errors except for where our values are very small. With a dataset of n=90, it is sufficiently large to account for the few lower values that do not fall along the line. Our Shapiro Wilk Normality Test also demonstrates that we fail to reject the null hypothesis demonstrates that we do draw from a normal population.

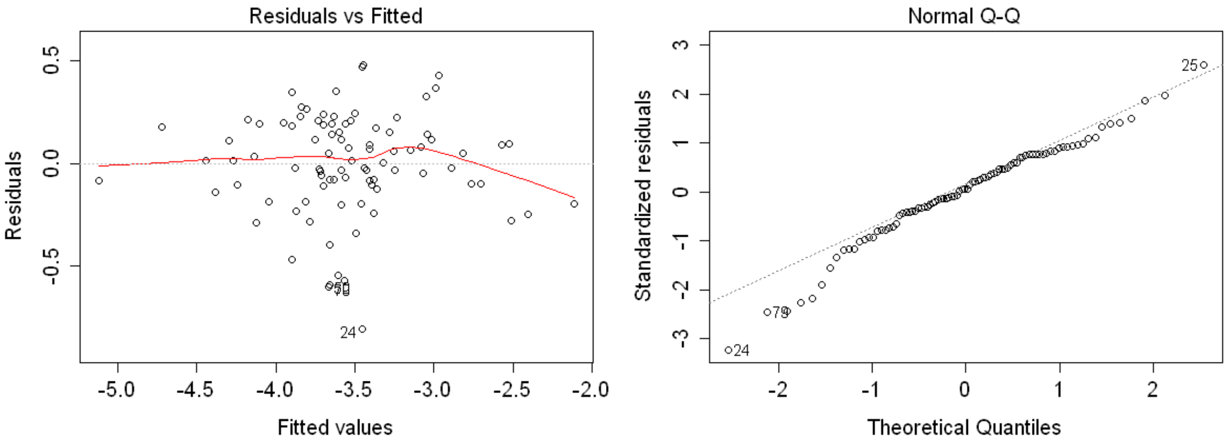
Exogeneity?

In [76]:

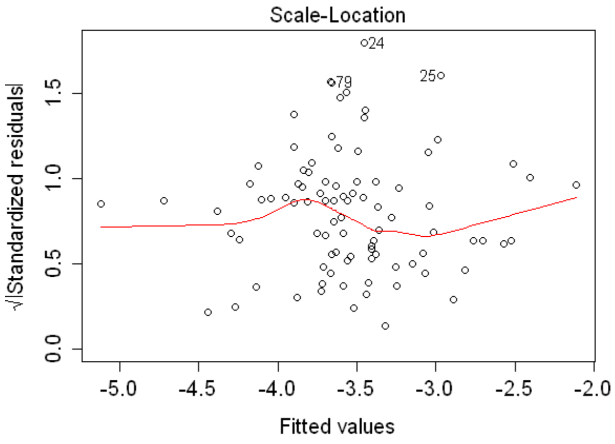
```
ncvTest(model_3)
```

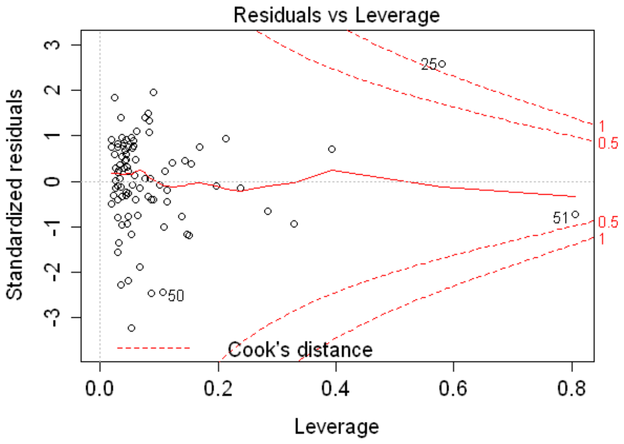
```
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 0.03557157    Df = 1     p = 0.850403
```

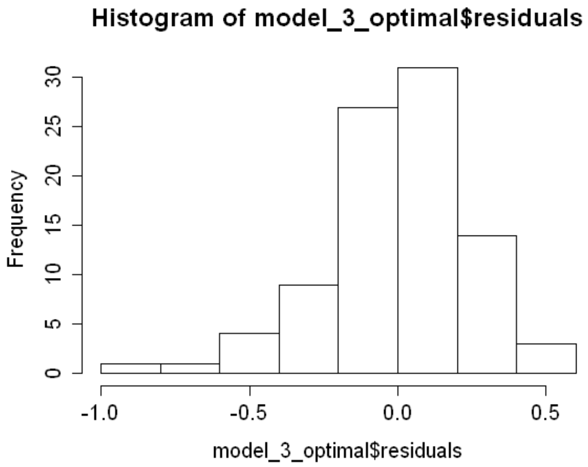
```
In [77]: par(mfrow=c(1,2))
         options(repr.plot.height = 4.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
         #plot(m1)
         #plot(m2)
         plot(model_3_optimal)
         hist(model_3_optimal$residuals)
         bptest(model_3_optimal)
         shapiro.test(model_3_optimal$residuals)
         par(mfrow=c(1,4))
         options(repr.plot.height = 4.0, repr.plot.width = 10.0, repr.plot.pointsize = 9.0)
```



```
	studentized Breusch-Pagan test

data:  model_3_optimal
BP = 13.962, df = 7, p-value = 0.05185


	Shapiro-Wilk normality test

data:  model_3_optimal$residuals
W = 0.96065, p-value = 0.008074
```







Extended Model Interpretation

(Moved to main body of paper.)

Additional Resources

https://www.americanprogress.org/issues/economy/reports/2012/06/19/11755/the-economic-benefits-of-reducing-violent-crime/ (https://www.americanprogress.org/issues/economy/reports/2012/06/19/11755/the-economic-benefits-of-reducing-violent-crime/) *Discusses the economic benefits of reducing crime.*

https://www.energy.gov/lm/services/environmental-justice/environmental-justice-history (https://www.energy.gov/lm/services/environmental-justice/environmental-justice-history) and https://www.bls.gov/news.release/history/annpay_091693.txt (https://www.bls.gov/news.release/history/annpay_091693.txt). *Details Warren County's background as poor justifying the reduction of the wser outlier value for the analysis.*

https://www2.census.gov/library/publications/decennial/1990/cp-1/cp-1-35.pdf (https://www2.census.gov/library/publications/decennial/1990/cp-1/cp-1-35.pdf). *We used the 1990 census to verify density values for the counties included and for the 10 counties that were not included in the dataset.*

https://www.continentaltelegraph.com/economy/increasing-the-minimum-wage-increases-crime-obviously-enough/ (https://www.continentaltelegraph.com/economy/increasing-the-minimum-wage-increases-crime-obviously-enough/) *Increases in wage seem to result in increased crime in other regions across the United States as well. We anticipate that our taxpc variable has a direct relationship with crime as a result. However, this is likely a correlational, not causational relationship. Better economies may result in a better ability to catch criminals.*